PSYCHOMETRIKA—VOL. 73, NO. 4, 595–624
DECEMBER 2008
DOI: 10.1007/S11336-008-9063-2

MODEL EVALUATION AND MULTIPLE STRATEGIES IN COGNITIVE DIAGNOSIS: AN ANALYSIS OF FRACTION SUBTRACTION DATA

JIMMY DE LA TORRE

DEPARTMENT OF EDUCATIONAL PSYCHOLOGY, RUTGERS, THE STATE UNIVERSITY OF NEW JERSEY

JEFFREY A. DOUGLAS

DEPARTMENT OF STATISTICS, UNIVERSITY OF ILLINOIS

This paper studies three models for cognitive diagnosis, each illustrated with an application to fraction subtraction data. The objective of each of these models is to classify examinees according to their mastery of skills assumed to be required for fraction subtraction. We consider the DINA model, the NIDA model, and a new model that extends the DINA model to allow for multiple strategies of problem solving. For each of these models the joint distribution of the indicators of skill mastery is modeled using a single continuous higher-order latent trait, to explain the dependence in the mastery of distinct skills. This approach stems from viewing the skills as the specific states of knowledge required for exam performance, and viewing these skills as arising from a broadly defined latent trait resembling the θ of item response models. We discuss several techniques for comparing models and assessing goodness of fit. We then implement these methods using the fraction subtraction data with the aim of selecting the best of the three models for this application. We employ Markov chain Monte Carlo algorithms to fit the models, and we present simulation results to examine the performance of these algorithms.

Key words: cognitive diagnosis, item response theory, latent class model, Markov chain Monte Carlo, goodness-of-fit.

1. Introduction

Latent variable models for cognitive diagnosis have been developed with the aim of diagnosing the presence or absence of multiple fine-grained skills required for solving problems on an examination. In the literature, the presence of a skill is often synonymous with "mastery" of the skill and the absence of a skill is referred to as "nonmastery". One motivation for fitting these multiple classification latent class models is that they may provide more diagnostic value than lower-dimensional item response models, and may lead to more efficient remediation. In these models mastery of particular skills can be represented by a vector of binary latent variables, indicating presence or absence of each element of a set of skills under diagnosis. A generic term for a skill or knowledge state is an "attribute", and we use this terminology. In this paper we consider several cognitive diagnosis models for analyzing fraction subtraction data. One of the models acknowledges the possibility that different strategies might be used to solve a problem. Different strategies might require different attributes to successfully respond, and we propose a model that addresses this situation. Because accurate classification of the attribute vector would be of interest, and because the probability of making a correct classification is likely to depend on the fit of the model, we also focus on methods for comparing the goodness of fit of these models.

The work reported here was performed under the auspices of the External Diagnostic Research Team funded by Educational Testing Service. Views expressed in this paper does not necessarily represent the views of Educational Testing Service.

Requests for reprints should be sent to Jimmy de la Torre, Graduate School of Education, Rutgers, The State University of New Jersey, 10 Seminary Place, New Brunswick, NJ 08901, USA. E-mail: j.delatorre@rutgers.edu

FIGURE 1.
Solving a fraction subtraction problem using Strategy A.

	Attribute required
$4\frac{4}{12} - 2\frac{7}{12}$	
$= 2\frac{4}{12} - \frac{7}{12}$	(3) separating whole number from fraction
$= 1\frac{16}{12} - \frac{7}{12}$	(4) borrowing one from whole number to fraction
$= 1\frac{9}{12}$	(1) performing basic fraction subtraction operation
$= 1\frac{3}{4}$	(2) simplifying/reducing

FIGURE 2.
Solving a fraction subtraction problem using Strategy B.

	Attribute required
$4\frac{4}{12} - 2\frac{7}{12}$	
$= \frac{52}{12} - \frac{31}{12}$	(6) converting mixed number to fraction
$= \frac{21}{12}$	(1) performing basic fraction subtraction operation
$= 1\frac{9}{12} = 1\frac{3}{4}$	(2) simplifying/reducing

The main objectives of this paper are to introduce a model for multiple strategies and consider ways to compare and contrast models for cognitive diagnosis based on competing assumptions. To illustrate these ideas we provide a thorough data analysis of fraction subtraction data. The goodness-of-fit measures we consider will apply to various cognitive diagnosis model, not just those used this paper.

The data we analyze include responses of 2144 examinees to 15 fraction subtraction items, and are a subset of the original data described by Tatsuoka (1990), and recently analyzed by Tatsuoka (2002) and de la Torre and Douglas (2004). Mislevy (1996) analyzed similar data using multiple strategies, and we define the attributes required for fraction subtraction as in his paper, albeit with some minor modifications. These attributes are defined as follows: (1) performing basic fraction subtraction operation; (2) simplifying/reducing; (3) separating whole number from fraction; (4) borrowing one from whole number to fraction; (5) converting whole number to fraction; (6) converting mixed number to fraction; and (7) column borrowing in subtraction.

According to Mislevy (1996), two alternative strategies each requiring five attributes can be culled from these attributes. One strategy, Strategy A, requires examinees to perform fraction subtraction with mixed numbers and involves attributes 1, 2, 3, 4, and 5. The other strategy, Strategy B, requires subtraction of fractions where mixed numbers are first changed to improper fractions and involves attributes 1, 2, 5, 6, and 7. For example, the correct answer for $4\frac{4}{12} - 2\frac{7}{12}$ can be obtained in two ways. Using the first strategy, which requires attributes 1, 2, 3 and 4 for the problem, arriving at the correct answer would involve the steps given in Fig. 1. Alternatively, the second strategy, which requires attributes 1, 2, and 6 for this problem, involves the steps outlined in Fig. 2.

Let $\boldsymbol{Y}$ denote a vector of binary item responses for the J items of an examination. Consistent with traditional latent variable models in psychometrics, the components of $\boldsymbol{Y}$ are modeled as statistically independent given the latent variable, which is an attribute vector $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_K)'$. The kth element, α_k, of $\boldsymbol{\alpha}$ is a binary indicator of an examinee's classification with regard to the

kth attribute. For instance, in the case of fraction subtraction α_k might denote mastery of converting a whole number to a fraction. Models for item responses in cognitive diagnosis often arise from constructing a sequence of unobservable responses to subtasks that must all be correct in order to correctly answer the item (Embretson, 1984, 1997; Maris, 1999). By recognizing that the performances of examinees cannot be precisely predicted from their attribute vectors, stochastic models allow for the possibility of "slips" and "guesses". A slip occurs when an examinee who possesses the required attributes fails to correctly perform a subtask, or fails to answer the item correctly. A guess refers to correctly answering an item or completing a subtask in the absence of one or more required attributes. The models we consider are largely defined by whether slips and guesses are allowed to take place at the subtask level or at the item level. These models are briefly introduced in the following, using the nomenclature for these models in Junker and Sijtsma (2001), with more mathematical descriptions given in the next section.

In the deterministic inputs, noisy "and" gate (DINA) model, slips and guesses occur at the item level (Junker & Sijtsma, 2001). Each item divides the population into two classes such that examinees within the same class have equal probabilities of answering correctly. The two classes are those who have all of the required attributes for the item and those who do not. Members of the class having the required attributes may still slip, and this occurs with a probability that is estimated from the data. Also, a parameter that indicates the probability of correctly answering the item for members of the class who lack at least one of the required attributes must be estimated. These slip and guessing parameters are allowed to change item by item. supposedly 2 super liit na values

In contrast with the DINA model, slipping and guessing in the noisy inputs, deterministic, "and" gate (NIDA) model, occur at the subtask level. In the NIDA model, introduced by Maris (1999), an item is answered correctly provided all subtasks are correctly performed. However, slips and guesses may take place for each subtask, depending on the examinee's attribute profile. If an item requires an attribute, an examinee who possesses that attribute will perform the subtask correctly provided the examinee does not slip, and an examinee who lacks that attribute still may guess correctly to complete the subtask. Thus, the model parameters are slip and guessing parameters that pertain to each attribute, rather than to each item.

Neither the DINA model nor the NIDA model consider the possibility that examinees may solve a problem in different ways. In his analysis of fraction subtraction data, Mislevy (1996) considered the notion of multiple strategies, in which a strategy refers to the set of required attributes. A straightforward extension of the DINA model allows for the incorporation of multiple strategies. The basic idea is to note whether an examinee's attribute pattern satisfies either strategy or not. In this way, an item still divides the population into two equivalence classes, but in a different manner than in the single-strategy DINA model described earlier.

In the next section the mathematical details of the NIDA model and the single-strategy and multiple-strategy DINA models are discussed, and the method of de la Torre and Douglas (2004) for parameterizing the joint distribution of a K-dimensional random attribute vector is described. The third section introduces several methods of assessing fit both globally and at the item level, which may be used for model selection. The fourth section includes results of simulation studies to examine the performance of the MCMC algorithms for parameter estimation. The fifth section provides an extensive study of the fraction subtraction data, and compares the three models using the proposed goodness-of-fit measures. The paper concludes with a discussion of the utility of the proposed models and recommendations for assessing goodness of fit.

2. Model Specification and Estimation

This section gives the mathematical details by which the NIDA model and the single-strategy and multiple-strategy DINA models relate the item response vector $\boldsymbol{Y}$ to the attribute vector $\boldsymbol{\alpha}$. All of these models require construction of a $\boldsymbol{Q}$-matrix (Embretson, 1984; Tatsuoka, 1985), which indicates the attributes needed for each item. $\boldsymbol{Q}$ is a $J \times K$ matrix with the j,k entry $q_{jk} = 1$ if the correct application of attribute k influences the probability of correctly answering the jth item, and equals 0 otherwise. A distinguishing feature of the multiple-strategy DINA model is that it incorporates multiple $\boldsymbol{Q}$-matrices in order to specify the different strategies that suffice to solve the examination problems.

2.1. Single-Strategy DINA Model

As previously discussed, the DINA model allows each item to divide the population into those who possess all the required attributes and those who do not. It can be viewed as a latent response model in which slips and guesses occur at the item level, rather than at the subtask level. Let η_{ij} denote whether the ith examinee possesses the attributes required for the jth item. This can be expressed by the equation $\eta_{ij} = \prod_{k=1}^{K} \alpha_{ik}^{q_{jk}}$.

The parameters for a correct response to item j are denoted by s_j and g_j. The parameter s_j refers to the probability of slipping and incorrectly answering the item when $\eta_{ij} = 1$, and g_j is the probability of correctly guessing the answer when $\eta_{ij} = 0$. The item response function for the jth item may then be written as

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

Assuming conditional independence as well as independence among N subjects, the joint likelihood function of the DINA model is

$$L(\boldsymbol{s}, \boldsymbol{g}; \boldsymbol{\alpha}) = \prod_{i=1}^{N} \prod_{j=1}^{J} \left[(1 - s_j)^{y_{ij}} s_j^{1-y_{ij}}\right]^{\eta_{ij}} \left[g_j^{y_{ij}} (1 - g_j)^{1-y_{ij}}\right]^{1-\eta_{ij}}. \tag{2}$$

The DINA model requires only two parameters for each item, and offers a clear interpretation. It is most appropriate when the conjunction of several equally important attributes is required. Applications of the DINA model along with MCMC algorithms for estimation are given in Junker and Sijtsma (2001), Tatsuoka (2002), and de la Torre and Douglas (2004). The DINA model is also discussed in Macready and Dayton (1977), Haertel (1989), and Doignon and Falmagne (1999).

