# Test Construction for Cognitive Diagnosis

Robert Henson and Jeff Douglas

*Applied Psychological Measurement* 2005 29: 262
DOI: 10.1177/0146621604272623

# Test Construction for Cognitive Diagnosis

**Robert Henson and Jeff Douglas, University of Illinois**

Although cognitive diagnostic models (CDMs) can be useful in the analysis and interpretation of existing tests, little has been developed to specify how one might construct a good test using aspects of the CDMs. This article discusses the derivation of a general CDM index based on Kullback-Leibler information that will serve as a measure of how informative an item is for the classification of examinees. The effectiveness of the index is examined for items calibrated using the deterministic input noisy "and" gate model (DINA) and the reparameterized unified model (RUM) by implementing a simple heuristic to construct a test from an item bank. When compared to randomly constructed tests from the same item bank, the heuristic shows significant improvement in classification rates.
*Index terms: cognitive diagnosis, cognitive diagnostic index, Kullback-Leibler information, test construction.*

Now, more than ever, there is a focus on measuring students' abilities to ensure appropriate grade placement and quality of education. However, unlike measurements that are made in the physical world, such as length and weight, the social sciences have always been challenged with measuring attributes that cannot be directly observed. Instead of measuring an attribute such as ability directly, the social sciences must measure the attribute based on a set of observable responses that are indicators of the attribute (Lord & Novick, 1968). Classical test theory (CTT) and item response theory (IRT) provide methods for developing instruments to measure constructs in the social sciences such as extraversion. In addition, CTT and IRT both provide methods for obtaining an examinee's score.

As an alternative, there may be instances when the estimation of an examinee's score is not the focus. For example, a grade school teacher may be interested in estimating students' profiles. The profile for each student specifies a set of dichotomous skills, or attributes, that a student has or has not mastered. A profile of discrete attributes provides the teacher with information about the instructional needs of groups of students (unlike multidimensional IRT, which provides a profile of scores). Cognitive diagnostic models (CDMs) can be used when the interest of a test is to estimate students' profiles, or attribute mastery patterns, instead of providing a general estimate of ability. Unlike CTT and IRT, CDMs are a special case of latent class models. Specifically, CDMs model the probability of correctly answering an item as a function of an attribute mastery pattern. Note that the terminology used here refers to mastery of binary skills or pieces of knowledge, generically referred to as attributes. However, the proposed methodology is equally applicable when the binary components of the attribute vector refer to the presence or absence of conditions, which might be related to medical, psychological, or psychiatric constructs. For instance, an application in


*Applied Psychological Measurement,* Vol. 29 No. 4, July 2005, 262–277
DOI: 10.1177/0146621604272623

psychiatry might involve an assessment to classify patients as having or not having particular anxiety disorders such as social anxiety disorder, generalized anxiety disorder, and panic disorders.

Because estimation of the mastery pattern no longer involves a continuous measure of ability, concepts initially introduced by CTT and IRT, such as reliability and information, do not apply. In this setting, few methods for test construction exist. The objective of this article is to discuss the concept of reliability, or *discrimination*, for CDMs to describe the ability of a test to distinguish among examinees' attribute patterns of mastery. In addition, given a measure of cognitive diagnostic discrimination, methods of test construction will be discussed for conjunctive CDMs that assume a discrete latent attribute space with invariant item parameters. As examples, the proposed method of test construction will be applied to items parameterized by the deterministic input noisy "and" gate model (DINA) and to items parameterized by the reparameterized unified model (RUM).

