Applied Psychological Measurement
Volume 33 Number 3
May 2009 163-183

10.1177/0146621608320523
http://apm.sagepub.com
hosted at
http://online.sagepub.com

A Cognitive Diagnosis Model for Cognitively Based Multiple-Choice Options

Jimmy de la Torre
Rutgers, The State University of New Jersey

Cognitive or skills diagnosis models are discrete latent variable models developed specifically for the purpose of identifying the presence or absence of multiple fine-grained skills. However, applications of these models typically involve dichotomous or dichotomized data, including data from multiple-choice (MC) assessments that are scored as right or wrong. The dichotomization approach to the analysis of MC data ignores the potential diagnostic information that can be found in the distractors and is therefore deemed diagnostically suboptimal. To maximize the diagnostic value of MC assessments, this article prescribes how MC options should be constructed to make them more cognitively diagnostic and proposes a cognitive diagnosis model for analyzing such data. The article discusses the specification of the proposed model and estimation of its parameters. Moreover, results of a simulation study evaluating the viability of the model and an estimation algorithm are presented. Finally, practical considerations concerning the proposed framework are discussed.

Keywords: *cognitive diagnosis; multiple choice; DINA; EM algorithm; parameter estimation*

The aim of assessment is "to educate and improve student performance and not merely to audit it" (Wiggins, 1998, p. xi). According to Stiggins (2002), assessment should be used not only to ascertain the status of learning but also to further learning. However, with the growing emphasis on accountability, more and more available resources are disproportionately allocated toward assessments that only audit learning but do not provide information that can facilitate instruction and learning. Cognizant of the disparity in resource allocation, the report of the Committee on the Foundations of Assessment recommended shifting more research and training investments toward assessments that can facilitate learning (National Research Council [NRC], 2001). Hence, assessments that can provide information deemed "interpretative, diagnostic, highly informative, and potentially prescriptive" (Pellegrino, Baxter, & Glaser, 1999, p. 335) need to be developed.

Shepard (NRC, 2003) noted that assessments typically used to support school and system accountability do not provide diagnostic information about individual students to support learning. These assessments are based on measurement models (unidimensional item response theory or IRT models) for building tests from homogeneous items and scoring procedures that submerge any distinct skills that may remain in a single reported value. Although scores from these assessments have been useful in establishing the relative order

Author's Note: The work reported here was supported by a National Academy of Education/Spencer Postdoctoral fellowship. Please address correspondence to Jimmy de la Torre, Graduate School of Education, 10 Seminary Place, New Brunswick, NJ 08901; e-mail: j.delatorre@rutgers.edu.

Table 1
Attributes in Fraction Subtraction

Attribute	Description
1	Borrow one from whole number to fraction
2	Basic fraction subtraction
3	Reduce/simplify
4	Separate whole number from fraction
5	Convert whole number to fraction

of students along a single latent proficiency continuum, the information contained in these scores neither permit evaluation of students' specific strengths and weaknesses necessary for targeted instruction nor can it be used as a feedback mechanism to allow teachers identify effective classroom methods and practices that can help students learn better. Attempts to provide more informative scores such as reporting cluster scores in a content strand have been made. However, aside from unreliability (e.g., de la Torre & Patz, 2005; Wainer et al., 2001), the information provided by cluster scores is superficial and may not reflect deeper underlying processes involved in learning and problem solving.

In contrast to the traditional models (i.e., IRT), *Knowing What Students Know* (NRC, 2001) discusses some measurement models that allow the merger of advances in cognitive and psychometric theories and facilitate inferences more relevant to learning. These psychometric models are known as cognitive diagnosis models (CDMs) and can be used to understand the skills, cognitive processes, and problem-solving strategies involved in an assessment. The CDMs are discrete latent variable models developed specifically for diagnosing the presence or absence of multiple fine-grained skills or processes required for solving problems on a test. With these models, profiles can be generated for a student or a group of students (e.g., class) to indicate which specific skills or processes the students have and have not mastered. Such information can be used to direct resources and tailor instruction to optimize student learning.

To illustrate the fundamental difference between IRT and CDM frameworks, consider the fraction-subtraction task $2\frac{4}{7} - \frac{7}{12}$. An IRT model might describe the performance on the task as a function of a global-subtraction proficiency and that students with higher proficiencies are expected to have higher probabilities of answering the item correctly. In contrast, a CDM might describe the performance as a function of the attributes listed 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, assessments developed and analyzed using CDMs provide information that is richer, more prescriptive, and more relevant to instruction and learning.

Although not without its share of criticisms, the multiple-choice (MC) format has been widely used for a variety of reasons, and the most important of which is its ability to sample and accommodate diverse contents (Nitko, 2001; Osterlind, 1998). In cognitive diagnosis

modeling, the most straightforward and common way of analyzing MC responses is to treat them as dichotomous data. For instance, this approach has been employed by Birenbaum, Tatsuoka, and Xin (2005) and Tatsuoka, Corter, and Tatsuoka (2004) to analyze Trends in International Mathematics and Science Study (TIMSS; 2003) data and by de la Torre (2006) to analyze the National Assessment of Educational Progress (NAEP; 2003) data. However, such an approach is suboptimal because it does not take into account the diagnostic insights about student difficulties and alternative conceptions that can be found in the distractors (Haertel & Wiley, 1993; Nitko, 2001; Sadler, 1998).

To maximize the diagnostic value of MC assessments, this article proposes a cognitive diagnosis framework for collecting and analyzing MC data. Specifically, the framework prescribes how options in MC items can be constructed to provide more diagnostic information and describes a psychometric model that can exploit such information. The details of the framework are laid out in the next section. The section on "An Expectation-Maximization (EM) Algorithm for the MC-DINA Model" discusses a method of estimating the model parameters, whereas the "Simulation Study" section evaluates the viability of the proposed framework. The final section discusses some practical considerations involved in implementing the framework.

A Cognitively Diagnostic MC Framework

Background

The proposed framework for MC data is based on the deterministic-input, noisy "and" gate (DINA; Junker & Sijtsma, 2001) model. Therefore, prior to the introduction of the framework, a discussion of the DINA model is in order.

Let the binary vectors $\mathbf{X}_i$ and $\boldsymbol{\alpha}_i$ denote student i's responses to J items and K latent attributes, respectively. Under the DINA model, the probability that the student will answer item j correctly is given by

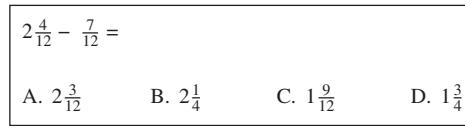
$$\begin{aligned} P\left(X_{ij}=1|\boldsymbol{\alpha}_i\right) &= P\left(X_{ij}=1|g_{ij}\right) \\ &= P_j(1|0)^{1-g_{ij}}P_j(1|1)^{g_{ij}} \quad (1) \\ &= P_j(1|0)^{1-g_{ij}}[1-P_j(0|1)]^{g_{ij}}, \quad (2) \end{aligned}$$

where P_j $(1|0)$ and P_j $(0|1)$ are the guessing and slip parameters, g_{ij}, the latent group classification of the examinee with respect to item j, is equal to $\prod_{k=1}^{K}\alpha_{ik}^{q_{jk}}$, P_j $(h|g)=P(X_{\cdot j}=h|g_{\cdot j}=g)$, $h,g=0,1$, and q_{jk} is the entry in the j^{th} row and k^{th} column of the Q-matrix (Embretson, 1984; Tatsuoka, 1983)—a $J\times K$ binary matrix that specifies the attribute requirements for each item. An alternative way of determining the value of g_{ij} is as follows:

$$g_{ij}=\begin{cases}1 & \text{if } \boldsymbol{\alpha}_i'\mathbf{q}_j=\mathbf{q}_j'\mathbf{q}_j \\ 0 & \text{otherwise}\end{cases}.$$

That is, student i is classified in the latent group 0 with respect to item j if and only if the student's attribute vector does not meet the attribute requirement of item j specified in the

Figure 1
A Fraction-Subtraction Item With Coded Options



vector $\mathbf{q}_j$. Finally, each attribute vector can represent a unique latent class or knowledge state; thus, given K attributes, the total number of latent classes is 2^K.