2.2. Multiple-Strategy DINA Model

The multiple-strategy DINA model is a straightforward extension of the single-strategy DINA model. Suppose that each item has as many as M distinct strategies that would suffice to solve it. A strategy is defined as a subset of the K attributes which could be used in conjunction to solve the problem. This may be coded by constructing M different matrices, $\boldsymbol{Q}_1, \boldsymbol{Q}_2, \ldots, \boldsymbol{Q}_M$. For examinee i and item j we modify the definition of η_{ij} to consider the M different strategies.

Let $\eta_{ijm} = \prod_{k=1}^{K} \alpha_{ik}^{q_{jkm}}$, for $m = 1, 2, \ldots, M$, where q_{jkm} denotes the element in the jth row and kth column of $\boldsymbol{Q}_m$. The variable η_{ijm} denotes if examinee i has the attributes to apply the mth strategy to the jth item. Then we check if at least one of the M strategies is satisfied by setting

$$\eta_{ij} = \max\{\eta_{ij1}, \eta_{ij2}, \ldots, \eta_{ijM}\}. \tag{3}$$

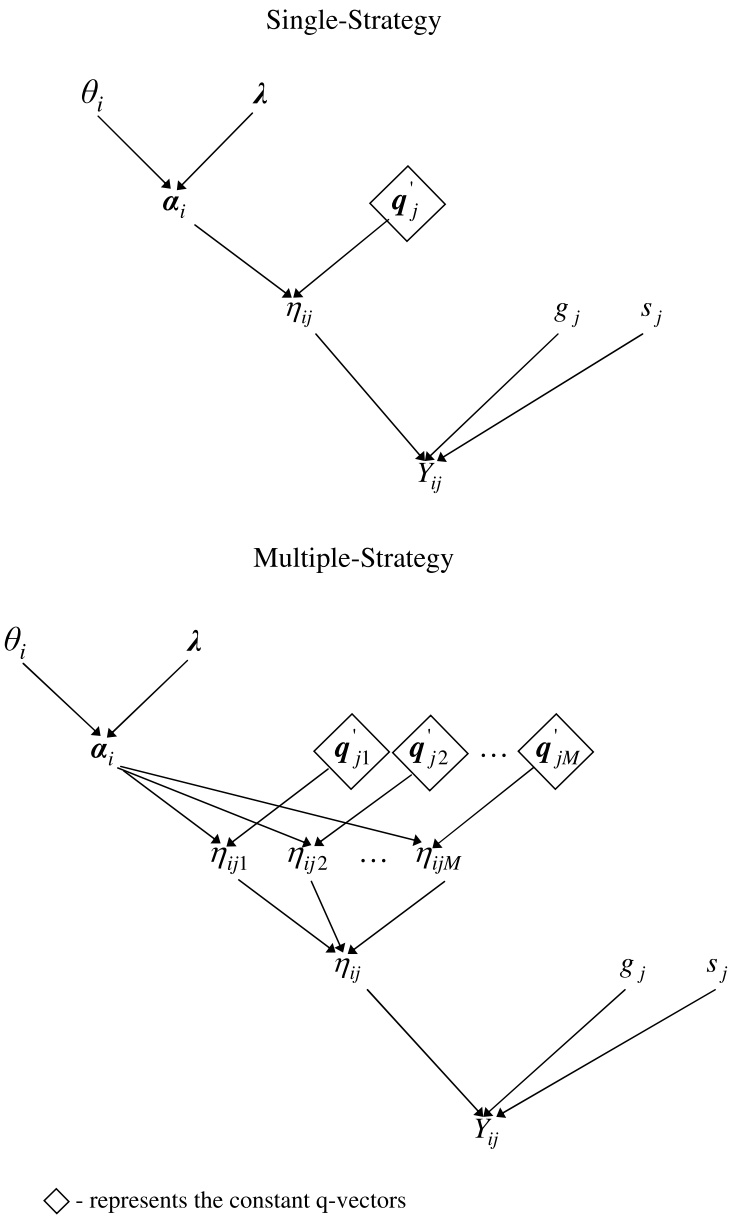

FIGURE 3.
Directed acylic graphs of the higher-order single- and multiple-strategy DINA model.

Once η_{ij} has been determined, the item response function is precisely the same as in (1), and the likelihood function is the same as in (2). This same approach of checking if any strategies may be satisfied by an examinee's attribute pattern and dividing the population into two equivalence classes accordingly has been applied in the POSET model in Tatsuoka (2002). A schematic differentiation of the single- and multiple-strategy DINA models is given in Fig. 3. The parameters θ and $\boldsymbol{\lambda}$ will be defined in Sect. 2.4.

The simplicity of the multiple-strategy DINA model is appealing, but it makes the assumption that a student will be able to identify a strategy to use successfully. Also, because the s and

g parameters are the same for different strategies, it assumes that the application of each strategy is equally difficult. Both of these points are in contrast to Mislevy's (1996) approach to modeling multiple strategies in mixed-number subtraction.

Although Mislevy (1996) mentioned extensions in which examinees may switch strategies, he focused more on a mixture modeling approach in which a student typically uses one strategy over the other, and allows for different strategies to be associated with different levels of difficulty. To make these distinctions clearer, one could consider how the multiple-strategy DINA model can be extended to allow for the strategies to have different s and g for each item. In addition to the current model, we associate each examinee with a particular strategy defined by a latent variable ω.

Suppose that each item can be described in terms of M strategies. To simplify the discussion we will assume that M is the same for all items although this approach can easily be generalized so that different items can be associated with different numbers of strategies. The attributes required for the jth item using strategy m are given in the jth row of $\boldsymbol{Q}_m$. Now suppose that for the ith examinee, the latent variable ω_i takes a value in the set $\{1, 2, \ldots, M\}$ indicating which type of strategy the ith examinee uses. Then the item response function for the jth item is given by

$$P[Y_{ij} = 1 \mid \boldsymbol{\alpha}_i, \omega_i] = \sum_{m=1}^{M} I[\omega_i = m](1 - s_{jm})^{\eta_{ijm}} g_{jm}^{1-\eta_{ijm}}, \tag{4}$$

where $I[\omega_i = m]$ is the indicator function that the ith examinee favors the mth strategy. In order to model the item response function as a function of $\boldsymbol{\alpha}$ alone, we need to mix over the conditional distribution of ω given $\boldsymbol{\alpha}_i$ which would require either assuming independence of ω and $\boldsymbol{\alpha}$, or modeling their joint distribution. We may then write

$$P[Y_{ij} = 1 \mid \boldsymbol{\alpha}_i] = \sum_{m=1}^{M} P[\omega_{ij} = m \mid \boldsymbol{\alpha}_i](1 - s_{jm})^{\eta_{ijm}} g_{jm}^{1-\eta_{ijm}}. \tag{5}$$

Compared to this mixture model, the multiple-strategy DINA model we have proposed does not allow for estimating a latent strategy being used. On the other hand, it does not constrain examinees to use a particular strategy throughout an examination (i.e., examinees can change their strategy from item to item).

As more strategies are involved, some confounding and loss of information can be expected for certain components of $\boldsymbol{\alpha}$. The degree of confounding and loss of information depend largely on the structure of the $\boldsymbol{Q}$-matrix. That is, for fixed test length and item quality, some $\boldsymbol{Q}$-matrices may allow better discrimination of the attributes patterns than others. As a practical matter, this issue should be addressed by examining the equivalence classes of $\boldsymbol{\alpha}$ patterns under the assumption of a deterministic multiple-strategy model in which all the guessing and slip parameters are assumed to be 0. If this deterministic model leads to the same theoretical response patterns for many distinct values of $\boldsymbol{\alpha}$, a simpler model might be required, using fewer strategies, or even a single strategy.

Following is an example of how different $\boldsymbol{Q}$-matrices can result in tests that have varying discrimination power. This example involves five attributes and two strategies, A and B. Strategy B is given in two alternative formulations, B_1 and B_2, and their rows are permutations of one another (see Table 1). Although the structure of the $\boldsymbol{Q}$-matrices of the two alternative strategies are identical, they produce $\boldsymbol{Q}$-matrices with different structures when used in conjunction with Strategy A. Using Strategies A and B_1 produces 16 equivalent classes, whereas using Strategies A and B_2 results in 23 equivalent classes. Thus, for this example, A and B_1 cannot discriminate between the attribute patterns $(1, 0, 0, 0, 1)'$, $(0, 0, 0, 1, 1)'$, and $(1, 0, 0, 1, 1)'$, but A and B_2 can.

TABLE 1.
$\boldsymbol{Q}$-matrices with two formulations of Strategy B.

Item	Strategy								
	A			B_1			B_2		
	1	2	3	3	4	5	3	4	5
1	1	0	0	0	1	0	1	0	1
2	0	1	0	1	0	0	1	1	0
3	0	0	1	0	0	1	0	0	1
4	1	1	0	1	0	1	0	1	0
5	1	0	1	1	1	0	1	0	0
6	0	1	1	0	0	1	0	0	1
7	1	1	1	1	1	1	1	1	1

This example underscores the fact that despite the limitations of the current formulation of the multiple-strategy DINA model, it can be used to effectively estimate $\boldsymbol{\alpha}$ in situations where the $\boldsymbol{Q}$-matrices can be well designed to ensure identification of most of the $\boldsymbol{\alpha}$ patterns.

As another critical point concerning the multiple-strategy DINA model, notice that a strategy is merely defined by the set of attributes required by a particular approach to solving a problem. One can imagine that a strategy might instead be determined by a set of attributes as well as a procedure and sequence for using them. In the example of fraction subtraction, one can think of the attributes as steps in solving the problem, making the set of attributes more or less equivalent with a strategy. However, depending on how the attributes are defined, this will not always be the case, and one must consider different methods of using the same attributes.

Finally, the within-person mixture feature of the multiple-strategy DINA model is consistent with the *overlapping waves theory* in cognitive science literature which states that individuals possess and adaptively use competing strategies that vary with situational demands (Opfer & Siegler, 2008). The presence of multiple strategies and their adaptive use are evident across different age groups (Siegler, 1988; Siegler et al., 1996), and have been found in diverse domains such as addition and subtraction (Siegler & Shrager, 1984), single-digit multiplication (LeFevre et al., 1996), question answering (Reder, 1987), and selection of causal rules (Shultz et al., 1986). However, adapting the multiple-strategy DINA model to these domains would require additional work. For example, to model the phenomenon that Siegler and Shrager (1984) found (i.e., different percentage of errors are associated with different strategies), the slip and guessing parameters of an item should be allowed to vary across the different strategies.

2.3. *NIDA Model*

The NIDA model, introduced in Maris (1999), considers slips and guesses at the subtask level (refer to Fig. 4). Let η_{ijk} indicate whether the ith subject correctly applied the kth attribute in completing the jth item. Slip and guessing parameters are indexed by attribute rather than by item and are defined by $s_k = P(\eta_{ijk} = 0 \mid \alpha_{ik} = 1, q_{jk} = 1)$ and $g_k = P(\eta_{ijk} = 1 \mid \alpha_{ik} = 0, q_{jk} = 1)$. As a technical device, we set $P(\eta_{ijk} = 1 \mid q_{jk} = 0)$ equal to 1, regardless of the value of α_{ik}. In the NIDA model an item response Y_{ij} will be equal to 1 if all η_{ijk}'s are equal to one, which can be expressed by the product $Y_{ij} = \prod_{k=1}^{K} \eta_{ijk}$. By assuming the η_{ijk}'s are independent conditional on $\boldsymbol{\alpha}_i$, the item response function which relates the probability of a successful response to the latent attribute pattern has the form

$$P(Y_{ij} = 1 \mid \boldsymbol{\alpha}_i, \boldsymbol{s}, \boldsymbol{g}) = \prod_{k=1}^{K} P(\eta_{ijk} = 1 \mid \alpha_{ik}, s_k, g_k) = \prod_{k=1}^{K} \left[(1 - s_k)^{\alpha_{ik}} g_k^{1-\alpha_{ik}}\right]^{q_{jk}}.$$

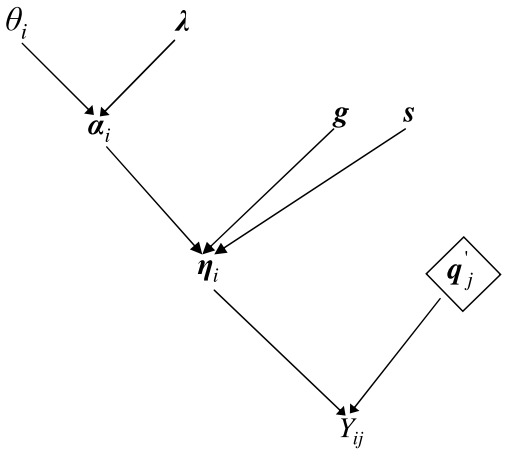

FIGURE 4.
Directed acylic graph of the higher-order NIDA model.