## Models for Cognitive Diagnosis

Many models have been developed for CDMs, each with their advantages and disadvantages. In this section, three particular models are discussed. All these models require specification of a $Q$-matrix (K. Tatsuoka, 1985). Given $J$ items and $K$ attributes, the $Q$-matrix has elements $q_{jk}$ that indicate whether mastery of the $k$th attribute is required by the $j$th item, written as

$$q_{jk} = \begin{cases} 1 & \text{if item } j \text{ requires attribute } k \\ 0 & \text{else.} \end{cases}$$

In addition, $\boldsymbol{\alpha}_i$ is defined as a vector of indicators for subject $i$'s attribute mastery for the $k$ attributes. The first model discussed is the DINA model. In the DINA model, items divide the population into two classes, those who have all required attributes and those who do not. Let $\xi_{ij}$ be an indicator of whether examinee $i$ has mastered all of the required attributes for item $j$,

$$\xi_{ij} = \prod_{k=1}^{K} \alpha_{ik}^{q_{jk}}.$$

Here, $\boldsymbol{\alpha}_i$ is a 0/1 vector such that the $k$th element indicates mastery or nonmastery of the $k$th attribute. Given $\xi_{ij}$, only two item parameters, $s_j$ and $g_j$, are required to model the probability of a correct response for an examinee. Equation (1) defines $s_j$ as the probability that an examinee misses an item when, in fact, he or she has all of the required attributes, a "slipping" parameter. As defined in equation (2), $g_j$ represents the probability that an individual gets the correct answer when he or she does not have all of the required attributes for that item, a "guessing" parameter.

$$s_j = P(X_{ij} = 0 | \xi_{ij} = 1), \tag{1}$$

$$g_j = P(X_{ij} = 1 | \xi_{ij} = 0). \tag{2}$$

Here, $X_{ij}$ is the response for the $i$th examinees for the $j$th item. Given the $j$th item's parameters and $\xi_{ij}$, the probability of a correct response can be written as

$$P(X_{ij} = 1 | \xi_{ij}, s_j, g_j) = (1 - s_j)^{\xi_{ij}} g_j^{(1-\xi_{ij})}. \tag{3}$$

Junker and Sijtsma (2001), C. Tatsuoka (2002), and de la Torre and Douglas (in press) discuss methods for estimation of the DINA model, which implement Markov Chain Monte Carlo (MCMC) algorithms. The DINA model has also been considered by Macready and Dayton (1977), Haertel (1989), and Doignon and Falmagne (1999). As the latent class part of a mixture model,

in which other examinees follow a latent trait model, Yamamoto (1987) employs a similar formulation.

One concern is that the DINA model partitions the population into only two equivalence classes per item, which might be viewed as too simple. In the DINA model, missing one attribute is equivalent to missing all required attributes. However, in some situations, it is realistic to think that an examinee lacking only one of the required attributes may have a higher probability of a correct response when compared to an examinee lacking all of the required attributes. Next, models that recognize this possibility are considered.

Junker and Sijtsma (2001) discuss the NIDA model (noisy input; deterministic "and" gate) as a model with such a quality. The NIDA model accounts for different contributions from each attribute by defining a "slipping" parameter, $s_k$, and "guessing" parameter, $g_k$, for each specified attribute, independent of the item. The probability of a correct response is the probability that all attributes are correctly applied. Specifically, because all slipping and guessing parameters are at the attribute level, a new latent variable $\eta_{ijk}$ is defined at the attribute level, such that $\eta_{ijk}$ is 1 if attribute $k$ was correctly applied and 0 otherwise. Now $s_k$ and $g_k$ can be defined in terms of $\eta_{ijk}$, given the $Q$-matrix and examinee's attribute mastery as

$$s_k = P(\eta_{ijk} = 0|\alpha_{ik} = 1, q_{jk} = 1)$$

and

$$g_k = P(\eta_{ijk} = 1|\alpha_{ik} = 0, q_{jk} = 1).$$

Therefore, the probability of a correct response is equal to the probability that all required attributes are correctly used. The NIDA model defines the probability of a correct response as

$$P(X_{ij} = 1|\boldsymbol{\alpha}_i, \mathbf{s}, \mathbf{g}) = \prod_{k=1}^{K}[(1 - s_k)^{\alpha_{ik}} g_k^{1-\alpha_{ik}}]^{q_{jk}}.$$