The formulation of the DINA shows that it is a conjunctive model in that a correct response to an item necessitates students to possess all the attributes required for the item. Moreover, the absence of a required attribute cannot be made up for by the presence of other attributes, required or otherwise. Consequently, examinees who lack one of the required attributes are not differentiated from those who lack several or all of the required attributes. This property allows the DINA model to be a parsimonious yet interpretable model. It only requires two parameters for each item that do not depend on K. This model is appropriate in application where the conjunction of several equally important attributes is required. Although labeled differently, other applications and discussions of the DINA model can be found in Doignon and Falmagne (1999), Haertel (1989), Macready and Dayton (1977), and Tatsuoka (2002).

Cognitively Based MC Options

For the MC format $X_{ij} = 1, 2, \ldots, H_j$, each number represents a different option and H_j is total number of options in item j. In a cognitively diagnostic framework for MC items, an option is said to be coded or cognitively based if it is constructed to correspond to one of the $2^K - 1$ latent classes (i.e., classes defined by all the attribute vectors except the null vector); otherwise, the option is considered noncoded or not cognitively based.

An example of a MC item with four coded options is given in Figure 1; the attribute specifications for these options are given in Table 2. In general, cognitive diagnosis modeling should always be viewed as a interdisciplinary endeavor. However, this recommendation is more salient for the MC-DINA model because, in addition to defining the attributes and determining the appropriate tasks, experts (e.g., domain experts) play a critical role in developing distractors that are both relevant and informative. For this example, the different options for this item were constructed in consultation with an experienced mathematics educator. The option with the largest number of required attributes (i.e., D) is the key. As can be seen from this example, in addition to the key, some distractors are also coded under this framework. The attribute specification for a coded option, say h, is represented by the Q-vector $\mathbf{q}_{jh}$, whereas the attribute specification for noncoded options is implicitly represented by the null vector.

Although the above example will be used throughout the article, a different domain will be discussed briefly to show that the MC-DINA model has sufficient generality for it to be

Table 2
Attributes Required for Each Option of a Fraction-Subtraction Item

	Attribute				
Option	1	2	3	4	5
A		✓			
B		✓	✓		
C	✓	✓			
D	✓	✓	✓		

Table 3
Attributes in the Balance Scale Problems

Attribute	Description
1	Weight and distance (Rule III)
2	Weight and distance with emphasis on weight (Rule II)
3	Weight and distance with emphasis on distance
4	Addition (compensation/buggy)
5	Multiplication (Rule IV)

useful in applications outside educational measurement. For example, with some modifications (i.e., allowing the attribute specifications for the distractors to be unconstrained), the MC-DINA model can be applied to the area of cognition and development, specifically the balance scale problems, which have been extensively used to study proportional reasoning in children (e.g., Halford, Andrews, Dalton, Boag, & Zielinski, 2002; Shultz, Mareschal, & Schmidt, 1994; Siegler, 1976, 1981). Jansen and van der Maas (2002) have shown that children can account for both weight and distance but not necessarily their combination at about the age of 9. Thus, when confined to older children (i.e., at least 9 years old), the attributes in Table 3 can be defined for the balance scale problems. Except for Attribute 3, all the attributes have been previously defined in the literature (Jansen & van der Maas, 2002; Siegler, 1976, 1981; van der Maas & Jansen, 2003). Attribute 3 is similar to Attribute 1 except that when conflict problems are encountered children emphasize the distance over the weight dimension. In applying the attributes to the eight example problems given by van der Maas and Jansen (2003), it can be shown that the five attributes have a hierarchical structure depicted in Figure 2. The structure does not indicate the order by which the attributes develop but rather a subsumption of the attributes with respect to the balance scale problems that are amenable to each attribute. This structure indicates that problems amenable to Attributes 1 through 4 are also amenable to Attribute 5. The structure also shows that although all problems amenable to Attribute 3 are also amenable to Attribute 4, some problems amenable to Attribute 2 are not amenable to Attribute 4 (see Figure 3 for an example of the latter).

The balance scale problems can be viewed as MC tasks that provide the same three options across all items: left, right, and balance. Figure 3 is an example of a conflict-weight problem similar to that given by van der Maas and Jansen (2003). From the MC-DINA

Figure 2
Balance Scale Problem Attribute Structure

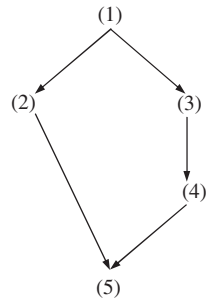
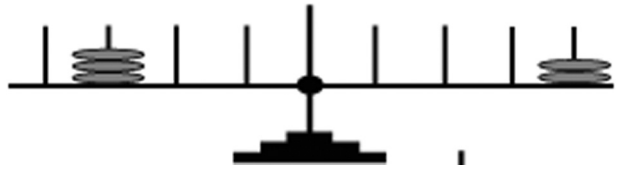


Figure 3
A Conflict-Weight Problem



model framework, the key (i.e., left) requires Attribute 2, the distractor right requires Attribute 3, and the distractor balance, Attribute 4. Children using Attributes 2 and 5 will give the correct response; children using Attribute 4 will respond with balance; children using Attribute 3 will respond with right; children using Attribute 1 will guess the answer to or muddle through this problem. By formulating the balance scale problems as MC-DINA model items, the incorrect responses can provide additional information that will allow attribute identification more efficiently.

In the educational measurement contexts, writing the distractors to correspond to specific latent classes is analogous to the works by Briggs, Alonzo, Schwab, and Wilson (2006) and Sadler (1998) except these works were from a unidimensional IRT perspective. In both works, the distractors can be placed onto the proficiency scale to allow examination of the relationship between proficiency level and propensity to choose specific options. The former utilizes construct maps also used in the Bear Assessment System (Wilson & Sloane, 2000) to create the options where mapping to proficiency levels is determined a priori. In comparison, the latter uses options based on popular alternative conceptions, and their locations on the scale are determined a posteriori. In this respect, the proposed framework is more akin to that of Briggs et al.

In this framework, distractors must be created to reflect the type of responses students who lack one or more of the required attributes for the correct response are likely to give. That is, the knowledge states represented by the distractors should be in the subset of the knowledge state that corresponds to the key. In the example above, option C, $1\frac{9}{12}$, was written to reflect

mastery of the attributes of borrowing from the whole and basic fraction subtraction but not that of reduction/simplification. By writing distractors in this manner, examinees from some latent classes are given higher propensities of choosing specific options.

Recall that the number of latent groups for each item using the DINA model is 2 (i.e., one group for examinees with all the required attributes and another for those who lack at least one of the prescribed attributes for the item). In contrast, the number of latent groups under the proposed framework is equal to $H_j^* + 1$, the number of coded options for item j plus one. That is, for item j, the original 2^K latent classes are classified into $H_j^* + 1$ latent groups. Thus, by coding some of the distractors, attribute patterns that do not meet the specification of the key can be further distinguished from one another, thereby affording the items additional diagnostic information. To the extent that the conjunctive property refers to the necessity of having all the required attributes, the property holds for all the coded options of an item. However, it is clear from the framework specification that, unlike the DINA model, this property does not always result in a single undifferentiated group for those who lack some of the required attributes for the correct response.

The MC-DINA Model

Coding the distractors in addition to the key represents only the first component of this framework. Without the appropriate tools (i.e., psychometric models), the potential diagnostic information in these options will not be fully realized. Hence, the complementary component of this framework presents a CDM that exploits the diagnostic information available in the distractors. Specifically, this article proposes a CDM based on the DINA model that is appropriate for MC items with cognitively based distractors and will be referred to hereon as the MC-DINA model.