By assuming conditional independence and independence among subjects, the likelihood function is

$$L(\boldsymbol{s},\boldsymbol{g};\boldsymbol{\alpha}) = \prod_{i=1}^{N}\prod_{j=1}^{J}\left\{\prod_{k=1}^{K}\left[(1-s_k)^{\alpha_{ik}} g_k^{1-\alpha_{ik}}\right]^{q_{jk}}\right\}^{Y_{ij}}\left\{1-\prod_{k=1}^{K}\left[(1-s_k)^{\alpha_{ik}} g_k^{1-\alpha_{ik}}\right]^{q_{jk}}\right\}^{1-Y_{ij}}. \quad (6)$$

This version of the NIDA model, also given in Junker and Sijtsma (2001) and de la Torre and Douglas (2004), is a simplified version of the conjunctive model introduced in Maris (1999). Maris allows slip and guessing parameters for each attribute to vary across the items. A further extension of the NIDA model is the Unified Model of DiBello et al. (1995). In the Unified Model $\boldsymbol{s}$ and $\boldsymbol{g}$ are allowed to vary across the items, and a unidimensional continuous latent trait is incorporated into the conditional distribution to account for attributes that the items have in common, which were not included in coding the $\boldsymbol{Q}$-matrix.

2.4. *Joint Distribution of Attributes*

After specifying the conditional distribution of $\boldsymbol{Y}$ given $\boldsymbol{\alpha}$, using the NIDA model or the single-strategy or multiple-strategy DINA models, an additional step in specifying the model is to consider the probability distribution of $\boldsymbol{\alpha}$. In this paper we use the higher-order latent trait structure introduced by de la Torre and Douglas (2004). This approach derives from the observation that despite the aim of obtaining specific cognitive diagnostic information, many of the examinations used for skills diagnosis could also be seen as primarily measuring a small number of general abilities. In these situations, the choice to use a diagnostic model or a unidimensional IRT model indicates the desire for formative or summative assessment, respectively. The higher-order latent trait model combines these approaches by assuming conditional independence of $\boldsymbol{Y}$ given $\boldsymbol{\alpha}$, and also by assuming that the components of $\boldsymbol{\alpha}$ are independent conditional on θ, a unidimensional latent trait representing general ability in the studied domain. This latent trait could be assumed to have any distribution, but de la Torre and Douglas (2004) used the standard normal distribution as its prior.

In the example of fraction subtraction given in a later section, specific rules for manipulating fractions and whole numbers and subtracting them are identified, and are used to define the attribute vector $\boldsymbol{\alpha}$. The probability model for $\boldsymbol{\alpha}$ conditional on θ is

$$P(\boldsymbol{\alpha} \mid \theta) = \prod_{k=1}^{K} P(\alpha_k \mid \theta), \tag{7}$$

where the probability of mastery is given by

$$P(\alpha_k = 1 \mid \theta) = \frac{\exp[1.7\lambda_{1k}(\theta - \lambda_{0k})]}{1 + \exp[1.7\lambda_{1k}(\theta - \lambda_{0k})]}, \tag{8}$$

where λ_{0k} and λ_{1k} represent the parameters in regressing the latent mastery or nonmastery of attribute k on the higher-order proficiency θ. The constant 1.7 was used to give the λs similar interpretations as the difficulty and discrimination parameters of item response models. The original formulation given by de la Torre and Douglas (2004) allows the joint distribution of the attributes to take other forms (e.g., multidimensional).

Modeling the joint distribution of $\boldsymbol{\alpha}$ in this way reduces the complexity of the saturated model for $\boldsymbol{\alpha}$ and is reasonable in applications in which the examination can be seen as measuring a general ability in addition to the specific skills that are indicated by $\boldsymbol{\alpha}$. The estimated value of $\boldsymbol{\lambda}$ relating $\boldsymbol{\alpha}$ to the higher-order latent trait θ may be used to estimate the population proportion for each of the 2^K possible attribute patterns. Finally, the higher-order model enables one to classify each α_k and obtain an estimate of θ in the same analysis.

2.5. *Parameter Estimation*

A fully Bayesian framework was adopted in estimating the parameters of the model. Samples from the joint posterior distribution were obtained using Markov chain Monte Carlo (MCMC) simulation. More specifically, samples from the full conditional distributions were iteratively drawn using the Metropolis–Hastings algorithm (Casella & George, 1992; Chib & Greenberg, 1995; Geman & Geman, 1984; Patz & Junker, 1999a, 1999b).

3. Model Assessment

Model fit was evaluated using various indices. Three indices were computed by comparing the expected and observed characteristics of the marginal and pairwise joint distributions of the items. Four other indices, Bayes factor, Bayesian Information Criterion (BIC; Schwarz, 1978) and Akaike Information Criterion (AIC; Akaike, 1973), and Deviance Information Criterion (DIC; Spiegelhalter et al., 2002) were also computed to provide global measures of the relative fit of the models.

For the marginal distributions of the items, the proportion of examinees correctly responding to each item was obtained and compared to the expected proportion computed under the estimated model. The pairwise relationships were assessed by comparing observed and expected product-moment correlations and the log-odds ratios. For items j and j', the log-odds ratio is computed as

$$\log\left[\frac{P(Y_j = 1, Y_{j'} = 1)P(Y_j = 0, Y_{j'} = 0)}{P(Y_j = 1, Y_{j'} = 0)P(Y_j = 0, Y_{j'} = 1)}\right]. \tag{9}$$

To compute the expected values of these indices under the estimated model, the Monte Carlo method was used. This was done by generating responses for 100,000 examinees under the estimated model parameters, and computing the sample values of the indices. Determining that item

pair associations are modeled correctly is a necessary, but not sufficient, condition for determining whether or not the association of the entire response vector is adequately modeled.

The Bayes factor is analogous to the likelihood ratio, but is used in a Bayesian context, and can be used in comparing models that may not be nested. The Bayes factor, which is the ratio of the marginal likelihoods of two competing models M_H and M_I (i.e., the likelihoods after integrating over the model parameters), is computed as

$$B_{HI} = \frac{P(\boldsymbol{Y}|M_H)}{P(\boldsymbol{Y}|M_I)}. \tag{10}$$

In (10),

$$P(\boldsymbol{Y}|M_m) = \int P(\boldsymbol{Y}|\boldsymbol{\lambda}_m, \boldsymbol{s}_m, \boldsymbol{g}_m, M_m) P(\boldsymbol{\lambda}_m, \boldsymbol{s}_m, \boldsymbol{g}_m|M_m)\, d\boldsymbol{\lambda}_m\, d\boldsymbol{s}_m\, d\boldsymbol{g}_m, \tag{11}$$

where $\boldsymbol{\lambda}_m$, $\boldsymbol{s}_m$, and $\boldsymbol{g}_m$ are the parameters under Model m, $P(\boldsymbol{\lambda}_m, \boldsymbol{s}_m, \boldsymbol{g}_m|M_m)$ is the prior density. The marginal likelihoods required for the Bayes factor were computed for the three models in the analysis of fraction subtraction data. In each case, the marginal likelihood was approximated using the Laplace–Metropolis estimator proposed by Raftery (1996). In addition, because the interest is in the structural parameters when choosing a model, the incidental parameters were integrated out of the marginal likelihood. Moreover, the posterior mode and variance were approximated by the posterior mean and sample covariance matrix of the simulation output. Finally, integration of continuous functions was approximated using quadrature nodes. For the computational details of the marginal likelihood, see de la Torre and Douglas (2004).

For the BIC and AIC, the maximized conditional log-likelihood given the structural parameters, $\log(P(\boldsymbol{Y}|\tilde{\boldsymbol{\lambda}}, \tilde{\boldsymbol{s}}, \tilde{\boldsymbol{g}}))$, was approximated by using the posterior means in place of the maximum likelihood estimates. In addition, similar to the marginal likelihood, the conditional log-likelihood was computed by integrating out the incidental parameters, θ and α. Finally, the DIC was computed based on DIC_4 for missing data models proposed by Celeux et al. (2006).

4. Simulation Studies

The primary objective of the simulation studies is to demonstrate that the MCMC algorithms for parameter estimation discussed in this paper can effectively recover model parameters. De la Torre and Douglas (2004) showed that the parameters of the higher-order, single-strategy DINA model can be estimated accurately using MCMC. This paper aims to show that the parameters of the higher-order, multiple-strategy DINA model and the higher-order NIDA model can also be accurately estimated. In particular, in the case of the multiple-strategy DINA model, we wish to see if parameters can be accurately estimated in scenarios similar to that of the real data analysis covered in the following section. A secondary objective of this section is to investigate the impact of model misspecification on parameter estimation and attribute classification.

4.1. Simulation Study I: The Higher-Order NIDA and Single-Strategy DINA Models

4.1.1. Method For the first simulation study, 25 data sets with five attributes, 20 items and 2000 examinees were simulated using the higher-order NIDA model. The structural parameters (i.e., $\boldsymbol{\lambda}$, $\boldsymbol{s}$, and $\boldsymbol{g}$) were fixed across the 25 replications, whereas the incidental parameters θ_i and α_{ik} were sampled from Normal(0,1) and Bernoulli($\{1+\exp[-1.7\lambda_1(\theta_i - \lambda_{0k})]\}^{-1}$), respectively, for each replication. Table 2 gives the $\boldsymbol{Q}$-matrix used in this simulation study. This $\boldsymbol{Q}$-matrix was constructed to be balanced in that each attribute appears in a pair, or in a triple the same number

TABLE 2.
The $\boldsymbol{Q}$-matrix for simulation study I.

Item	Attribute					Item	Attribute				
	1	2	3	4	5		1	2	3	4	5
1	1	1	0	0	0	11	1	1	1	0	0
2	1	0	1	0	0	12	1	1	0	1	0
3	1	0	0	1	0	13	1	1	0	0	1
4	1	0	0	0	1	14	1	0	1	1	0
5	0	1	1	0	0	15	1	0	1	0	1
6	0	1	0	1	0	16	1	0	0	1	1
7	0	1	0	0	1	17	0	1	1	1	0
8	0	0	1	1	0	18	0	1	1	0	1
9	0	0	1	0	1	19	0	1	0	1	1
10	0	0	0	1	1	20	0	0	1	1	1

TABLE 3.
Mean and SD of $\boldsymbol{\lambda}$ estimates (over 25 replications).