Maris (1999) extends the NIDA model such that the "slip" parameters, $s_{jk}$, and "guessing" parameters, $g_{jk}$, are estimated separately for each item; therefore, the probability of a correct response is

$$P(X_{ij} = 1|\boldsymbol{\alpha}_i, \mathbf{s}, \mathbf{g}) = \prod_{k=1}^{K}[(1 - s_{jk})^{\alpha_{ik}} g_{jk}^{1-\alpha_{ik}}]^{q_{jk}}.$$

Finally, the reparameterized unified model (Hartz, 2002) is considered, which extends the NIDA model by incorporating a continuous latent variable $\theta_i$ to account for any attributes not otherwise specified in the $Q$-matrix (DiBello, Stout, & Roussos, 1995). In addition, the reparameterized unified model uses a parameterization that eliminates a source of unidentifiability present in Maris's (1999) extension of the NIDA model. In particular, to solve the identifiability problem, Hartz (2002) reparameterizes the unified model (DiBello et al., 1995) such that there is a parameter that defines the probability of getting an item correct, given that all required attributes have been mastered, denoted by $\pi_j^*$. Using the parameters of the extended NIDA model,

$$\pi_j^* = \prod_{k=1}^{K}(1 - s_{jk})^{q_{jk}}. \tag{4}$$

In addition, a "penalty" for each attribute that is not mastered for the $j$th item, $r_{jk}^*$, is defined as

$$r_{jk}^* = \frac{g_{jk}}{1 - s_{jk}}. \tag{5}$$

The reparameterized unified model allows for the possibility that not all required attributes have been explicitly specified in the $Q$-matrix by incorporating a general ability measure, $P_{c_j}(\theta_i)$.

Specifically, using the reparameterized unified model, the probability of a correct response can be written as

$$P(X_{ij} = 1|\boldsymbol{\alpha}_i, \theta_i) = \pi_j^* \prod_{k=1}^{K} r_{jk}^{*(1-\alpha_{ik})q_{jk}} P_{c_j}(\theta_i), \tag{6}$$

where $P_{c_j}$ is the Rasch model item characteristic curve with difficulty parameter $c_j$, and $\theta_i$ is a general measure of examinee $i$'s knowledge not otherwise specified by the $Q$-matrix.

Notice that for each required attribute not mastered, $\pi_j^*$ is reduced by a factor of $r_{jk}^*$. Estimation of the reparameterized unified model was accomplished by Hartz (2002) by casting it into a hierarchical Bayes model, called the fusion model, and applying an MCMC algorithm.

## Cognitive Diagnostic Information

To identify a method for test construction, it is useful to explore possible characteristics of indices that would indicate a good test. In some cases, it may be advantageous to generalize the indices used previously in educational assessment. However, in many ways, it does not make sense to generalize results from CTT. For example, there are no invariant item parameters in CTT, and Cronbach's $\alpha$, one of the most common measures in CTT, assumes unidimensionality. In CDMs, unidimensionality does not hold because there is a set of $k$ attributes that influence the outcome of the test. One last characteristic of CTT methods for defining a good test that creates some problems, as discussed previously, is that the reliability for a test is not a linear function of the reliability for each item. Although some suggestions have been developed to accommodate such complications, if one were to develop an index for test construction, it would be more convenient, if possible, to construct an index such that the test's index is a linear function of its item indices.

IRT often uses Fisher information as a criterion for test construction. The information in the test is the sum of the information of each item; therefore, optimization routines such as integer programming are easily implemented (van der Linden, 1998). As the information increases, the variability of the maximum likelihood estimate for an examinee's ability decreases. When considering criteria for test construction with CDMs, one should strive for similar characteristics.