For notational convenience, assume $H_j = H$, $\forall j$. As mentioned earlier, the H q-vectors of item j (some of which are null) divide the examinees into $H_j^* + 1$ groups. In addition, for the purposes of this article, only Q-vectors that allow attribute patterns to be uniquely classified into one of the $H_j^* + 1$ groups will be discussed. The general case, which allows examinees to be classified into more than one nonzero group and entails some modifications in notation and estimation, can be considered at a later time. In the fraction-subtraction example above, the Q-vectors representing three distractors, each requiring one of the first three attributes, will be excluded because examinees who have exactly two of the three required attributes will be classified under two latent groups.

Define $\mathbf{q}_{j0} = 0$. Examinee i's latent group classification with respect to item j, denoted by g_{ij}, is

$$g_{ij} = \arg\max_{h'} \left\{ \boldsymbol{\alpha}_i' \mathbf{q}_{jh'} | \boldsymbol{\alpha}_i' \mathbf{q}_{jh'} = \mathbf{q}_{jh'}' \mathbf{q}_{jh'} \right\} \tag{3}$$

for $h' = 0, \ldots, H$. Recall that the q-vector for noncoded options is also 0. It follows from equation (3) that α_i will be classified as $g_{ij} = 0$ if and only if the attribute vector does not meet the attribute specification of at least one of the coded options.

Some features of the MC-DINA model are illustrated using the fraction-subtraction example given earlier. Let A = 1, B = 2, C = 3, and D = 4. For this item, $H_j^* = 4$, representing the

number of coded options. For examinees who possess the first three attributes, $\alpha_i'\mathbf{q}_{jh'} = 1$ for $h = 1, 2, 3, 4$. Because q_{j4} is the q-vector with the largest h where $\alpha_i'\mathbf{q}_{jh'} = 1$, these examinees will be classified under latent Group 4. In contrast, for examinees who possess the first two of the three required attributes, $\alpha_i'\mathbf{q}_{jh'} = 1$ only for $h = 1, 3$. These examinees will be classified under latent Group 3. Finally, for examinees who possess none or only the first or the third of the three required attributes, $\alpha_i'\mathbf{q}_{jh'} = 0$ for $h = 1, 2, 3, 4$. Thus, given this set of options, these examinees will be classified under latent Group 0.

For the MC-DINA model, the probability that examinee i will choose option h of item j is given by

$$P_{jh}(\boldsymbol{\alpha}_i) = P(X_{ij} = h|\boldsymbol{\alpha}_i) = P(X_{ij} = h|g_{ij} = g) = P_j(h|g), \tag{4}$$

where $P_j\ (h|g)$ is defined as the probability of an examinee in group g choosing option h of item j and $g \in G_j$, which contains 0 and a subset of $\{1, 2, \ldots, H\}$. For a fixed g, $\sum_{h=1}^{H} P_j(h|g) = 1$. Therefore, the model has a total of $\sum_{j=1}^{J} H(H_j^* + 1)$ parameters, $\sum_{j=1}^{J} H_j^* + J$ of which are not free to vary.

Under a strict definition, the correct response of item j refers to the key, say, h_j^*. Only examinees in group $g_j = h_j^*$ are expected to choose this option, and they do so with the probability of $P_j\ (h_j^*|h_j^*)$. Although not expected to choose the correct options, examinees in other groups can arrive at this answer by guessing, and the probability that an examinee will guess is $P_j\ (h_j^*|g)$, for $g \neq h_j^*$. However, under a less stringent definition, a correct response can refer to any coded option, although it conditional on a specific latent group. That is, an option, say h, is considered the correct response with respect only to examinees in group h. Consequently, the correct response for examinees in various groups, except Group 0, refers to the different options. To minimize confusion, the correct response for group g (i.e., option g) will be referred to as the expected response for the group. Examinees in Group 0 are not expected to choose any particular option; in addition, they need not choose from the options with equal probability.

It should be noted that the MC-DINA model can still be used even if only the key is coded so long as the distractors are distinguished from each other (i.e., they are not all scored as 0). Under this condition, $P_j\ (h|g)$ for $g = 0, 1$ and $h = 1, \ldots, H$ can be computed. These probabilities can provide information regarding the differential attractiveness of the options to each group. If no distinctions are made between the distractors, the MC-DINA model is equivalent to the DINA model. However, because the correct answer is provided in the MC format, guessing in its literal sense is more likely to take place.

An Expectation-Maximization (EM) Algorithm for the MC-DINA Model

Parameter Estimation

As in traditional item response models (IRMs), inconsistent estimates of the item parameters may be obtained when maximization jointly involves the structural and incidental

parameters (Baker, 1992; Neyman & Scott, 1948). To avoid this problem, estimation of the item parameters can be based on the marginalized likelihood of the data.

Let the marginalized likelihood of the response vector of examinee i be

$$L(\mathbf{X}_i) = \sum_{l=1}^{L} L(\mathbf{X}_i|\boldsymbol{\alpha}_l)p(\boldsymbol{\alpha}_l), \tag{5}$$

where

$$L(\mathbf{X}_i|\boldsymbol{\alpha}_l) = \prod_{j=1}^{J}\prod_{h=1}^{H} P_{jh}(\boldsymbol{\alpha}_l)^{X_{ijh}}, \tag{6}$$

$p(\boldsymbol{\alpha}_l)$ is the prior probability of the attribute vector $\boldsymbol{\alpha}_l$, $L = 2^K$, and

$$X_{ijh} = \begin{cases} 1 & \text{if } X_{ij} = h \\ 0 & \text{otherwise} \end{cases}$$

For fixed l and j, $\sum_{h=1}^{H}(\boldsymbol{\alpha}_l) = 1$ and $\sum_{h=1}^{H} X_{ijh} = 1$. That is, the probability that an examinee with the attribute pattern α_l will select one of the H options is 1, and exactly one option will be selected. Therefore, equation (6) can be written as

$$L(\mathbf{X}_i|\boldsymbol{\alpha}_l) = \prod_{j=1}^{J}\left(\prod_{h=1}^{H-1} P_{jh}(\boldsymbol{\alpha}_l)^{X_{ijh}}\right)\left(1 - \sum_{h=1}^{H-1} P_{jh}(\boldsymbol{\alpha}_l)\right)^{1-\Sigma_{h=1}^{H-1}X_{ijh}}. \tag{7}$$

To obtain the marginalized maximum likelihood estimate (MMLE) of $P_j\,(h|g)$, maximize

$$l(\mathbf{X}) = log\prod_{i=1}^{I} L(\mathbf{X}_i) = \sum_{i=1}^{I} \log L(\mathbf{X}_i) \tag{8}$$

with respect to $P_j\,(h|g)$. The computational details of the algorithm are given in the appendix. It is shown that the estimate of $P_j\,(h|g)$ is given by

$$\hat{P}_j(h|g) = \frac{I_{j(h|g)}}{\sum_{h=1}^{H} I_{j(h|g)}}, \tag{9}$$

where $I_{j(h|g)}$ is the expected number of examinees in group g choosing option h of item j.

Step 1 of the algorithm starts with initial values for $P_j\,(h|g)$, $j = 1, \ldots, J$, $h = 1, \ldots, H-1$, and $g \in G_j$. In Step 2, the expected counts $I_{j(h|g)}$ are computed based on the current values of $P_j(h|g)$. Step 3 involves finding $\hat{P}_j(h|g)$ using equation (9). Finally, Steps 2 and 3 are repeated until convergence. The algorithm in this article was implemented using an empirical Bayes method (Carlin & Louis, 2000). Specifically, the prior distribution of the latent classes, which is initially uniform, is updated after each iteration based on the posterior distributions of the examinees.