Parameter	Fitted model	
	NIDA	DINA
λ_{01}: −2.00	−1.98 (0.10)	−1.81 (0.12)
λ_{02}: −1.50	−1.46 (0.08)	−1.37 (0.08)
λ_{03}: −1.00	−0.97 (0.06)	−0.95 (0.06)
λ_{04}: −0.50	−0.50 (0.05)	−0.54 (0.05)
λ_{05}: 0.00	0.01 (0.04)	−0.09 (0.04)
λ_{1}: 1.00	1.02 (0.04)	1.11 (0.05)

of times as other attributes. The simulated data were fitted using the higher-order NIDA model and the higher-order single-strategy DINA model.

The same prior distributions were used for both models. For $\boldsymbol{\lambda}$ and $\boldsymbol{\theta}$, the parameters μ and σ^2 of the prior distributions were set to 0 and 1, and for the guessing and slip parameters, a 4-Beta$(0, 0.9, 1.5, 2)$ prior was used. (The four-parameter beta distribution 4-Beta(ν, ω, a, b) is a generalization of the beta distribution Beta(ν, ω), and has the interval (a, b), rather than $(0, 1)$, as its support.) The multivariate potential scale reduction factor (MPSRF; Brooks and Gelman, 1998) was used as the criterion to verify stationarity of the chains for all structural parameters. By running five parallel chains with a burn-in of 1000 iterations, followed by 4000 iterations for the first simulated data set, the MPSRFs for the NIDA and DINA models were computed as $\hat{R}^{16} = 1.17$, $\hat{R}^{46} = 1.14$.

4.1.2. Results For this simulation study, a common discrimination parameter ($\lambda_{1k} = \lambda_1$, for all k) was used to model the joint distribution of the attributes. Table 3 gives the mean and standard deviation of the parameter estimates using the NIDA and single-strategy DINA models across the 25 replications. Results show that in estimating $\boldsymbol{\lambda}$, using the correct model, in this case, NIDA model, provided estimates that are both more accurate and less variable. Nonetheless, differences in estimates of $\boldsymbol{\lambda}$ had little effect on estimating the higher-order latent trait θ. The mean correlation and RMSE between the true and estimated abilities in Table 4 indicate that although the correct model gave better estimates of θ (i.e., higher correlation and lower RMSE), the differences from estimates given by the misspecified model are very slight.

Table 5 provides the percent of the attributes correctly classified using the NIDA and DINA models. Results show that correct classifications using the correct model were consistently higher

TABLE 4.
Mean correlation and RMSE between θ and $\hat{\theta}$ (over 25 replications).

Model	Correlation	RMSE
NIDA	0.75	0.66
DINA	0.74	0.67

TABLE 5.
Percent of correct attribute classification (over 25 replications).

Model	Attribute				
	α_1	α_2	α_3	α_4	α_5
NIDA	96.47	94.78	94.43	94.29	94.88
DINA	95.19	93.53	93.20	92.86	93.34

TABLE 6.
Estimation of the parameters of the NIDA model (over 25 replications).

Attribute	g	$1-s$	$\hat{g}$	SD($\hat{g}$)	$1-\hat{s}$	SD($1-\hat{s}$)
1	0.15	0.88	0.16	(0.03)	0.88	(0.00)
2	0.20	0.90	0.21	(0.02)	0.90	(0.01)
3	0.25	0.93	0.25	(0.01)	0.92	(0.01)
4	0.30	0.95	0.30	(0.01)	0.95	(0.01)
5	0.35	0.98	0.35	(0.01)	0.98	(0.01)

across the five attributes. However, for the particular example, attribute classification using the DINA model was not substantially inferior to the correct model. Finally, Table 6 shows that following the MCMC algorithm developed for this paper, efficient estimates of the guessing and slip parameters of the NIDA model can be obtained.

4.2. Simulation Study II: The Higher-Order, Single-Strategy and Higher-Order, Multiple-Strategy DINA Models

4.2.1. Method This two-part simulation study is a comparison between the higher-order, single-strategy and higher-order, multiple-strategy DINA models. For the first part, 25 data sets were generated using the single-strategy DINA model. For the second part, another 25 data sets were generated using the multiple-strategy DINA model. All the data sets were analyzed using the two models. The $\boldsymbol{Q}$-matrix corresponding to Strategy A in Table 7 was used for part one, whereas the $\boldsymbol{Q}$-matrices corresponding to Strategies A and B were used for part 2. Although each strategy requires five attributes, three common attributes are present in both strategies. Hence, the multiple-strategy DINA model involved seven distinct attributes, with attributes 6 and 7 used only for the multiple-strategy DINA model. For this simulation study, in addition to a common discrimination parameter a common intercept ($\lambda_{0k} = -1.00$, for all k), was used for the seven attributes, and the guessing and slip parameters were set at 0.20. Except for the specific models, and the numbers of iterations and burn-in (which were twice as long for the multiple-strategy DINA model), the remaining steps (i.e., the data generation, parameter estimation, convergence verification) proceeded in the same manner as the above simulation study. For data generated using the single-strategy DINA model, the MPSRFs for the single- and multiple-strategy DINA models $\hat{R}^{46}$ and $\hat{R}^{48}$, were 1.20 and 1.19. These factors were equal to 1.14 and 1.16 for the data generated under the multiple-strategy DINA model.

TABLE 7.
The $\boldsymbol{Q}$-matrices for simulation study II.

Item	Attribute									
	Strategy A					Strategy B				
	1	2	3	4	5	3	4	5	6	7
1	1	1	0	0	0	0	1	0	1	1
2	1	0	1	0	0	0	0	1	1	1
3	1	0	0	1	0	0	1	1	0	1
4	1	0	0	0	1	0	1	1	1	0
5	0	1	1	0	0	1	0	0	1	1
6	0	1	0	1	0	1	1	0	0	1
7	0	1	0	0	1	1	1	0	1	0
8	0	0	1	1	0	1	0	1	0	1
9	0	0	1	0	1	1	0	1	1	0
10	0	0	0	1	1	1	1	1	0	0
11	1	1	1	0	0	0	0	0	1	1
12	1	1	0	1	0	0	1	0	0	1
13	1	1	0	0	1	0	1	0	1	0
14	1	0	1	1	0	0	0	1	0	1
15	1	0	1	0	1	0	0	1	1	0
16	1	0	0	1	1	0	1	1	0	0
17	0	1	1	1	0	1	0	0	0	1
18	0	1	1	0	1	1	0	0	1	0
19	0	1	0	1	1	1	1	0	0	0
20	0	0	1	1	1	1	0	1	0	0

4.2.2. Results Table 8 gives the estimates of $\boldsymbol{\lambda}$ using the correct and misspecified models. For both parts, using the correct model yielded estimates that are less biased. Other impacts of model misspecification depend on the generating model used. For the data generated using the single-strategy DINA model, misspecification resulted in overestimation of the intercepts of the attributes. However, the biases for the attributes common to both strategies (i.e., those used in generating the data) are much smaller compared to the biases for attributes unique to Strategy B which were ignored in the data generation. In addition to larger biases, estimates of these unique attributes have high variabilities. The high estimates for intercepts of attributes not used in the test indicate a very low prevalence of these attributes in the population. This implies that for items requiring either α_6 or α_7, η is largely determined by Strategy A.

For data generated using the multiple-strategy DINA model, intercepts of attributes unique to Strategy A were underestimated, whereas intercepts of attributes shared by the two strategies were overestimated using the misspecified model. In addition, misspecification and use of the single-strategy DINA model resulted in a high estimate of the discrimination parameter. This demonstrates the need to evaluate goodness of fit to determine the more appropriate model.

The impact of model misspecification on θ estimation and attribute classification is presented in Tables 9 and 10. For θ estimates, using the correct model generally provided estimates that correlated more highly with the true θ, and have lower RMSE. However, as in the previous section, the results obtained using misspecified models were not considerably inferior. Similarly, better attribute classification rates can be obtained using the correct model across all the relevant attributes. It can be noted because the multiple-strategy DINA model has higher complexity, the percent of correct attribute classification is generally lower when data are generated using this model. When data are generated using the single-strategy model, attributes common to both strategies have slightly lower correct classification using the multiple-strategy model. In contrast,

TABLE 8.
Mean and SD of $\boldsymbol{\lambda}$ estimates for data generated using the single- and multiple-strategy DINA models (over 25 replications; fitted model in braces).

Parameter	Generating model			
	Single-strategy		Multiple-strategy	
	Single	Multiple	Single	Multiple
λ_{01}	−0.99 (0.06)	−0.92 (0.06)	−1.18 (0.10)	−1.01 (0.10)
λ_{02}	−1.01 (0.08)	−0.92 (0.07)	−1.14 (0.09)	−1.03 (0.11)
λ_{03}	−1.00 (0.07)	−0.79 (0.06)	−0.76 (0.07)	−1.02 (0.07)
λ_{04}	−0.99 (0.06)	−0.81 (0.05)	−0.93 (0.06)	−1.01 (0.05)
λ_{05}	−1.01 (0.05)	−0.83 (0.04)	−0.86 (0.07)	−1.04 (0.08)
λ_{06}	–	3.31 (2.99)	–	−1.01 (0.07)
λ_{07}	–	2.90 (2.78)	–	−1.02 (0.10)
λ_{1}	1.01 (0.06)	1.18 (0.08)	1.67 (0.13)	0.97 (0.06)

TABLE 9.
Mean correlation and RMSE between θ and $\hat{\theta}$ (over 25 replications).

Fitted model	Generating model			
	Single		Multiple	
	Correlation	RMSE	Correlation	RMSE
Single	0.73	0.69	0.70	0.71
Multiple	0.72	0.69	0.72	0.69

TABLE 10.
Percent of correct attribute classification (over 25 replications).

Generating model	Fitted model	Attribute						
		α_1	α_2	α_3	α_4	α_5	α_6	α_7
Single	Single	91.34	91.28	90.91	91.23	91.16	–	–
	Multiple	90.62	90.54	89.08	89.52	89.93	–	–
Multiple	Single	83.81	83.66	89.08	89.76	89.04	–	–
	Multiple	86.02	85.44	92.87	92.97	92.80	85.85	86.00

attributes common to both strategies can be classified with higher accuracy when the data are generated using the multiple-strategy model. This is so because these attributes are measured by more items.

The estimates of the DINA model parameters are given in Tables 11 and 12. Note that the parameter estimates using the correct model are consistently more accurate for both generating models. When the data are generated from the multiple-strategy DINA model and fitted using the single-strategy DINA model, a large number of guessing parameters are overestimated, and when the data are generated from the single-strategy DINA model and fitted using the multiple-strategy DINA model, estimates are generally close to the generating parameters except for a few slip parameters that are overestimated. Overestimation of the guessing parameter is a consequence of treating alternative strategies as a part of the guessing mechanism, whereas overestimation of the slip parameter is due to attributing skills to examinees that they do not possess.

TABLE 11.
DINA model parameter estimates for data generated using the single-strategy DINA model (over 25 replications).