### Kullback-Leibler Information

Kullback-Leibler information or, more appropriately, the Kullback-Leibler information for discrimination (Lehmann & Casella, 1998) is most commonly thought of as a measure of distance between any two probability distributions, $f(x)$ and $g(x)$. Formally, Kullback-Leibler information is defined as

$$d[f, g] = E_f\left[\log\left[\frac{f(x)}{g(x)}\right]\right], \tag{7}$$

where the measure $d[f, g]$ is equal to the expectation with respect to $f(x)$ of the log-likelihood ratio of any two probability density functions $f(x)$ and $g(x)$. Although $d[f, g]$ is sometimes referred to as a distance measure between two distributions, it is not a symmetric function (i.e., $d[f, g] \neq d[g, f]$). However, it does have some similar interpretations of a distance measure. Specifically, as $d[f, g]$ increases, it is easier to statistically discriminate between the two distributions (Lehmann & Casella, 1998; Rao, 1962). In addition, $d[f, g] \geq 0$ with equality when and only when $f$ equals $g$.

Kullback-Leibler information is not new to assessment. Madigan and Almond (1995) use the Kullback-Leibler information, which they call the expected weight of evidence, for strategies of test selection for belief networks that can be directly applied to a decision-theoretic framework. Tatsouka and Ferguson (2003) use Kullback-Leibler information and Shannon entropy for sequential item selection with possible applications for computer-adaptive tests in CDMs. Chang and Ying (1996) suggest using Kullback-Leibler information instead of Fisher information as a more effective index for item selection in computer-adaptive tests based on unidimensional IRT models. Veldkamp and van der Linden (2002) also use Kullback-Leibler information as part of an integer programming approach to multidimensional IRT test construction. Kullback-Leibler information in IRT can be thought of as global information in which Fisher information is local. More specifically, Fisher information describes the ability to differentiate among abilities that are within a small range of one another. Specialized to the unidimensional IRT setting, Kullback-Leibler information is defined for all pairs $\theta$ and $\theta'$ (Chang & Ying, 1996). Unlike Fisher information, Kullback-Leibler information does not require that the parameter space is a continuum and is hence suitable for CDMs in which the attribute, $\boldsymbol{\alpha}$, is a discrete parameter. Therefore, this study intends to generalize Chang and Ying's results using Kullback-Leibler information as a basis for test construction with CDMs.

As in IRT, for the CDMs considered here, the item response, $X$, is a dichotomous variable (i.e., an examinee either gets the item right or wrong). In addition, the probability distribution of $X$, $P_{\boldsymbol{\alpha}}(X)$, depends on the pattern of attribute mastery, $\boldsymbol{\alpha}$, and therefore the results from Chang and Ying (1996) easily generalize to CDMs. According to Kullback-Leibler information, an item is most useful in determining the difference between an attribute mastery pattern, $\boldsymbol{\alpha}$, and an alternative attribute mastery pattern, $\boldsymbol{\alpha}^*$, if Kullback-Leibler information for the comparison of $P_{\boldsymbol{\alpha}}(X)$ and $P_{\boldsymbol{\alpha}^*}(X)$,

$$d_j[\boldsymbol{\alpha}, \boldsymbol{\alpha}^*] = E_{\boldsymbol{\alpha}}\left[\log\left[\frac{P_{\boldsymbol{\alpha}}(X_j)}{P_{\boldsymbol{\alpha}^*}(X_j)}\right]\right], \tag{8}$$

is large, where $P_{\boldsymbol{\alpha}}(X_j)$ and $P_{\boldsymbol{\alpha}^*}(X_j)$ are the probability distributions of $X_j$ conditional on $\boldsymbol{\alpha}$ and $\boldsymbol{\alpha}^*$, respectively.

Because $X$ is dichotomous, equation (8) can be written as

$$\sum_{x=0}^{1} P_{\boldsymbol{\alpha}}(x_j) \log\left[\frac{P_{\boldsymbol{\alpha}}(x_j)}{P_{\boldsymbol{\alpha}^*}(x_j)}\right],$$

namely,

$$P_{\boldsymbol{\alpha}}(1) \log\left[\frac{P_{\boldsymbol{\alpha}}(1)}{P_{\boldsymbol{\alpha}^*}(1)}\right] + P_{\boldsymbol{\alpha}}(0) \log\left[\frac{P_{\boldsymbol{\alpha}}(0)}{P_{\boldsymbol{\alpha}^*}(0)}\right].$$

$P_{\boldsymbol{\alpha}}(1)$ and $P_{\boldsymbol{\alpha}^*}(1)$ are defined in either equation (3) or equation (6) depending on whether the items have been parameterized using the DINA model or RUM, respectively, and $P_{\boldsymbol{\alpha}}(0)$ is equal to $1 - P_{\boldsymbol{\alpha}}(1)$.