Computing the Standard Errors (*SEs*)

Let $\mathbf{P}$ denote the vector of the MC-DINA model parameters. The information matrix of the estimator $\mathbf{P}$ is $\mathbf{I}(\mathbf{P}) = -E\{\partial^2 l(\mathbf{X})/\partial \mathbf{P}^2\}$. Let P and P' denote $P_j(h|g)$ and $P_{j'}(h'|g')$, respectively. The computational details in the appendix show that the second derivative of $l(\mathbf{X})$ with respect to P and P' can be written as follows:

$$-\sum_{i=1}^{I}\left[\sum_{l=1}^{L} p(\alpha_l|\mathbf{X}_i)\sum_{h=1}^{H-1}\left(\frac{X_{ijh}}{P_{jh}(\alpha_l)}-\frac{X_{ijH}}{P_{jH}(\alpha_l)}\right)\frac{\partial P_{jh}(\alpha_l)}{\partial P}\right]$$
$$\times\left[\sum_{l=1}^{L} p(\alpha_l|\mathbf{X}_i)\sum_{h=1}^{H-1}\left(\frac{X_{ijh}}{P_{jh}(\alpha_l)}-\frac{X_{ijH}}{P_{jH}(\alpha_l)}\right)\frac{\partial P_{jh}(\alpha_l)}{\partial P'}\right]. \quad (10)$$

$\mathbf{I}(\mathbf{P})$ is approximated by $\mathbf{I}(\hat{\mathbf{P}})$, which is computed using the estimated $\mathbf{P}$ and the realized values of $\mathbf{X}$, and $SE(\hat{\mathbf{P}})$ by the root of the diagonal element of $\mathbf{I}^{-1}(\hat{\mathbf{P}})$.

Simulation Study

Design

The simulation study involved 1,000 examinees drawn from a distribution where all attributes patterns are equally likely, 30 four-option MC items and 5 attributes. The correct options for the first, second, and last 10 items have one, two, and three required attributes, respectively. In addition, for this particular example, the attribute specifications for options of items with more than one required attribute have an exhaustive (hierarchical) linear structure. That is, the number of coded options is equal to the number of required attributes for the key, and their specifications can be represented by a series of subsets of required attributes. For example, if the key requires three attributes, then two distractors are coded with one distractor requiring two attributes and the other distractor one attribute. Moreover, if the attributes required in the one-, two-, and three-attribute options can be represented by the sets A_1, A_2, and A_3, then $A_1 \subset A_2 \subset A_3$. As stated earlier, all the distractors need not be coded, nor attribute specifications for coded distractors need form subsets of one another. The attribute specifications for the options in Table 2 form a hierarchical albeit nonlinear structure (i.e., for options B and C, one specification is not a subset of the other). For the options of this item to form an exhaustive linear specification, either option B or C must be replaced by a noncoded option.

The Q-matrices for the different coded options are combined in the modified Q-matrix given in Table 4. The entry in each cell indicates the number of times an attribute is specified in the options. For example, the modified Q-vector for Item 11, [2 1 0 0 0], indicates that the correct option requires α_1 and α_2, whereas the only coded distractor requires α_1. In the simulation study, the probability that an examinee in group g will choose option h of item j is given by

$$P_j(h|g) = \begin{cases} 0.25 & \text{if } g = 0 \\ 0.82 & \text{if } g > 0 \text{ and } g = h \\ 0.06 & \text{if } g > 0 \text{ and } g \neq h \end{cases}.$$

Table 4
Modified Q-Matrix for the Simulated Data

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

That is, for examinees whose attribute patterns do not meet the requirements of any of the coded options, choices are made at random (i.e., with equal probability). Examinees who meet the requisites of at least one of the coded options will choose the expected options 82% of the time and choose from the remaining options randomly.

One hundred data sets were generated, and parameter estimates and their *SE*s were obtained for each replication. In addition, the mean estimates, mean *SE* (root mean squared *SE*), and empirical *SE* (standard deviation of the estimates) across the replications were computed. Finally, to assess the relative efficiency of the proposed framework, the attribute classification rate using the MC-DINA model was compared with the classification rate using the traditional DINA model (i.e., analysis using dichotomized data). The MMLE algorithms for estimating the parameters of the DINA and the MC-DINA models via MMLE were implemented in Ox (Doornik, 2002). The console version of Ox can be downloaded free of charge, whereas the code developed for this article can be obtained by contacting the author. Using a Pentium-4 3 GHz processor and a convergence criterion of 0.001, each replication took an average of under 3 s to estimate.

Results

Item parameter estimation. Table 5 gives the mean, bias, and empirical *SE* across the 100 replications. These results represent the average across the parameters of similar items. For example, $P(\cdot|1)$ is the average across 30 items with options that require one attribute, whereas $P(\cdot|3)$ is the average across 10 items with options that requires three attributes. The results show that regardless of the parameters concerned (i.e., $P(h|0)$, $P(h|h)$, and $P(\not h|h)$) the algorithm produced very accurate estimates—The maximum mean bias across these estimates was only 0.001 in absolute terms. In addition, mean and

Table 5
Bias, Mean, and Empirical *SE* Across 30 Items (Negative Values in Parenthesis)

	True		*SE*	
Probability	$P(h\|g)$	Bias	Empirical	M
.25	$P(h\|0)$	.000	.019	.020
.82	$P(1\|1)$	.000	.024	.024
	$P(2\|2)$	.000	.031	.031
	$P(3\|3)$	(.001)	.037	.036
.06	$P(1\|1)$	.001	.015	.015
	$P(2\|2)$	.000	.020	.020
	$P(3\|3)$	.000	.022	.023

empirical *SE*s were very similar to each other indicating that the estimated *SE* faithfully reflected the true sampling variability. This table also shows that the precision of the estimate of the probability of choosing an expected or a nonexpected response varied as a function of the number of required attributes. Specifically, parameters that required more attributes had lower precision.

To further examine the impact of the number of required attributes on parameter estimation, the results in Table 5 were disaggregated according to item classification—Items with the same number of required attributes for the key were classified as the same item type. That is, Items 1-10, 11-20, and 21-30 were considered of the same types. The disaggregated results in Table 6 indicate the algorithm can provide accurate parameter and *SE* estimates across the different item types. In addition, for a fixed parameter $P(h|g)$, $g > 0$, although the estimates were the same, the variabilities of the estimates were not the same across the item types—the *SE*s of parameters for one-attribute items were always the smallest—whereas, except for sampling errors, the *SE*s of two-attribute items were less than or equal to those of three-attribute items. For example, for $P(\cdot|1)$, one-attribute items have the smallest *SE*. For the same parameters, the *SE*s of two- and three-attribute items are approximately equal (the actual difference between the two *SE*s was less than 0.001), but for $P(\cdot|2)$, the *SE* of the former was smaller than the latter. No such pattern can be observed for Group 0.

The pattern of results in Table 6 can be better understood by taking into account the design of the study, particularly the exhaustive and hierarchical linear structure of the option specifications. In general, a specific parameter (e.g., $P_j(h|g)$ with fixed h and g) can be estimated more precisely when the expected number of examinees in the group g with respect of item j, $I(g_j)$, is larger. Given the design of the study, if options of items j and j' require the same attributes, but the items are of different types, then $I(g_j) \leq I(g_{j'})$ when the key to item j' has a greater number of specified attributes.