Item	Fitted DINA model							
	Single				Multiple			
	$\hat{g}$	SD($\hat{g}$)	$1-\hat{s}$	SD($1-\hat{s}$)	$\hat{g}$	SD($\hat{g}$)	$1-\hat{s}$	SD($1-\hat{s}$)
1	0.20	(0.02)	0.80	(0.01)	0.20	(0.02)	0.80	0.01
2	0.20	(0.02)	0.80	(0.01)	0.22	(0.02)	0.80	0.01
3	0.20	(0.02)	0.80	(0.01)	0.21	(0.02)	0.80	0.01
4	0.20	(0.02)	0.80	(0.01)	0.21	(0.02)	0.81	0.01
5	0.21	(0.02)	0.80	(0.01)	0.23	(0.02)	0.81	0.01
6	0.20	(0.02)	0.80	(0.01)	0.22	(0.02)	0.80	0.01
7	0.20	(0.02)	0.80	(0.01)	0.21	(0.02)	0.80	0.01
8	0.20	(0.02)	0.80	(0.01)	0.24	(0.02)	0.81	0.01
9	0.20	(0.02)	0.79	(0.01)	0.24	(0.02)	0.80	0.01
10	0.20	(0.02)	0.80	(0.01)	0.23	(0.02)	0.80	0.01
11	0.20	(0.01)	0.80	(0.01)	0.20	(0.01)	0.81	0.01
12	0.20	(0.01)	0.80	(0.01)	0.19	(0.01)	0.80	0.01
13	0.20	(0.01)	0.80	(0.01)	0.20	(0.01)	0.80	0.01
14	0.20	(0.01)	0.80	(0.01)	0.21	(0.01)	0.80	0.01
15	0.20	(0.01)	0.80	(0.01)	0.21	(0.01)	0.80	0.01
16	0.20	(0.01)	0.80	(0.01)	0.18	(0.02)	0.75	0.01
17	0.21	(0.01)	0.80	(0.01)	0.21	(0.01)	0.80	0.01
18	0.20	(0.01)	0.80	(0.01)	0.21	(0.01)	0.81	0.01
19	0.20	(0.01)	0.80	(0.01)	0.18	(0.02)	0.71	0.01

TABLE 12.
DINA model parameter estimates for data generated using the multiple-strategy DINA model (over 25 replications).

Item	Fitted DINA model							
	Single				Multiple			
	$\hat{g}$	SD($\hat{g}$)	$1-\hat{s}$	SD($1-\hat{s}$)	$\hat{g}$	SD($\hat{g}$)	$1-\hat{s}$	SD($1-\hat{s}$)
1	0.21	(0.03)	0.78	(0.01)	0.21	(0.03)	0.80	0.01
2	0.31	(0.02)	0.80	(0.01)	0.21	(0.02)	0.80	0.01
3	0.19	(0.02)	0.77	(0.01)	0.20	(0.02)	0.80	0.01
4	0.21	(0.02)	0.78	(0.01)	0.20	(0.02)	0.80	0.01
5	0.24	(0.02)	0.78	(0.01)	0.20	(0.02)	0.80	0.01
6	0.21	(0.02)	0.77	(0.01)	0.21	(0.02)	0.80	0.01
7	0.28	(0.02)	0.79	(0.01)	0.20	(0.02)	0.80	0.01
8	0.27	(0.02)	0.79	(0.01)	0.21	(0.02)	0.80	0.01
9	0.21	(0.02)	0.77	(0.01)	0.20	(0.02)	0.80	0.01
10	0.19	(0.02)	0.76	(0.01)	0.20	(0.02)	0.80	0.01
11	0.35	(0.02)	0.79	(0.01)	0.21	(0.02)	0.80	0.01
12	0.25	(0.02)	0.78	(0.01)	0.20	(0.02)	0.80	0.01
13	0.32	(0.02)	0.79	(0.01)	0.20	(0.02)	0.80	0.01
14	0.37	(0.02)	0.80	(0.01)	0.20	(0.02)	0.80	0.01
15	0.32	(0.02)	0.79	(0.01)	0.20	(0.02)	0.80	0.01
16	0.22	(0.02)	0.77	(0.01)	0.20	(0.02)	0.80	0.01
17	0.31	(0.02)	0.79	(0.01)	0.20	(0.02)	0.80	0.01
18	0.32	(0.02)	0.79	(0.01)	0.21	(0.02)	0.80	0.01
19	0.30	(0.02)	0.78	(0.01)	0.20	(0.02)	0.80	0.01
20	0.27	(0.02)	0.78	(0.01)	0.20	(0.02)	0.80	0.01

5. Fraction Subtraction Data

5.1. Data

The data we considered include responses by 2144 middle school students to 15 fraction subtraction items. The data set is a subset of the data originally used and described by Tatsuoka (1990), and recently analyzed by Tatsuoka (2002) and de la Torre and Douglas (2004). Stout et al. (2003) and Yan et al. (2003) analyzed a similar data set, using the Fusion model and Bayesian network, respectively. The three models, the higher-order NIDA, and the single- and multiple-strategy DINA models, were used to analyze the data. The $\boldsymbol{Q}$-matrices, given in Table 13 were adapted by the authors from a similar analysis by Mislevy (1996). Strategies A and B described earlier were considered for the multiple-strategy DINA model. With the two alternative strategies measuring three common attributes, the multiple-strategy DINA model measured seven unique attributes. Only the first five attributes were needed for the NIDA and single-strategy DINA models.

5.2. Results

The prior distributions used to analyze the fraction subtraction data were the same as those used in the simulation studies. To verify convergence, five chains with random starting points over a broad range of reasonable values were run for the three models. For all the NIDA and single-strategy DINA and models, chains with burn-ins of 5000 iterations and total iterations of 25,000 were used. Using these chain lengths and burn-ins, the MPSRFs for the NIDA, single-strategy DINA models were $\hat{R}^{16} = 1.13$ and $\hat{R}^{36} = 1.12$, respectively. For the multiple-strategy DINA model, the burn-in and total iterations were 25,000 and 250,000, and the corresponding MPSRF was $\hat{R}^{38} = 1.17$. The MPSRFs indicate that the chains have reached approximate stationarity. Finally, parameter estimates were based on the mean estimates across the five parallel chains, and the corresponding posterior standard deviations were obtained by taking the root of the mean posterior variances across the five separate chains.

TABLE 13.
The $\boldsymbol{Q}$-matrices for the analysis of fraction subtraction data.

Item	Attribute								
		Strategy A				Strategy B			
	1	2	3	4	5	2	5	6	7
1	1	0	0	0	0	0	0	0	0
2	1	1	1	1	0	0	0	1	0
3	1	0	0	0	0	0	0	0	0
4	1	1	1	1	1	0	1	1	0
5	0	0	1	0	0	1	1	1	1
6	1	1	1	1	0	1	0	1	0
7	1	1	1	1	0	1	0	1	0
8	1	1	0	0	0	1	0	0	0
9	1	0	1	0	0	0	0	1	0
10	1	0	1	1	1	1	1	0	0
11	1	0	1	0	0	1	0	1	0
12	1	0	1	1	0	1	0	1	0
13	1	1	1	1	0	1	0	1	1
14	1	1	1	1	1	1	1	1	1
15	1	1	1	1	0	1	0	1	1

Estimates of the higher-order structural parameters for the three models are given in Table 14. The estimates of λ_0 for the NIDA model show that attribute 1 is the most prevalent attribute in the examinee population, whereas attribute 3 is the least prevalent. In addition, the negative coefficients of the location parameters indicate that each of the attributes is possessed by more than half of the population. The guessing parameter estimates of the NIDA model in Table 15 indicate that examinees lacking attributes 2 and 3 are able to guess correctly more than 60% of the time. The small slip parameter estimates indicate that examinees who possess the required attributes are able to apply the attributes correctly almost all the time, particularly for attributes 2, 4, and 5.

For λ parameters common to the single- and multiple-strategy DINA models, the estimates indicate that more than half of the population possess attributes 1, 2, and 3, and less than half possess attributes 4 and 5. The guessing and slip parameter estimates listed in Table 16 for the two models are quite similar for all items except for the guessing parameter of item 10. Whereas similarity in parameter estimates favor the simpler model (i.e., single-strategy DINA model) in that increased model complexity by introducing additional attributes does not result in markedly different estimates, the higher estimate of the guessing parameter using the single-strategy DINA model indicates that examinees may have used the alternative strategy in responding to this item.

The observed indices (proportion correct, log-odds ratio, and correlation) for the 15 items are given in Table 17. The residuals between the observed and predicted indices provided by the three models and the corresponding standard errors are given in Tables 18–20 and 21–23, respectively. The residuals were obtained by comparing the observed indices and the predicted indices, where the latter were computed from a simulated response matrix using a large number of generated examinees (i.e., 100,000) and the means of the chains of the structural parameters. The standard errors of the residuals were computed using the standard deviations of the residuals obtained from generating response matrices of simulated examinees and 100 randomly selected draws from the chains. We have verified for the NIDA model that randomly sampling 20 draws from each of the five chains yielded residuals with means and variances that are almost identical to those obtained

TABLE 14.
Mean and posterior standard deviation of $\hat{\boldsymbol{\lambda}}$ for the fraction subtraction data.

Parameter	Fitted model		
	NIDA	Single-DINA	Multiple-DINA
λ_{01}	−0.76 (0.05)	−0.89 (0.04)	−0.91 (0.04)
λ_{02}	−0.88 (0.54)	−0.76 (0.06)	−0.83 (0.06)
λ_{03}	−0.02 (0.07)	−0.88 (0.06)	−0.91 (0.06)
λ_{04}	−0.10 (0.05)	0.04 (0.03)	0.22 (0.10)
λ_{05}	−0.33 (0.08)	0.22 (0.05)	0.11 (0.04)
λ_{06}	–	–	0.35 (0.12)
λ_{07}	–	–	−0.42 (0.28)
λ_1	1.57 (0.10)	2.99 (0.38)	2.49 (0.20)

TABLE 15.
NIDA model parameter estimates for the fraction subtraction data.

Attribute	$\hat{g}$	SD($\hat{g}$)	$1-\hat{s}$	SD($1-\hat{s}$)
1	0.23	(0.02)	0.93	(0.00)
2	0.76	(0.12)	0.99	(0.00)
3	0.62	(0.02)	0.96	(0.01)
4	0.21	(0.01)	1.00	(0.00)
5	0.13	(0.04)	0.99	(0.01)

TABLE 16.
DINA model parameter estimates for the fraction subtraction data.

Item	Fitted DINA model							
	Single				Multiple			
	$\hat{g}$	SD($\hat{g}$)	$1-\hat{s}$	SD($1-\hat{s}$)	$\hat{g}$	SD($\hat{g}$)	$1-\hat{s}$	SD($1-\hat{s}$)
1	0.00	(0.00)	0.72	(0.01)	0.00	(0.00)	0.72	(0.01)
2	0.21	(0.01)	0.88	(0.01)	0.21	(0.01)	0.88	(0.01)
3	0.13	(0.03)	0.96	(0.01)	0.13	(0.03)	0.96	(0.01)
4	0.13	(0.01)	0.87	(0.01)	0.12	(0.01)	0.83	(0.02)
5	0.23	(0.03)	0.75	(0.01)	0.22	(0.03)	0.76	(0.01)
6	0.03	(0.01)	0.77	(0.01)	0.03	(0.01)	0.77	(0.01)
7	0.07	(0.01)	0.92	(0.01)	0.07	(0.01)	0.93	(0.01)
8	0.15	(0.02)	0.95	(0.01)	0.15	(0.02)	0.95	(0.01)
9	0.09	(0.02)	0.94	(0.01)	0.09	(0.02)	0.94	(0.01)
10	0.17	(0.01)	0.93	(0.01)	0.06	(0.01)	0.93	(0.01)
11	0.11	(0.02)	0.89	(0.01)	0.12	(0.02)	0.90	(0.01)
12	0.04	(0.01)	0.87	(0.01)	0.04	(0.01)	0.87	(0.01)
13	0.13	(0.01)	0.84	(0.01)	0.14	(0.01)	0.85	(0.01)
14	0.02	(0.00)	0.80	(0.02)	0.02	(0.00)	0.77	(0.02)
15	0.01	(0.00)	0.82	(0.01)	0.01	(0.00)	0.83	(0.01)

using all the draws. In addition, for all the models, the residuals calculated using the mean of the posterior distribution are highly similar to the mean of the residuals obtained by averaging the residuals from the 100 draws. The absolute maximum differences for the NIDA, single-strategy DINA, and multiple-strategy DINA models are 0.05, 0.06, and 0.07, respectively. Relative to the expected log-odds ratio that produced these differences, 1.58, 2.10, and 3.46, the maximum discrepancies between the two methods are deemed small, and do not alter the interpretation of the results. In comparing the residuals and their standard errors, it can be noted that a large proportion of the residuals are significantly different from zero because of the sample size ($N = 2144$), and the fact that no model can be entirely correct. In this light, the standard errors should be used as a guide in evaluating the relative size and not in determining the significance of the residuals.