It is also possible to compute Kullback-Leibler information at the test level. Kullback-Leibler information for a test compares the probability distribution for a test vector of $I$ item responses, $\mathbf{X}$, given an attribute pattern, $\boldsymbol{\alpha}$, when compared to the probability distribution of $\mathbf{X}$, given an alternative attribute pattern, $\boldsymbol{\alpha}^*$. The Kullback-Leibler information can be written as

$$d_{\bullet}[\boldsymbol{\alpha}, \boldsymbol{\alpha}^*] = E_{\boldsymbol{\alpha}}\left[\log\left[\frac{P_{\boldsymbol{\alpha}}(\mathbf{X})}{P_{\boldsymbol{\alpha}^*}(\mathbf{X})}\right]\right]. \tag{9}$$

Because one assumption of latent cognitive diagnostic models is independence among items conditional on the attribute $\boldsymbol{\alpha}$, equation (9) can be written as

$$d_{\bullet}[\boldsymbol{\alpha}, \boldsymbol{\alpha}^*] = E_{\boldsymbol{\alpha}}\left[\sum_{j=1}^{J} \log\left[\frac{P_{\boldsymbol{\alpha}}(X_j)}{P_{\boldsymbol{\alpha}^*}(X_j)}\right]\right],$$

which simplifies to

$$d_{\bullet}[\boldsymbol{\alpha}, \boldsymbol{\alpha}^*] = \sum_{j=1}^{J} E_{\boldsymbol{\alpha}}\left[\log\left[\frac{P_{\boldsymbol{\alpha}}(X_j)}{P_{\boldsymbol{\alpha}^*}(X_j)}\right]\right]. \tag{10}$$

Equation (10) is the sum of the Kullback-Leibler information for each item in the exam. Thus, the Kullback-Leibler test information is additive over items, an important and useful property. Note that $d_{\bullet}[\boldsymbol{\alpha}, \boldsymbol{\alpha}^*]$ has an interpretation related to the power of the likelihood ratio test for the null hypothesis that the true parameter is $\boldsymbol{\alpha}$ versus the alternative hypothesis that the true parameter is $\boldsymbol{\alpha}^*$, conducted at a fixed significance level (Rao, 1962). To be specific, if $\beta_J(\boldsymbol{\alpha}, \boldsymbol{\alpha}^*)$ denotes the probability of a Type II error for an assessment of length $J$, the following relationship holds:

$$\lim_{J\to\infty} \frac{\log[\beta_J(\boldsymbol{\alpha}, \boldsymbol{\alpha}^*)]}{-d_J[\boldsymbol{\alpha}, \boldsymbol{\alpha}^*]} = 1.$$

Thus, the Kullback-Leibler information for discriminating between $\boldsymbol{\alpha}$ and $\boldsymbol{\alpha}^*$ is nearly monotonically related to the power of the most powerful test of $\boldsymbol{\alpha}$ versus $\boldsymbol{\alpha}^*$.