To illustrate this point, assume that the key to items j, j', and j'' requires the attributes $\{\alpha_1\}$, $\{\alpha_1, \alpha_2\}$, and $\{\alpha_1, \alpha_2, \alpha_3\}$, respectively. In addition, assume that the specifications for the distractors can also be found from these sets. Item j'' divides the examinees into four latent groups, say, $g = 0, 1, 2, 3$. The number of examinees in these groups can be

Table 6
Bias, Mean, and Empirical *SE* by Item Classification[a]
(Negative Values in Parenthesis)

True		Item		*SE*	
Probability	$P(h\|g)$	Type	Bias	Empirical	*M*
.25	$P(h\|0)$	1	.000	.019	.020
		2	.000	.019	.020
		3	.000	.020	.020
.82	$P(1\|1)$	1	(.001)	.018	.018
		2	.001	.027	.026
		3	.000	.026	.026
	$P(2\|2)$	2	.000	.026	.026
		3	.001	.036	.036
	$P(3\|3)$	3	.001	.037	.036
.06	$P(1'\|1)$	1	.000	.012	.011
		2	.000	.017	.016
		3	(.001)	.016	.016
	$P(2'\|2)$	2	.000	.016	.016
		3	(.001)	.023	.023
	$P(3'\|3)$	3	(.001)	.022	.023

a. Classification based on maximum attributes required for the key.

denoted by $I(0)$, $I(1)$, $I(2)$, and $I(3)$. In comparison, item j' divides the examinees into three latent groups, and the numbers of examinees corresponding to these groups are $I(0)$, $I(1)$, and $I(2)+I(3)$. The two latent groups in item j have $I(0)$ and $I(1)+I(2)+I(3)$ examinees. As can be seen from these partitions, when $g=0$, $P(h|g)$ is always based on $I(0)$ regardless of the item type; when $g>0$, $P(\cdot|1)$ is based on $I(1)+I(2)+I(3)$ for item j and only $I(1)$ for items j' and j'', whereas $P(\cdot|2)$ is based on $I(2)+I(3)$ for item j' but $I(2)$ only for item j''. The differences in sample sizes in the latent groups account for the observed differences in the *SE*s of the parameter estimates. These results underscore the importance not only of the overall sample size but also the expected numbers of examinees in the latent groups in determining the precision of the estimates.

Attribute classification. Table 7 gives the accuracy of attribute classifications using the MC-DINA and DINA models. For the purposes of this article, an examinee was classified as having a mastery of attribute k if and only if the marginal probability for the attribute based on the examinee's posterior distribution is greater than or equal to 0.5. The attribute vector classification was a concatenation of the individual attribute classifications. The values in the table were computed by comparing the true and estimated classifications, and represent the percentage of correct classifications across the replications and examinees, and are presented by individual attribute and attribute pattern. In classifying the attributes individually, the DINA model was correct about 91% of the time. By taking into account the information in the distractors using the MC-DINA model, the classification

Table 7
Percentage of Correctly Classified Attributes

	Attribute					
Model	α_1	α_2	α_3	α_4	α_5	α
MC-DINA	97.06	97.53	97.23	97.93	97.40	89.71
DINA	90.35	91.92	92.37	90.32	90.68	69.58

Note: MC-DINA = multiple-choice deterministic-input, noisy "and" gate.

accuracy improved by 6%, on the average. The difference between the classification rates of the DINA and MC-DINA models was more stark (i.e., a 20% improvement) when classification accuracy was based on the entire attribute pattern. These results demonstrate that, by infusing the distractors with cognitive information and using an appropriate model to analyze the data, dramatic improvement in classification accuracy can be achieved over conventional tests and analyses.

Summary and Discussion

As the issue of school and system accountability becomes more prominent, the need for assessments that can inform classroom instruction and student learning becomes more imperative. Without a conscious effort to buck the trend, assessments designed only to audit student learning will continue to relegate to the margin assessments that can provide information deemed interpretative and diagnostic, and that can facilitate learning. This article proposes a framework that allows tests using a MC format to be more diagnostically informative. The first component of the framework prescribes a method of constructing MC options such that the distractors, in addition to the key, become a source of diagnostic information. The complementary component of the framework proposes the MC-DINA model as a mechanism to synthesize and optimally utilize information that can be found from various sources.

The simulation study shows that the EM algorithm developed in this article provides accurate estimates of the MC-DINA model parameters. In addition, the *SE*s of the estimates faithfully reflect the sampling variability of the parameter estimates. Compared to the traditional DINA model, the MC-DINA model provided correct classification rates that are dramatically better, particularly when classification of attribute vectors are involved. From a psychometric point of view, these results unequivocally indicate the feasibility of the framework and the benefit that can be gained from using it.

Inherent in cognitive diagnosis modeling is the simultaneous measurement of multiple attributes. For these attributes to be measured with sufficient reliability, the CDM framework requires assessments that are at least of moderate test lengths (e.g., 15 or 20 items). In addition, because the MC-DINA model evaluates the actual responses to the items, not simply their accuracy, a contingency table that cross-classifies the responses would be very sparse for the conventional likelihood ratio test to be reliable in most, if not all, testing situations. Thus, no general goodness-of-fit test is available for the MC-DINA model

unless the number of examinees is extremely large. This is the same problem encountered by Thissen and Steinberg (1997) in implementing a unidimensional IRM for MC data. Although direct global fit of the MC-DINA model is practically not possible, fit at the item level can be investigated. With a moderately large sample size (i.e., sample size sufficiently large for calibration purposes), the chi-square goodness-of-fit test can be used to compare the model-based expected responses against the observed responses to determine whether the lower-level marginal distributions (i.e., single items and item pairs) can be fitted adequately. The item-level tests can be used to identify nonfitting items. In addition, these tests, as necessary conditions, can provide an indirect measure of the model's global fit.

Promising as it is, it should be noted the proposed framework represents only the psychometric aspect of developing and implementing cognitively based MC assessments. The complete process is a multidisciplinary endeavor that requires collaboration between experts from various fields, such as learning science, cognitive science, subject domains, didactics, and psychometrics. Given a domain, the definition and grain size of the attributes that need to be included, the type of tasks that best measure them, and the misconceptions and difficulties typically encountered in the domain are some of the issues beyond the expertise of most psychometricians. The inferences culled from these types of assessments will only be practically useful to the extent that contributions from relevant fields are solicited and incorporated in the process of test development and analysis.

The work in this article represents only the initial steps in understanding how MC assessments can be made more useful for diagnostic purposes and serves as an impetus for additional work to be done in this area. One research direction could involve extending the generality of the results found in here. The findings in this article are based on a specific design of the simulation study. Whether these results generalize to other conditions (e.g., varying sample size, number and quality of items and coded options, and number of attributes) remains an open question. Results from this line of research can offer some practically useful guidelines regarding the sample size requirements for tests with various specifications to obtain reasonable estimates. It would also be interesting to investigate the impact of the type of attribute specification on parameter estimation and attribute classification. In particular, attribute specifications that are neither exhaustive, linear, nor hierarchical can be considered.

As the results of the simulation study indicate, precision of the parameter estimates is a function not only of the overall sample size I but also of the expected number of examinees in a latent group $I_{j(\cdot|g)}$. As more options of an item are coded, I is partitioned into more groups. This increases the likelihood of encountering a small $I_{j(\cdot|g)}$, which can have dire implications on the stability of the estimates. Thus, in deciding the number of options to be coded, one needs to carefully weigh the trade-offs between the additional diagnostic information and the possibly unstable estimates resulting from coding more options.

From its specification, the MC-DINA model has a large number of parameters to be estimated (i.e., $(H_j^* + 1)(H - 1)$ for item j) and would require a relatively large sample size that may not be available in some applications. In these situations, some simplifying assumptions about the model can be made to reduce its computational requirement. One mild assumption that can be invoked presupposes that nonexpected responses are chosen equiprobably. This assumption is reasonable when the probability of choosing the expected responses is moderately high and when invoked reduces the number parameters to be

estimated to $H_j^* + H - 1$. A stronger assumption posits that examinees in group $g_j = 0$ choose each of the options with probability $1/H$. When both assumptions are invoked simultaneously, there are only H_j^* parameters to be estimated. However, the extent of the impact of invoking these assumptions on attribute classification has not been established. This is a research area that could provide much needed guidelines in using the MC-DINA model when the sample size is small.

Finally, as noted earlier, the MC-DINA model can be applied to traditional MC assessments. Assuming that an appropriate Q-matrix can be constructed for the keys and that the sample size is sufficiently large, it would be interesting to investigate the type of insights one can obtain when the distractors are kept distinct, and how these insights compare to those derived when all the distractors are treated only as incorrect responses.