In general, it can be said that the NIDA model poorly fits these data. The large absolute residuals for the proportion correct given in Table 18 (maximum of 0.17 and greater than 0.05 for more than half of the items) indicate that the model does an inadequate job of predicting even the first moment of the item responses. In addition, several large residuals for the log-odds ratio and the correlation involving different items, most notably items 1 and 9, can also be found in the same table. For these reasons, the NIDA model cannot be considered a viable model for these data.

In contrast, both the single-strategy and multiple-strategy DINA models fit the first moment quite accurately, as can be seen in Tables 19 and 20. In examining residuals for the log-odds ratio and correlation, it can be seen that many of the worst fits involve item 1 for both models. Item 1, $\frac{3}{4} - \frac{3}{8}$, which requires only attribute 1 (performing basic fraction subtraction operation) for both strategies, is the only problem that requires the examinees to find the least common denominator. Because this problem requires a more complex skill that cannot be subsumed under attribute 1, item 1 cannot be fit adequately. To achieve better fit, either a new attribute (i.e., finding a common denominator) must be defined, or the item be discarded. The similarity in parameter estimates obtained using the single- and multiple-strategy DINA models resulted in similar residuals for both models except for item 10, where the multiple-strategy DINA model provided smaller residuals. The overestimation of the guessing parameter for the single-strategy

TABLE 17.
Observed indices (diagonal—proportion correct; upper triangle—log-odds ratio; lower triangle—correlation).

Item	Item														
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	**0.58**	1.90	3.58	2.79	1.11	2.74	2.31	2.05	2.50	2.83	2.22	2.36	2.50	2.88	2.32
2	0.44	**0.53**	1.20	2.01	0.35	2.49	2.80	1.28	1.22	2.03	1.24	2.51	1.90	2.59	2.62
3	0.53	0.54	**0.79**	2.63	1.31	2.13	2.37	3.03	3.25	2.48	3.00	2.51	2.57	2.55	2.18
4	0.26	0.52	0.50	**0.40**	1.32	2.30	2.35	2.03	2.52	3.39	1.98	2.42	2.18	3.27	2.28
5	0.42	0.50	0.57	0.47	**0.65**	1.17	0.86	1.16	1.53	1.20	1.42	1.20	1.08	1.33	0.98
6	0.50	0.53	0.48	0.47	0.23	**0.38**	3.42	2.38	2.53	2.47	2.24	2.99	2.46	2.65	3.38
7	0.44	0.08	0.51	0.60	0.27	0.26	**0.47**	2.37	2.43	2.32	2.06	3.61	2.92	3.35	3.98
8	0.46	0.28	0.54	0.44	0.48	0.54	0.35	**0.73**	2.85	2.32	2.71	2.38	2.50	2.45	2.54
9	0.26	0.30	0.37	0.57	0.60	0.37	0.56	0.36	**0.72**	2.54	3.05	2.82	2.82	3.02	2.82
10	0.38	0.29	0.31	0.28	0.51	0.52	0.35	0.40	0.68	**0.45**	2.22	2.74	2.25	3.43	2.61
11	0.37	0.53	0.49	0.63	0.51	0.25	0.20	0.25	0.34	0.27	**0.70**	2.31	2.32	2.40	2.22
12	0.32	0.27	0.25	0.26	0.22	0.66	0.38	0.39	0.54	0.39	0.62	**0.43**	2.99	3.18	3.85
13	0.53	0.55	0.68	0.43	0.44	0.52	0.41	0.71	0.62	0.59	0.72	0.58	**0.47**	2.31	2.91
14	0.41	0.56	0.40	0.44	0.34	0.39	0.44	0.61	0.45	0.47	0.37	0.42	0.42	**0.31**	3.27
15	0.59	0.51	0.62	0.56	0.42	0.44	0.36	0.39	0.63	0.60	0.73	0.47	0.60	0.64	**0.39**

TABLE 18.
Residuals: NIDA model (diagonal—proportion correct; upper triangle—log-odds ratio; lower triangle—correlation).

Item	Item														
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	**−0.17**	0.10	1.16	0.96	0.44	0.93	0.53	−0.26	0.64	1.01	0.36	0.59	0.69	1.06	0.52
2	0.11	**0.09**	−0.59	−0.52	−0.84	−0.04	0.26	−0.53	−0.41	−0.49	−0.37	−0.01	−0.62	0.07	0.06
3	0.03	0.23	**0.05**	0.83	0.64	0.32	0.57	0.72	1.39	0.65	1.15	0.74	0.75	0.73	0.37
4	0.13	0.19	0.18	**0.03**	−0.01	−0.22	−0.17	0.19	0.85	0.34	0.32	−0.09	−0.32	0.23	−0.25
5	−0.06	0.11	0.26	0.08	**−0.14**	0.00	−0.34	0.47	0.56	−0.12	0.46	0.03	−0.12	0.00	−0.22
6	0.17	0.20	0.17	0.14	−0.10	**−0.06**	0.89	0.55	0.91	−0.04	0.62	0.48	−0.06	0.14	0.83
7	−0.10	−0.13	−0.04	0.04	−0.07	−0.10	**0.04**	0.58	0.80	−0.19	0.45	1.07	0.39	0.85	1.42
8	−0.09	−0.08	−0.02	−0.12	−0.06	−0.03	0.05	**0.02**	1.05	0.50	0.94	0.61	0.69	0.60	0.72
9	0.14	−0.03	0.04	0.10	0.22	0.06	0.17	0.03	**0.11**	0.85	1.43	1.21	1.19	1.34	1.17
10	0.05	−0.01	−0.02	0.06	−0.03	−0.03	0.02	0.05	0.04	**0.07**	0.55	0.20	−0.25	0.38	0.08
11	0.02	−0.01	−0.05	0.00	−0.04	0.04	−0.02	0.12	0.14	0.05	**0.08**	0.73	0.72	0.75	0.59
12	0.12	0.05	0.03	0.04	0.00	0.10	0.02	0.03	−0.01	0.03	0.07	**−0.02**	0.47	0.65	1.32
13	−0.03	0.01	0.12	0.08	0.08	−0.02	0.05	0.15	0.07	0.05	0.16	0.19	**0.03**	−0.18	0.36
14	0.08	0.17	0.06	0.09	0.01	0.05	0.08	0.24	0.09	0.11	0.02	0.06	0.07	**−0.06**	0.74
15	0.05	−0.03	−0.02	0.01	0.07	0.09	0.01	0.03	0.07	0.06	0.18	−0.07	0.04	0.09	**−0.05**

TABLE 19.
Residuals: single-strategy DINA model (diagonal—proportion correct; upper triangle—log-odds ratio; lower triangle—correlation).

Item	Item														
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	**0.00**	1.13	0.66	2.10	0.37	1.82	1.32	0.22	0.52	2.15	0.51	1.37	1.69	2.02	1.33
2	0.25	**0.00**	−0.19	0.25	−0.26	0.00	0.17	−0.25	−0.25	0.30	−0.03	−0.12	−0.20	0.29	−0.14
3	0.05	0.37	**0.00**	1.29	0.09	0.12	0.39	0.23	0.13	1.22	0.23	0.41	1.03	0.62	−0.08
4	0.09	0.31	0.27	**0.00**	0.78	0.33	0.09	0.57	1.12	0.94	0.80	0.25	0.35	0.32	0.11
5	0.04	0.08	0.41	0.09	**0.00**	0.44	0.10	0.05	0.11	0.66	0.16	0.41	0.45	0.65	0.20
6	0.26	0.34	0.29	0.24	−0.03	**0.00**	0.14	0.17	0.49	0.48	0.58	−0.08	−0.05	0.32	0.38
7	0.04	−0.06	0.00	0.03	−0.05	−0.05	**0.00**	0.22	0.40	0.09	0.37	0.03	0.17	0.37	0.27
8	0.06	−0.01	−0.02	−0.04	0.03	−0.02	0.12	**0.00**	0.20	0.93	0.37	0.19	0.81	0.34	0.06
9	0.02	0.01	0.04	0.03	0.02	0.14	0.03	0.03	**0.01**	1.19	0.01	0.66	1.22	1.07	0.55
10	0.11	0.03	−0.01	0.16	0.06	0.01	0.07	0.13	0.14	**0.00**	1.08	0.56	0.45	0.20	0.41
11	0.12	0.05	0.07	0.04	0.02	0.09	0.02	0.01	0.02	0.14	**0.00**	0.55	0.96	0.85	0.41
12	0.03	0.09	0.10	0.11	0.04	0.01	0.01	0.04	0.09	0.07	−0.01	**0.00**	0.26	0.46	0.41
13	−0.01	0.05	0.05	0.03	0.05	0.02	0.06	0.00	0.03	0.03	0.02	0.03	**0.00**	0.05	0.14
14	0.13	0.06	0.02	0.10	0.02	0.00	0.16	0.00	0.06	0.15	0.06	0.04	0.17	**0.00**	0.68
15	0.10	0.09	0.02	0.07	0.07	0.15	0.08	0.05	0.04	0.06	0.05	0.01	0.02	0.09	**0.00**

TABLE 20.
Residuals: multiple-strategy DINA model (diagonal—proportion correct; upper triangle—log-odds ratio; lower triangle—correlation).

Item	Item														
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	**0.00**	1.15	0.63	2.10	0.41	1.83	1.34	0.21	0.57	1.86	0.56	1.38	1.70	2.04	1.35
2	0.25	**0.00**	−0.20	0.23	−0.26	0.02	0.14	−0.20	−0.24	0.28	−0.01	−0.14	−0.18	0.35	−0.11
3	0.04	0.37	**0.00**	1.26	0.16	0.07	0.35	0.21	0.20	0.47	0.28	0.34	1.00	0.57	−0.13
4	0.10	0.31	0.27	**0.00**	0.78	0.31	0.07	0.57	1.12	0.97	0.79	0.22	0.36	0.64	0.08
5	0.04	0.09	0.34	0.10	**0.00**	0.44	0.07	0.11	0.12	0.46	0.19	0.42	0.43	0.67	0.21
6	0.26	0.34	0.29	0.25	−0.03	**0.00**	0.07	0.24	0.49	0.41	0.60	−0.10	−0.03	0.33	0.38
7	0.04	−0.06	0.00	0.02	−0.04	−0.05	**0.00**	0.21	0.39	0.02	0.36	−0.02	0.14	0.36	0.20
8	0.05	0.00	−0.02	−0.03	0.04	−0.02	0.12	**0.00**	0.27	0.23	0.42	0.21	0.83	0.41	0.11
9	0.03	0.01	0.04	0.03	0.03	0.04	0.04	0.03	**0.00**	0.57	0.05	0.66	1.23	1.11	0.55
10	0.11	0.03	−0.01	0.16	0.06	0.01	0.07	0.13	0.15	**−0.01**	0.60	0.49	0.40	0.26	0.31
11	0.12	0.04	0.07	0.09	0.01	0.09	0.02	0.02	0.03	0.10	**0.00**	0.57	0.98	0.87	0.42
12	0.04	0.09	0.09	0.12	0.04	0.01	0.02	0.05	0.07	0.07	−0.01	**0.00**	0.31	0.47	0.39
13	0.00	0.05	0.05	0.03	0.05	0.00	0.06	0.00	0.02	0.03	0.02	0.05	**0.00**	0.03	0.08
14	0.03	0.07	0.02	0.11	0.03	0.01	0.06	0.01	0.06	0.15	0.07	0.04	0.09	**0.00**	0.61
15	0.08	0.08	0.03	0.05	0.08	0.15	0.09	0.05	0.05	0.06	0.05	0.00	0.01	0.08	**0.00**

TABLE 21.
Standard errors of residuals: NIDA model (diagonal—proportion correct; upper triangle—log-odds ratio; lower triangle—correlation).