One complication of Kullback-Leibler information is that it only compares two attribute patterns when there are $2^K$ possible attribute mastery patterns. Because Kullback-Leibler information is not symmetric, there are a total of $2^K(2^K - 1)$ possible comparisons. To organize the $2^K(2^K - 1)$ comparisons of all attribute pairs for the $j$th item, it is natural to define a $(2^K \times 2^K)$ matrix, $\mathbf{D}_j$, such that each $u, v$ element equals

$$D_{juv} = E_{\boldsymbol{\alpha}_u}\left[\log\left[\frac{P_{\boldsymbol{\alpha}_u}(x_j)}{P_{\boldsymbol{\alpha}_v}(x_j)}\right]\right]. \tag{11}$$

For example, if the DINA model is used, then

$$\begin{aligned} D_{juv} = {} & (1 - s_j)^{\xi_{ju}} g_j^{(1-\xi_{ju})} \log\left[\frac{(1 - s_j)^{\xi_{ju}} g_j^{(1-\xi_{ju})}}{(1 - s_j)^{\xi_{jv}} g_j^{(1-\xi_{jv})}}\right] \\ & + (s_j)^{\xi_{ju}} (1 - g_j)^{(1-\xi_{ju})} \log\left[\frac{(s_j)^{\xi_{ju}} (1 - g_j)^{(1-\xi_{ju})}}{(s_j)^{\xi_{jv}} (1 - g_j)^{(1-\xi_{jv})}}\right], \end{aligned}$$

and if RUM is used,

$$\begin{aligned} D_{juv} = {} & \pi_j^* \prod_{k=1}^{K} r_{jk}^{*(1-\alpha_{uk})q_{jk}} \log\left[\frac{\pi_j^* \prod_{k=1}^{K} r_{jk}^{*(1-\alpha_{uk})q_{jk}}}{\pi_j^* \prod_{k=1}^{K} r_{jk}^{*(1-\alpha_{vk})q_{jk}}}\right] \\ & + \left(1 - \pi_j^* \prod_{k=1}^{K} r_{jk}^{*(1-\alpha_{uk})q_{jk}}\right) \log\left[\frac{1 - \pi_j^* \prod_{k=1}^{K} r_{jk}^{*(1-\alpha_{uk})q_{jk}}}{1 - \pi_j^* \prod_{k=1}^{K} r_{jk}^{*(1-\alpha_{vk})q_{jk}}}\right], \end{aligned}$$

where $\alpha_{uk}$ represents the $k$th element of the attribute mastery vector indicated by $u$. Notice that if one has an item bank of $N$ items, $\mathbf{D}_j$ can be computed for each item. The elements of each $\mathbf{D}_j$ that are large indicate the attribute patterns that the $j$th item is most useful in discriminating among. In addition, the total Kullback-Leibler information matrix can be defined for any exam of $J$ items, $\mathbf{D}_\bullet$, by simply summing across the $\mathbf{D}_j$s of the items selected. To construct a test, one might choose items such that all of the elements in $\mathbf{D}_\bullet$ are large; therefore, the power to discriminate between any two attribute patterns is high.

One concern is that, as $K$ increases, the number of required computations increases exponentially, and therefore the time required also increases. However, the time required is not unreasonable, and such computations must only be performed once for each item. For example, using a program written in Matlab and run on a Pentium IV computer with a 2.4G processor and 1M of RAM, the calculation of the $\mathbf{D}_j$s for 1000 items, parameterized to measure four attributes using RUM, used 2.5 minutes of computing time.

## A CDM Information Index (CDI)

It may not be desirable to focus simultaneously on all the elements $\mathbf{D}_j$ because the number of elements increases exponentially with $K$. Therefore, this study intends to suggest a possible index, or summary, of the elements of $\mathbf{D}_j$. One natural summary of a set of numbers is its mean. Equation (12) defines the mean of the off-diagonal elements of $\mathbf{D}_j$, giving the average Kullback-Leibler distance between any two attribute patterns for item $j$.