Appendix

Computational Details of an EM Algorithm for the MC-DINA Model

Parameter Estimation

Let

$$\frac{\partial l(\mathbf{X})}{\partial P_j(\cdot|\cdot)} = \sum_{i=1}^{I} \frac{1}{L(\mathbf{X}_i)} \frac{\partial L(\mathbf{X}_i)}{\partial P_j(\cdot|\cdot)} = \sum_{i=1}^{I} \frac{1}{L(\mathbf{X}_i)} \sum_{l=i}^{L} p(\boldsymbol{\alpha}_l) \frac{\partial L(\mathbf{X}_i|\boldsymbol{\alpha}_l)}{\partial P_j(\cdot|\cdot)}. \tag{11}$$

Now,

$$\begin{aligned} \frac{\partial L(\mathbf{X}_i|\boldsymbol{\alpha}_l)}{\partial P_j(\cdot|\cdot)} = & \left[\prod_{j' \neq j} \left(\prod_{h=1}^{H-1} P_{j'h}(\boldsymbol{\alpha}_l)^{X_{ij'h}} \right) \left(1 - \sum_{h=1}^{H-1} P_{j'h}(\boldsymbol{\alpha}_l) \right)^{1 - \sum_{h=1}^{H-1} X_{ij'h}} \right] \\ & \times \frac{\partial \left(\prod_{h=1}^{H-1} P_{jh}(\boldsymbol{\alpha}_l)^{X_{ijh}} \right) \left(1 - \sum_{h=1}^{H-1} P_{jh}(\boldsymbol{\alpha}_l) \right)^{1 - \sum_{h=1}^{H=1} X_{ijh}}}{\partial P_j(\cdot|\cdot)}. \end{aligned} \tag{12}$$

The derivative in the right-hand side (RHS) of equation (12) is equal to the following:

$$\begin{aligned} & \frac{\partial \left(P_{j1}(\boldsymbol{\alpha}_l)^{X_{ij1}} \cdots P_{j1}(\boldsymbol{\alpha}_l)^{X_{ij(H-1)}} \right) \left(1 - \sum_{h=1}^{H-1} P_{jh}(\boldsymbol{\alpha}_l) \right)^{1 - \sum_{h=1}^{H-1} X_{ijh}}}{\partial P_j(\cdot|\cdot)} \\ & = \sum_{h=1}^{H-1} \left[\prod_{h' \neq h} P_{jh'}(\boldsymbol{\alpha}_l)^{X_{ijh'}} \cdot X_{ijh} \cdot P_{jh}(\boldsymbol{\alpha}_l)^{X_{ijh}-1} \frac{\partial P_{jh}(\boldsymbol{\alpha}_l)}{\partial P_j(\cdot|\cdot)} \left(1 - \sum_{h=1}^{H-1} P_{jh}(\boldsymbol{\alpha}_l) \right)^{1 - \sum_{h=1}^{H-1} X_{ijh}} \right] \\ & \quad + \prod_{h=1}^{H-1} P_{jh}(\boldsymbol{\alpha}_l)^{X_{ijh}} \left(1 - \sum_{h=1}^{H-1} X_{ijh} \right) \left(1 - \sum_{h=1}^{H-1} P_{jh}(\boldsymbol{\alpha}_l) \right)^{1 - \sum_{h=1}^{H-1} X_{ijh} - 1} \frac{-\partial \sum_{h=1}^{H-1} P_{jh}(\alpha_l)}{\partial P_j(\cdot|\cdot)} \end{aligned}$$

$$
\begin{aligned}
&= \sum_{h=1}^{H-1}\left[\prod_{h'=1}^{H-1} P_{jh'}(\boldsymbol{\alpha}_l)^{X_{ijh'}}\left(1-\sum_{h'=1}^{H-1} P_{jh}(\boldsymbol{\alpha}_l)\right)^{1-\sum_{h'=1}^{H-1} X_{ijh'}} \frac{X_{ijh}}{P_{jh}(\boldsymbol{\alpha}_l)}\frac{\partial P_{jh}(\boldsymbol{\alpha}_l)}{\partial P_j(\cdot\,|\,\cdot)}\right] \\
&\quad - \prod_{h=1}^{H-1} P_{jh}(\boldsymbol{\alpha}_l)^{X_{ijh}}\left(1-\sum_{h=1}^{H-1} P_{jh}(\boldsymbol{\alpha}_l)\right)^{1-\sum_{h=1}^{H-1} X_{ijh}} \frac{1-\sum_{h=1}^{H-1} X_{ijh}}{1-\sum_{h=1}^{H-1} P_{jh}(\boldsymbol{\alpha}_l)} \sum_{h=1}^{H-1}\frac{\partial P_{jh}(\boldsymbol{\alpha}_l)}{\partial P_j(\cdot\,|\,\cdot)} \\
&= \left[\prod_{h=1}^{H-1} P_{jh}(\boldsymbol{\alpha}_l)^{X_{ijh}}\left(1-\sum_{h=1}^{H-1} P_{jh}(\boldsymbol{\alpha}_l)\right)^{1-\sum_{h=1}^{H-1} X_{ijh}}\right] \\
&\quad \times \left[\sum_{h=1}^{H-1}\frac{X_{ijh}}{P_{jh}(\boldsymbol{\alpha}_l)}\frac{\partial P_{jh}(\boldsymbol{\alpha}_l)}{\partial P_j(\cdot\,|\,\cdot)} - \frac{1-\sum_{h=1}^{H-1} X_{ijh}}{1-\sum_{h=1}^{H-1} P_{jh}(\boldsymbol{\alpha}_l)}\sum_{h=1}^{H-1}\frac{\partial P_{jh}(\boldsymbol{\alpha}_l)}{\partial P_j(\cdot\,|\,\cdot)}\right].
\end{aligned}
\tag{13}
$$

The second factor of equation (13) can be written as follows:

$$
\begin{aligned}
&\sum_{h=1}^{H-1}\frac{X_{ijh}}{P_{jh}(\boldsymbol{\alpha}_l)}\frac{\partial P_{jh}(\boldsymbol{\alpha}_l)}{\partial P_j(\cdot\,|\,\cdot)} - \sum_{h=1}^{H-1}\frac{1-\sum_{h=1}^{H-1} X_{ijh}}{1-\sum_{h=1}^{H-1} P_{jh}(\boldsymbol{\alpha}_l)}\frac{\partial P_{jh}(\boldsymbol{\alpha}_l)}{\partial P_j(\cdot\,|\,\cdot)} \\
&\quad = \sum_{h=1}^{H-1}\left[\frac{X_{ijh}}{P_{jh}(\boldsymbol{\alpha}_l)} - \frac{1-\sum_{h=1}^{H-1} X_{ijh}}{1-\sum_{h=1}^{H-1} P_{jh}(\boldsymbol{\alpha}_l)}\right]\frac{\partial P_{jh}(\boldsymbol{\alpha}_l)}{\partial P_j(\cdot\,|\,\cdot)}.
\end{aligned}
\tag{14}
$$

Substituting equation (14) in to the second factor of equation (13) and equation (13) in to the derivative in the RHS of equation (12),

$$
\begin{aligned}
&\frac{\partial L(\mathbf{X}_i|\boldsymbol{\alpha}_l)}{\partial P_j(\cdot\,|\,\cdot)} \\
&\quad = \left[\prod_{j=1}^{J}\left(\prod_{h=1}^{H-1} P_{jh}(\boldsymbol{\alpha}_l)^{X_{ijh}}\right)\left(1-\sum_{h=1}^{H-1} P_{jh}(\boldsymbol{\alpha}_l)\right)^{1-\sum_{h=1}^{H-1} X_{ijh}}\right] \\
&\quad \times \sum_{h=1}^{H-1}\left[\frac{X_{ijh}}{P_{jh}(\boldsymbol{\alpha}_l)} - \frac{1-\sum_{h=1}^{H-1} X_{ijh}}{1-\sum_{h=1}^{H-1} P_{jh}(\boldsymbol{\alpha}_l)}\right]\frac{\partial P_{jh}(\boldsymbol{\alpha}_l)}{\partial P_j(\cdot\,|\,\cdot)} \\
&\quad = L(\mathbf{X}_i|\boldsymbol{\alpha}_l)\sum_{h=1}^{H-1}\left[\frac{X_{ijh}}{P_{jh}(\boldsymbol{\alpha}_l)} - \frac{1-\sum_{h=1}^{H-1} X_{ijh}}{1-\sum_{h=1}^{H-1} P_{jh}(\boldsymbol{\alpha}_l)}\right]\frac{\partial P_{jh}(\boldsymbol{\alpha}_l)}{\partial P_j(\cdot\,|\,\cdot)}.
\end{aligned}
\tag{15}
$$