Item	Item														
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	**0.01**	0.05	0.08	0.06	0.04	0.05	0.05	0.07	0.05	0.06	0.05	0.05	0.05	0.06	0.05
2	0.01	**0.01**	0.05	0.05	0.06	0.05	0.05	0.05	0.04	0.05	0.04	0.04	0.05	0.05	0.05
3	0.01	0.01	**0.01**	0.06	0.04	0.05	0.05	0.07	0.05	0.06	0.05	0.05	0.05	0.06	0.05
4	0.01	0.01	0.01	**0.01**	0.07	0.05	0.05	0.07	0.05	0.07	0.05	0.05	0.05	0.07	0.05
5	0.01	0.01	0.01	0.01	**0.01**	0.06	0.06	0.04	0.05	0.06	0.05	0.05	0.06	0.06	0.06
6	0.01	0.01	0.01	0.01	0.01	**0.01**	0.05	0.05	0.04	0.05	0.04	0.05	0.05	0.05	0.05
7	0.01	0.01	0.01	0.01	0.01	0.01	**0.01**	0.05	0.04	0.05	0.04	0.05	0.05	0.05	0.05
8	0.01	0.01	0.01	0.01	0.01	0.01	0.01	**0.01**	0.05	0.06	0.05	0.05	0.05	0.06	0.05
9	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	**0.01**	0.05	0.05	0.04	0.04	0.05	0.04
10	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	**0.01**	0.05	0.05	0.05	0.08	0.05
11	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	**0.01**	0.04	0.04	0.05	0.04
12	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	**0.01**	0.05	0.05	0.05
13	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	**0.01**	0.05	0.05
14	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	**0.01**	0.05
15	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	**0.01**

TABLE 22.
Standard errors of residuals: single-strategy DINA model (diagonal—proportion correct; upper triangle—log-odds ratio; lower triangle—correlation).

Item	Item														
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	**0.01**	0.05	0.15	0.04	0.06	0.06	0.06	0.10	0.10	0.05	0.10	0.06	0.05	0.05	0.06
2	0.01	**0.01**	0.06	0.07	0.05	0.10	0.10	0.07	0.06	0.08	0.07	0.10	0.09	0.10	0.11
3	0.02	0.01	**0.01**	0.07	0.08	0.11	0.08	0.13	0.14	0.07	0.13	0.11	0.07	0.11	0.12
4	0.01	0.01	0.01	**0.01**	0.04	0.07	0.08	0.07	0.06	0.09	0.06	0.07	0.07	0.14	0.07
5	0.02	0.02	0.01	0.02	**0.01**	0.06	0.06	0.08	0.10	0.04	0.09	0.06	0.05	0.05	0.06
6	0.01	0.01	0.01	0.01	0.01	**0.01**	0.12	0.13	0.10	0.08	0.09	0.10	0.08	0.08	0.09
7	0.01	0.01	0.02	0.02	0.01	0.01	**0.01**	0.10	0.08	0.08	0.08	0.10	0.10	0.11	0.12
8	0.02	0.01	0.02	0.02	0.02	0.02	0.01	**0.01**	0.11	0.07	0.11	0.11	0.08	0.13	0.15
9	0.02	0.01	0.01	0.02	0.02	0.01	0.02	0.01	**0.01**	0.07	0.12	0.10	0.07	0.10	0.12
10	0.01	0.01	0.01	0.01	0.01	0.02	0.01	0.01	0.02	**0.01**	0.06	0.09	0.07	0.14	0.08
11	0.01	0.01	0.01	0.02	0.01	0.01	0.01	0.02	0.02	0.01	**0.01**	0.10	0.07	0.09	0.10
12	0.02	0.01	0.01	0.01	0.01	0.02	0.02	0.01	0.02	0.01	0.01	**0.01**	0.09	0.09	0.11
13	0.01	0.01	0.01	0.01	0.01	0.02	0.01	0.01	0.02	0.02	0.01	0.02	**0.01**	0.08	0.09
14	0.01	0.02	0.01	0.01	0.01	0.01	0.01	0.02	0.01	0.01	0.01	0.01	0.01	**0.01**	0.09
15	0.02	0.01	0.02	0.02	0.02	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.02	**0.01**

TABLE 23.
Standard errors of residuals: Multiple-strategy DINA model (diagonal—proportion correct; upper triangle—log-odds ratio; lower triangle—correlation).

Item	Item														
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	**0.01**	0.05	0.14	0.04	0.05	0.05	0.06	0.08	0.10	0.06	0.09	0.06	0.05	0.05	0.05
2	0.01	**0.01**	0.07	0.08	0.04	0.11	0.10	0.07	0.07	0.09	0.07	0.10	0.09	0.09	0.11
3	0.01	0.01	**0.01**	0.07	0.07	0.11	0.10	0.13	0.13	0.14	0.12	0.11	0.08	0.12	0.13
4	0.01	0.01	0.01	**0.01**	0.04	0.06	0.08	0.07	0.07	0.10	0.06	0.07	0.07	0.10	0.06
5	0.02	0.02	0.01	0.02	**0.01**	0.05	0.06	0.06	0.08	0.06	0.08	0.06	0.05	0.05	0.05
6	0.01	0.01	0.01	0.01	0.01	**0.01**	0.12	0.12	0.11	0.09	0.09	0.10	0.09	0.08	0.10
7	0.01	0.01	0.02	0.02	0.01	0.01	**0.01**	0.11	0.10	0.10	0.09	0.12	0.11	0.11	0.13
8	0.02	0.01	0.01	0.02	0.01	0.02	0.01	**0.01**	0.09	0.13	0.10	0.12	0.08	0.11	0.14
9	0.01	0.01	0.01	0.02	0.02	0.02	0.02	0.01	**0.01**	0.13	0.12	0.11	0.08	0.10	0.11
10	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.02	**0.01**	0.10	0.10	0.09	0.14	0.10
11	0.01	0.01	0.01	0.02	0.01	0.01	0.01	0.01	0.02	0.01	**0.01**	0.09	0.07	0.08	0.09
12	0.02	0.01	0.01	0.01	0.01	0.02	0.01	0.01	0.02	0.01	0.01	**0.01**	0.11	0.09	0.11
13	0.02	0.01	0.02	0.02	0.01	0.02	0.01	0.01	0.02	0.01	0.01	0.02	**0.01**	0.09	0.10
14	0.02	0.02	0.01	0.01	0.01	0.01	0.02	0.02	0.01	0.01	0.01	0.01	0.02	**0.01**	0.09
15	0.02	0.02	0.02	0.02	0.01	0.01	0.01	0.01	0.02	0.01	0.01	0.02	0.01	0.01	**0.01**

DINA model for item 10, $2 - \frac{1}{3}$, can be traced to specification of more attributes than necessary to solve this item (i.e., using only attributes 1, 4, and 5 without attribute 3 may be sufficient for this item). Alternatively, this may be an indication that use of an alternative strategy may provide a better fit for this item.

Further comparison of the single- and multiple-strategy DINA models is given in Table 24. Essentially identical values (i.e., maximum difference is 0.01) were obtained when these statistics were computed using posterior predictive checks. The table compares the different models using all the 15 items. With all the items included, the mean and maximum absolute residuals for the log-odds ratio and correlation show that the multiple-strategy DINA model provides a slightly better fit than the single-strategy DINA model. This minor difference is due mainly to differences in the parameter estimates of item 10. When this item is excluded in the computation, the mean absolute residuals for both models are virtually identical. Although the NIDA model does not fit the data adequately, its mean and maximum absolute residuals are not always inferior to those of the DINA models. This can be attributed to the presence of item 1 that cannot be fit well by any of the models using the current definitions of the attributes. Excluding this item dramatically lowers the mean and maximum absolute residuals of the DINA models while keeping those of the NIDA model relatively unchanged. Thus, the first- and second-moment statistics can be used to unequivocally choose the DINA models over the NIDA model.

As previously noted, the standardized residuals obtained from dividing residuals in Tables 19 and 20 by their corresponding standard errors in Tables 22 and 23, respectively, indicate significant departures from 0 in most cases. Nevertheless, the absolute magnitude of the unstandardized residuals provide confidence that the relatively parsimonious versions of the higher-order DINA model can do an adequate job of explaining first and second moments. By inspecting such residuals one can distinguish between poor models and promising models in a very similar fashion as is done in confirmatory factor analysis. Of course, with binary data, unlike data arising from linear factor analysis models with normal errors, first and second moments are not sufficient to describe the joint distribution of the item responses. One could consider residuals based on higher moments. However, at this stage of model scrutiny, a more efficient approach is to use tools developed for inference and model assessment using the full information available through the likelihood function and posterior distribution.

Table 24 also shows the log-marginal likelihoods, BIC, AIC, and DIC of the three models. The large differences between the log-marginal likelihood, BIC, AIC, and DIC of the NIDA model and log-marginal likelihoods of the single- and multiple-strategy DINA models indicate an overwhelming evidence for the DINA models over the NIDA model, validating the observation above. Although parameter estimates are similar for the single- and multiple-strategy DINA models, a difference exists between the log-marginal likelihoods, BIC, AIC, and DIC of the two models. Whereas comparison of the two models based on the parameter estimates alone resulted in very similar log-conditional likelihoods (i.e., log-likelihoods when the incidental parameters are integrated out), comparison of the models in their entirety (i.e., the joint posterior distribution of the parameters) or comparison of the models that accounts for their dimensionality and complexity resulted in log-marginal likelihood difference and criteria that indicate a consistent evidence for the single-strategy DINA model. It can be noted that the Bayes factor is found by exponentiating the difference between the log-marginal likelihoods. This is consistent with the results one would arrive at using the BIC and AIC where the dimensions of the models are taken into account. The Schwarz criterion, which can be computed from the BIC, is one rough approximation of the Bayes factor (Kass, 1993; Kass & Raftery, 1995). In this example, the Bayes factors computed using the approximate log-marginal likelihood or approximated by the Schwarz criterion lead to the same conclusion. To the extent that BIC approximates the Bayes factor, the difference of 15.42 in the BICs of the single- and multiple-strategy DINA models indicate a strong evidence for the former based on the interpretation suggested by Raftery (1996).

TABLE 24.
Log-marginal likelihood and residual summary statistics.