Therefore, equation (11) can be written as follows:

$$
\begin{aligned}
&\sum_{i=1}^{I}\frac{1}{L(\mathbf{X}_i)}\sum_{l=i}^{L} p(\boldsymbol{\alpha}_l)L(\mathbf{X}_i|\boldsymbol{\alpha}_l)\sum_{h=1}^{H-1}\left[\frac{X_{ijh}}{P_{jh}(\boldsymbol{\alpha}_l)} - \frac{1-\sum_{h=1}^{H-1} X_{ijh}}{1-\sum_{h=1}^{H-1} P_{jh}(\boldsymbol{\alpha}_l)}\right]\frac{\partial P_{jh}(\boldsymbol{\alpha}_l)}{\partial P_j(\cdot\,|\,\cdot)} \\
&= \sum_{l=i}^{L}\sum_{i=1}^{I}\frac{L(\mathbf{X}_i|\boldsymbol{\alpha}_l)p(\boldsymbol{\alpha}_l)}{L(\mathbf{X}_i)}\sum_{h=1}^{H-1}\left[\frac{X_{ijh}}{P_{jh}(\boldsymbol{\alpha}_l)} - \frac{1-\sum_{h=1}^{H-1} X_{ijh}}{1-\sum_{h=1}^{H-1} P_{jh}(\boldsymbol{\alpha}_l)}\right]\frac{\partial P_{jh}(\boldsymbol{\alpha}_l)}{\partial P_j(\cdot\,|\,\cdot)} \\
&= \sum_{l=i}^{L}\sum_{i=1}^{I} p(\boldsymbol{\alpha}_l|\mathbf{X}_i)\sum_{h=1}^{H-1}\left[\frac{X_{ijh}}{P_{jh}(\boldsymbol{\alpha}_l)} - \frac{1-\sum_{h=1}^{H-1} X_{ijh}}{1-\sum_{h=1}^{H-1} P_{jh}(\boldsymbol{\alpha}_l)}\right]\frac{\partial P_{jh}(\boldsymbol{\alpha}_l)}{\partial P_j(\cdot\,|\,\cdot)},
\end{aligned}
\tag{16}
$$

where $p(\boldsymbol{\alpha}_l|\mathbf{X}_i)$ is the posterior probability of the attribute pattern $\boldsymbol{\alpha}_l$ given the response vector $\mathbf{X}_i$. By rearranging the summations further and distributing the posterior probabilities,

$$\begin{aligned}\frac{\partial l(\mathbf{X})}{\partial P_j(\cdot|\cdot)} &= \sum_{l=i}^{L}\sum_{h=1}^{H-1}\frac{\partial P_{jh}(\boldsymbol{\alpha}_l)}{\partial P_j(\cdot|\cdot)}\sum_{i=1}^{I}p(\boldsymbol{\alpha}_l|\mathbf{X}_i)\left[\frac{X_{ijh}}{P_{jh}(\boldsymbol{\alpha}_l)}-\frac{1-\sum_{h=1}^{H-1}X_{ijh}}{1-\sum_{h=1}^{H-1}P_{jh}(\boldsymbol{\alpha}_l)}\right]\\ &= \sum_{l=i}^{L}\sum_{h=1}^{H-1}\frac{\partial P_{jh}(\boldsymbol{\alpha}_l)}{\partial P_j(\cdot|\cdot)}\left[\frac{1}{P_{jh}(\boldsymbol{\alpha}_l)}\sum_{i=1}^{I}p(\boldsymbol{\alpha}_l|\mathbf{X}_i)X_{ijh}\right.\\ &\quad\left.-\frac{1}{1-\sum_{h=1}^{H-1}P_{jh}(\boldsymbol{\alpha}_l)}\sum_{i=1}^{I}p(\boldsymbol{\alpha}_l|\mathbf{X}_i)\left(1-\sum_{h=1}^{H-1}X_{ijh}\right)\right].\end{aligned} \tag{17}$$

Let l_g denote $l\in\{\boldsymbol{\alpha}_l : g_{lj}=g\}$. For all attribute patterns in the same group, $P_{jh}(\boldsymbol{\alpha}_{l_g})=P_j(h|g)$. This allows (17) to be written as follows:

$$\begin{aligned}&\sum_{\forall g\in G_j}\sum_{h=1}^{H-1}\frac{\partial P_j(h|g)}{\partial P_j(\cdot|\cdot)}\left[\frac{1}{P_j(h|g)}\sum_{i=1}^{I}p(g|\mathbf{X}_i)X_{ijh}\right.\\ &\quad\left.-\frac{1}{1-\sum_{h=1}^{H-1}P_j(h|g)}\sum_{i=1}^{I}p(g|\mathbf{X}_i)\left(1-\sum_{h=1}^{H-1}X_{ijh}\right)\right],\end{aligned} \tag{18}$$

where $p(g|\mathbf{X}_i)$ is the posterior probability of the examinee i being in group g. Because

$$\frac{\partial P_j(h'|g')}{\partial P_j(h|g)}=\begin{cases}1 & \text{if } g=g' \text{ and } h=h'\\ 0 & \text{otherwise}\end{cases},$$

equation (18) can be evaluated one group and one option at a time, and for group g and h it reduces to the following:

$$\begin{aligned}&\frac{1}{P_j(h|g)}\sum_{i=1}^{I}p(g|\mathbf{X}_i)X_{ijh}-\frac{1}{1-\sum_{h=1}^{H-1}P_j(h|g)}\sum_{i=1}^{I}p(g|\mathbf{X}_i)\left(1-\sum_{h=1}^{H-1}X_{ijh}\right)\\ &=\frac{1}{P_j(h|g)}I_{j(h|g)}-\frac{1}{1-\sum_{h=1}^{H-1}P_j(h|g)}I_{j(H|g)}\\ &=\frac{I_{j(h|g)}}{P_j(h|g)}-\frac{I_{j(H|g)}}{1-\sum_{h=1}^{H-1}P_j(h|g)},\end{aligned} \tag{19}$$

where $\sum_{i=1}^{I}p(g|\mathbf{X}_i)X_{ijh}=I_{j(h|g)}$. Because $\partial l(\mathbf{X})$ with respect to specific items and groups across the different options share common parameters, the solution to the gradient

$$\nabla[l(\mathbf{X})]_{jg}=\left(\frac{\partial l(\mathbf{X})}{\partial P_j(1|g)},\ldots,\frac{\partial l(\mathbf{X})}{\partial P_j(H-1|g)}\right), \tag{20}$$

(i.e., $\nabla[l(\mathbf{X})]_{jg}=0$) needs to be obtained.

Finding the solution to equation (20) is identical to finding the maximum likelihood estimate of the parameters of a multinomial distribution except that, instead of the observed counts, the expected counts for each option of an item given in a particular group are involved. Therefore,

$$\hat{P}_j(h|g) = \frac{I_{j(h|g)}}{\sum_{h=1}^{H} I_{j(h|g)}}. \tag{21}$$

Computing the *SE*s

With $P = P_j(h|g)$ and $P' = P'_j(h'|g')$,

$$\frac{\partial l(\mathbf{X})}{\partial P} = \sum_{i=1}^{I} \frac{1}{L(\mathbf{X}_i)} \frac{\partial L(\mathbf{X}_i)}{\partial P}, \tag{22}$$

$$\frac{\partial L(\mathbf{X}_i)}{\partial P} = \sum_{l=1}^{L} p(\alpha_l) \frac{\partial L(\mathbf{X}_i|\alpha_l)}{\partial P}, \tag{23}$$

$$\frac{\partial L(\mathbf{X}_i|\boldsymbol{\alpha}_l)}{\partial P} = L(\mathbf{X}_i|\alpha_l) \sum_{h=1}^{H-1} \left(\frac{X_{ijh}}{P_{jh}(\alpha_l)} - \frac{X_{ijH}}{P_{jH}(\alpha_l)} \right) \frac{\partial P_{jh}(\alpha_l)}{\partial P}. \tag{24}$$

The second derivative of the marginalized log likelihood with respect to P and P' is equal to

$$\frac{\partial l(\mathbf{X})}{\partial P \partial P'} = \sum_{i=1}^{I} \frac{\partial l(\mathbf{X}_i)}{\partial P \partial P'} = \sum_{i=1}^{I} \left[L^{-1}(\mathbf{X}_i) \frac{\partial^2 L(\mathbf{X}_i)}{\partial P \partial P'} + \frac{\partial L(\mathbf{X}_i)}{\partial P} \frac{\partial L^{-1}(\mathbf{X}_i)}{\partial P'} \right]. \tag{25}$$

The expected value of the first term of equation (25) disappears. Therefore, $\partial^2 l(\mathbf{X}) / \partial P \partial P'$ reduces to

$$\begin{aligned}
&\sum_{i=1}^{I} \frac{\partial L(\mathbf{X}_i)}{\partial P} \frac{\partial L^{-1}(\mathbf{X}_i)}{\partial P'} \\
&= \sum_{i=1}^{I} \left[\sum_{l=1}^{L} p(\alpha_l) L(\mathbf{X}_i|\alpha_l) \sum_{h=1}^{H-1} \left(\frac{X_{ijh}}{P_{jh}(\alpha_l)} - \frac{X_{ijH}}{P_{jH}(\alpha_l)} \right) \frac{\partial P_{jh}(\alpha_l)}{\partial P} \right] \\
&\quad \times \left[-\frac{1}{L^2(\mathbf{X}_i)} \sum_{l=1}^{L} p(\alpha_l) L(\mathbf{X}_i|\alpha_l) \sum_{h=1}^{H-1} \left(\frac{X_{ijh}}{P_{jh}(\alpha_l)} - \frac{X_{ijH}}{P_{jH}(\alpha_l)} \right) \frac{\partial P_{jh}(\alpha_l)}{\partial P'} \right] \\
&= -\sum_{i=1}^{I} \left[\sum_{l=1}^{L} \frac{p(\alpha_l) L(\mathbf{X}_i|\alpha_l)}{L^2(\mathbf{X}_i)} \sum_{h=1}^{H-1} \left(\frac{X_{ijh}}{P_{jh}(\alpha_l)} - \frac{X_{ijH}}{P_{jH}(\alpha_l)} \right) \frac{\partial P_{jh}(\alpha_l)}{\partial P} \right] \\
&\quad \times \left[\sum_{l=1}^{L} \frac{p(\alpha_l) L(\mathbf{X}_i|\alpha_l)}{L^2(\mathbf{X}_i)} \sum_{h=1}^{H-1} \left(\frac{X_{ijh}}{P_{jh}(\alpha_l)} - \frac{X_{ijH}}{P_{jH}(\alpha_l)} \right) \frac{\partial P_{jh}(\alpha_l)}{\partial P'} \right] \\
&= -\sum_{i=1}^{I} \left[\sum_{l=1}^{L} p(\alpha_l|\mathbf{X}_i) \sum_{h=1}^{H-1} \left(\frac{X_{ijh}}{P_{jh}(\alpha_l)} - \frac{X_{ijH}}{P_{jH}(\alpha_l)} \right) \frac{\partial P_{jh}(\alpha_l)}{\partial P} \right] \\
&\quad \times \left[\sum_{l=1}^{L} p(\alpha_l|\mathbf{X}_i) \sum_{h=1}^{H-1} \left(\frac{X_{ijh}}{P_{jh}(\alpha_l)} - \frac{X_{ijH}}{P_{jH}(\alpha_l)} \right) \frac{\partial P_{jh}(\alpha_l)}{\partial P'} \right].
\end{aligned} \tag{26}$$

References

Baker, F. B. (1992). *Item response theory: Parameter estimation techniques*. New York: Dekker.

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.

Briggs, D., Alonzo, A., Schwab, C., & Wilson, M. (2006). Diagnostic assessment with ordered multiple-choice items. *Educational Assessment, 11*, 33-63.

Carlin, B. P., & Louis, T. A. (2000). *Bayes and empirical Bayes methods for data analysis*. New York: Chapman & Hall.

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

de la Torre, J., & Patz, R. (2005). Making the most of what we have: A practical application of multidimensional item response theory in test scoring. *Journal of Educational and Behavioral Statistics, 30*, 295-311.

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

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

Embretson, S. (1984). A general latent trait model for response processes. *Psychometrika, 49*, 175-186.

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

Haertel, E. H., & Wiley, D. E. (1993). Representations of ability structures: Implications for testing. In N. Frederiksen, R. J. Mislevy, & I. Bejar (Eds.), *Test theory for a new generation of tests* (pp. 359-384). Hillsdale, NJ: Erlbaum.

Halford, G. S., Andrews, G., Dalton, C., Boag, C., & Zielinski, T. (2002). Young children's performance on the balance scale: The influence of relational complexity. *Journal of Experimental Child Psychology, 81*, 417-445.

Jansen, B. R. J., & van der Maas, H. L. J. (2002). The development of children's rule use on the balance scale task. *Journal of Experimental Child Psychology, 81*, 383-416.

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.

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

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, NJ: Erlbaum.

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

National Research Council. (2003). *Strategic education research partnership/Committee on a strategic education research partnership; M. S. Donovan, A. K. Wigdor, and C. E. Snow, Editors*. Washington, DC: National Academies Press.

Neyman, J., & Scott, E. L. (1948). Consistent estimates based on partially consistent observations. *Econometrika, 16*, 661-679.

Nitko, A. J. (2001). *Educational assessment of students* (3rd ed.). Columbus, OH: Merrill Prentice Hall.

Osterlind, S. J. (1998). *Constructing test items: Multiple choice, constructed-response, performance and other formats* (2nd ed.). Boston: Kluwer Academic.

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.

Sadler, P. M. (1998). Psychometric models of student conceptions in science: Reconciling qualitative studies and distractor-driven assessment instruments. *Journal of Research in Science Teaching, 35*, 265-296.

Shultz, T. R., Mareschal, D., & Schmidt, W. C. (1994). Modeling cognitive development on balance scale phenomena. *Machine Learning, 16*, 57-86.

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

Siegler, R. S. (1981). Developmental sequences within and between concepts. *Monographs of the Society for Research in Child Development, 46* (2, Serial No. 189).

Stiggins. R. J. (2002). Assessment crisis: The absence of assessment for learning. *Phi Delta Kappan, 83*, 758-765.

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.

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, & Safto, M. (Eds.), *Monitoring skills and knowledge acquisition* (pp. 453-488). 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.

Thissen, D., & Steinberg, L. (1997). A response model for multiple-choice items. In W. van der Linden & R. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 51-65). New York: Springer.

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

Wainer, H., Vevea, J. L., Camacho, F., Reeve, B., Rosa, K., Nelson, L., et al. (2001). Augmented scores—"Borrowing strength" to compute scores based on small number of items. In D. Thissen & H. Wainer (Eds.), *Test scoring* (pp. 343-387). Mahwah, NJ: Lawrence Erlbaum.

Wiggins, G. (1998). *Educative assessment: Designing assessment to inform and improve performance*. San Francisco: Jossey-Bass.

Wilson, M., & Sloane, K. (2000). From principles to practice: An embedded assessment system. *Applied Measurement in Education, 13*, 181-208.