		NIDA	DINA-S	DINA-M
	−Log-marginal likelihood	21333.47	20927.41	20930.16
	BIC	42676.93	41873.38	41888.80
	AIC	42586.20	41669.24	41673.32
	DIC	39552.72	33367.15	35886.03
Mean absolute residual	Proportion	0.07	0.00	0.00
	Log-odds ratio	0.56	0.51	0.49
	Correlation	0.08	0.08	0.07
Maximum absolute residual	Proportion	0.17	0.01	0.01
	Log-odds ratio	1.43	2.15	2.10
	Correlation	0.26	0.41	0.37

6. Discussion

Applications of cognitive diagnosis, as in the case of the fraction subtraction data, are aimed at diagnosing mastery of fine-grained skills rather than measuring general abilities. Cognitive diagnosis models differ in how they describe the process by which skills are applied to respond to items. These competing theories that define models suggest that some models may result in a better model fit than other models. The goodness of fit of models for cognitive diagnosis has not been adequately discussed in the literature. By viewing cognitive diagnosis models as analogous to models for confirmatory factor analysis, with discrete latent variables and discrete latent responses, we can employ simple techniques based on residuals of first and second moments to investigate model adequacy. In addition, the Bayesian approach to model fitting with MCMC naturally leads to straightforward computation of standard errors of residuals and global measures of relative model fit such as the Bayes factor, AIC, BIC, and DIC. This paper illustrates how the first and second moments can be used not only to identify items that cannot be fit adequately but, in some instances, to also choose among several competing models. Models that provide very disparate fits can be differentiated using the first moment alone, whereas models that provide similar fits can be further differentiated using the second moment. However, because moments are computed using solely the parameter estimates, equivocal results may be obtained for models that provide similar estimates but differ in other respects. Hence, there is a need for more global measures such as the Bayes factor, AIC, BIC, and DIC that take into account the models in their entirety. As the example shows, despite the similar parameter estimates, the global measures differentiated between the single- and multiple-strategy DINA models based on the differences in their joint posterior distributions and dimensionality. It can be argued that for the example discussed here, even if the global measures do not provide any evidence for either DINA model that the single-strategy DINA model is to be preferred over the multiple-strategy DINA model based on principle of parsimony.

An interesting theoretical issue that arises in a Bayesian analysis is whether the AIC is appropriate when the posterior-mean estimator derived from MCMC is plugged in, in place of the marginal maximum likelihood estimator. Spiegelhalter et al. (2002), proposed the DIC as a preferred criterion for expressing model complexity and fit when conducting a Bayesian MCMC analysis. Like the AIC, the DIC includes a term for fidelity to the data and a penalty term for model complexity. The goodness-of-fit term is simply the deviance measure integrated over the posterior distribution of the model parameters. The penalty term is the difference between this mean deviance and the deviance calculated at the posterior mean of the model parameters. In

hierarchical Bayesian models both the goodness-of-fit and penalty terms of the DIC depend on which parameters are considered focal to the model and which are integrated out of the posterior.

The dependence of the DIC on the choice of the focus has drawn some criticism, especially when used with hierarchical models such as random effects models and models like those considered here. In the cognitive diagnosis model that we consider, estimation of the latent attribute $\boldsymbol{\alpha}$ can be considered a focus. Thus, in computing the DIC, all structural parameters, parameters of the high-order latent trait model, and $\boldsymbol{\alpha}$ are included. However, because the subject-specific parameters can alternatively be viewed as missing data, and because viewing them as model parameters can lead to uncertain asymptotics, we employed a variation of the DIC proposed by Celeux et al. (2006), which they labeled DIC_4. The DIC_4 computes the DIC for the complete data, but integrates this over the posterior distribution of the missing data. Of the many variations of the DIC they considered for missing data, this appeared to achieve the best performance.

The discussion following Spiegelhalter et al. (2002) was far from unanimous about the theoretical justification and interpretation of the DIC. However, no clear justification for plugging in the posterior mean and using the AIC appears present in the literature, leaving the practitioner with a couple of choices but no definite answers. A suggested interpretation is that marginal likelihood analysis with the AIC is appropriate when we are concerned primarily with the structural parameters and predicting characteristics of the population. However, the DIC is more appropriate when we are concerned with individuals and "random effects" such as $\boldsymbol{\alpha}$ and θ, particularly when making adjustments such as with the DIC_4. An additional index, BIC, can also be used and differs in interpretation from the DIC in that it is primarily concerned with identification of the correct Bayesian model, but not in predictions concerning the incidental parameters. The practice of plugging in the posterior mean estimator in the BIC is also not clearly justified in the literature.

In this paper, we considered three models derived from different theories for responding to items, and analyzed how appropriate they were for fitting the fraction subtraction data. The higher-order NIDA model assumes that the probabilities of slipping and guessing are constant across all items for each attribute. This assumption may be a reasonable approximation, but we saw that it resulted in poor item fit statistics for a few items. The higher-order single-strategy DINA model assumes deterministic latent responses, and slips and guesses at the item level with probabilities of slipping and guessing allowed to vary across the items. In the fraction subtraction data, this provided the best fit among the models considered. Despite the simplicity of the model, its goodness of fit under the various indices was better than the two competing models.

The higher-order, multiple-strategy DINA model allows for several different methods of solution for each item. This model has deterministic latent responses after it is checked whether or not an examinee's attribute pattern satisfies at least one of the strategies. Slipping and guessing probabilities are constant for all strategies, but are allowed to vary across the items. Allowing for multiple strategies in the fraction subtraction data did not improve the fit compared with the single-strategy DINA model. This may indicate that a single dominant strategy suffices for this problem, or the somewhat poorer fit of the multiple-strategy DINA model may indicate that the assumption of equal slipping and guessing probabilities for each strategy is too restrictive.

A variant of the multiple-strategy DINA model is the mixture model given in (5). This approach has the advantage that, if the model parameters are identifiable, one can obtain posterior probabilities in order to classify ω into a particular strategy. Although, we have considered this approach, we have not yet resolved the necessary identifiability issues to be confident in fitting it, especially when the slip and guessing parameters are allowed to vary for different strategies.

The higher-order latent trait approach to modeling the joint distribution of the attributes is appropriate whenever it is reasonable to assume the existence of a general ability in addition to fine-grained skills. This is arguably the case for the fraction subtraction data. We have developed algorithms to fit the higher-order component for several cognitive diagnosis models. The simulation study given here and in de la Torre and Douglas (2004) indicate that these algorithms

implemented with the software Ox (Doornik, 2003) can be used to obtain accurate parameter estimates.

Finally, in our analyses we just considered fixed $\boldsymbol{Q}$-matrices, rather than conduct an exploratory search for the correct one or consider a handful of other expert-chosen candidates. In applications it is certainly possible that content experts who identify the attributes and the particular attributes or strategies employed by examinees on certain items may differ from one another. This could yield a handful of choices that would need to be sorted through. Models corresponding to each of the $\boldsymbol{Q}$-matrices could be evaluated under precisely the same model fit statistics we propose, and those that seem to fit could be evaluated under global statistics to arrive at a final model and a corresponding $\boldsymbol{Q}$-matrix.

However, there is one complication that concerns us when using such techniques to select one among several $\boldsymbol{Q}$-matrices. Indices such as the AIC and BIC penalize for complexity, though the complexity might not be what it seems when a $\boldsymbol{Q}$-matrix is seen as fixed and not a model parameter. Different $\boldsymbol{Q}$-matrices actually result in different levels of complexity because of the way they define equivalence classes of attributes for different items. If some sort of $\boldsymbol{Q}$-matrix estimation is to be done rather formally, the issue of complexity of a $\boldsymbol{Q}$-matrix would need to be addressed so that its complexity could be quantified accurately. Notwithstanding this complication, the $\boldsymbol{Q}$-matrix, which is an integral component of a model specification in cognitive diagnosis, cannot be assumed correct after its construction, and should be subjected to empirical scrutiny (de la Torre, 2008). In future studies, it would be worthwhile to investigate how different specifications of the $\boldsymbol{Q}$-matrix interact with various CMDs, and the extent to which these interactions affect different model-data fit indices.

References

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

Brooks, S.P., & Gelman, A. (1998). General methods of monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, *7*, 434–455.

Casella, G., & George, E.I. (1992). Explaining the Gibbs sampler. *The American Statistician*, *46*, 167–174.

Celeux, G., Forbers, F., Robert, C.P., & Titterington, D.M. (2006). Deviance information criteria for missing data models. *Bayesian Analysis*, *1*, 651–674.

Chib, S., & Greenberg, E. (1995). Understanding the Metropolis-Hastings algorithm. *The American Statistician*, *49*, 327–335.

de la Torre, J. (2008, in press). An empirically-based method of $\boldsymbol{Q}$-matrix validation for the DINA model: Development and applications. *Journal of Educational Measurement*.

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

DiBello, L.V., Stout, W.F., & Roussos, L.A. (1995). Unified cognitive/psychometric diagnostic assessment likelihood-based classification techniques. In P.D. Nichols, S.F. Chipman, & R.L. Brennan (Eds.), *Cognitively diagnostic assessment* (pp. 361–389). Hillsdale, NJ: Erlbaum.

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

Doornik, J.A. (2003). *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.

Embretson, S. (1997). Multicomponent response models. In W.J. van der Linden & R.K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 305–321). New York: Springer.

Geman, S., & Geman, D. (1984). Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *6*, 721–741.

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

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

Kass, R.E. (1993). Bayes factor in practice. *The Statistician*, *42*, 551–560.

Kass, R.E., & Raftery, A.E. (1995). Bayes factor. *Journal of the American Statistical Association*, *430*, 773–795.

LeFevre, J., Bisanz, J., Daley, K., Buffone, L., Greenham, S.L., & Sadesky, G.S. (1996). Multiple routes to solution of single-digit multiplication problems. *Journal of Experimental Psychology: General*, *125*, 284–306.

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.J. (1996). Test theory reconceived. *Journal of Education al Measurement*, *33*, 379–416.
Opfer, J.E., & Siegler, R.S. (2007). Representational change and children's numerical estimation. *Cognitive Psychology*, *55*, 169–195.
Patz, R.J., & Junker, B.W. (1999a). A straightforward approach to Markov chain Monte Carlo methods for item response theory. *Journal of Educational and Behavioral Statistics*, *24*, 146–178.
Patz, R.J., & Junker, B.W. (1999b). Applications and extensions of MCMC in IRT: Multiple item types, missing data, and rated responses. *Journal of Educational and Behavioral Statistics*, *24*, 342–366.
Raftery, A.E. (1996). Hypothesis testing and model selection. In R.W. Gilks, S. Richardson, & D.J. Spiegelhalter (Eds.), *Markov chain Monte Carlo in practice* (pp. 163–187). London: Chapman & Hall.
Reder, L.M. (1987). Strategy selection in question answering. *Cognitive Psychology*, *19*, 90–138.
Shultz, T., Fisher, G., Pratt, C., & Rulf, S. (1986). Selection of causal rules. *Child Development*, *57*, 143–152.
Siegler, R.S. (1988). Individual differences in strategy choices: Good students, not-so-good students, and perfectionists. *Child Development*, *59*, 833–851.
Siegler, R.S., Adolph, K.E., & Lemaire, P. (1996). Strategy choices across the lifespan. In L. Reder (Ed.), *Implicit memory and metacognition* (pp. 79–121). Hillsdale, NJ: Erlbaum.
Siegler, R.S., & Shrager, J. (1984). Strategy choices in addition and subtraction: How do children know what to do? In C. Sophian (Ed.), *The origins of cognitive skills* (pp. 229–293). Hillsdale, NJ: Erlbaum.
Spiegelhalter, D.J., Best, N.G., Carlin, B.P., & van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society Series B*, *64*, 583–639.
Stout, W., Roussos, L., & Hartz, S. (2003). *A demonstration of the Fusion Model skills diagnostic system: An analysis of mixed-number subtraction data*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Chicago, IL.
Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, *6*, 461–464.
Tatsuoka, K. (1985). A probabilistic model for diagnosing misconceptions in the pattern classification approach. *Journal of Educational Statistics*, *12*, 55–73.
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. (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, NJ: Erlbaum.
Yan, D., Almond, R., Mislevy, R., & Simpson, M.A. (2003). *A rule space approach to modeling skills-pattern of two different pedagogical groups in mathematics assessment*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Chicago, IL.

Manuscript Received: 1 MAR 2004
Final Version Received: 5 FEB 2008
Published Online Date: 29 MAR 2008