$$\overline{D_j} = \frac{1}{2^K(2^K - 1)} \sum_{u \neq v} D_{juv}. \tag{12}$$

Although such an index would be possible, Chang and Ying (1996) suggest in the IRT setting that a general index be computed over a specific interval instead of computed across all possible comparisons. Although a specific interval does not directly apply to the latent classes identified in CDMs, it is important to notice that some comparisons are more important than others. For example, an examinee that has not mastered any of the attributes measured by a test is easily discriminated from a examinee who has mastered all attribute patterns. On the other hand, attribute patterns that differ by only one component are usually the most difficult to discriminate; therefore, $D_{juv}$s for those comparisons require more attention. If a test discriminates well between attribute patterns that are similar, it will discriminate well between those attribute patterns that are not similar. If $\overline{D_j}$ is used as a summary of the elements in $\mathbf{D}_j$, it is possible that high Kullback-Leibler information between highly dissimilar attribute mastery patterns inflates the index when many attribute patterns of mastery cannot be at all well discriminated (i.e., Kullback-Leibler information approximately equals zero). Therefore, a weighted average should be used such that each element is first weighted by the similarity, or inverse "distance" between the attribute patterns. Thus, a larger emphasis is placed on those comparisons of attribute patterns that are more similar.

One common measure that can be used to determine the similarity of any two attribute patterns, $\boldsymbol{\alpha}$ and $\boldsymbol{\alpha}^*$, is the squared Euclidean distance,

$$h(\boldsymbol{\alpha}, \boldsymbol{\alpha}^*) = \sum_{k=1}^{K} (\alpha_k - \alpha_k')^2.$$

Because any attribute pattern is a vector of 1s (masters) and 0s (nonmasters), the squared Euclidean distance is equivalent to the Hamming distance, which is a count of the nonidentical components of $\boldsymbol{\alpha}$. Therefore, when $h(\boldsymbol{\alpha}, \boldsymbol{\alpha}^*)$ is small, the two attribute patterns are similar and should be given a greater emphasis than those attribute pairs with high $h(\boldsymbol{\alpha}, \boldsymbol{\alpha}^*)$.

Using the inverse of $h(\boldsymbol{\alpha}, \boldsymbol{\alpha}^*)$, a weighted mean can be computed as a cognitive diagnostic index ($CDI_j$) of the discriminating power among attribute patterns for the $j$th item. As shown in equation (13), attribute patterns that are more similar, and therefore more difficult to discriminate, are weighted higher than those attribute patterns that are easily distinguished.

$$CDI_j = \frac{1}{\sum_{u \neq v} h(\boldsymbol{\alpha}_u, \boldsymbol{\alpha}_v)^{-1}} \sum_{u \neq v} h(\boldsymbol{\alpha}_u, \boldsymbol{\alpha}_v)^{-1} D_{juv}. \tag{13}$$

Lastly, $CDI_j$ is an indication of the discriminating power of a single item. However, the purpose is to construct a test with high discriminating power, which is identical to making the cognitive diagnostic index of the test, $CDI_\bullet$, as large as possible. The computation of $CDI_\bullet$ is identical to the computation of $CDI_j$ given in equation (13), only now the Kullback-Leibler matrix for the test, $\mathbf{D}_\bullet$, is used, and therefore

$$CDI_\bullet = \frac{1}{\sum_{u \neq v} h(\boldsymbol{\alpha}_u, \boldsymbol{\alpha}_v)^{-1}} \sum_{u \neq v} h(\boldsymbol{\alpha}_u, \boldsymbol{\alpha}_v)^{-1} D_{\bullet uv}.$$

In addition, because all $D_{\bullet uv}$s are linear functions of the $D_{juv}$s for each item, and $h(\boldsymbol{\alpha}_u, \boldsymbol{\alpha}_v)$ is constant between any two specific attribute patterns $\boldsymbol{\alpha}_u$ and $\boldsymbol{\alpha}_v$, it is seen that

$$CDI_\bullet = \sum_{j=1}^{J} CDI_j. \tag{14}$$

It is now possible to summarize the discriminating power of each item into a single index. In addition, to construct a test with a large $CDI_\bullet$ and hence a test that provides good discrimination between masters and nonmasters, items with large $CDI_j$ should be selected, as implied by equation (14). In this way, the criterion for test construction is analogous to the common procedures for test construction used in IRT. In the next section, a heuristic to construct exams from a given item bank based on the index $CDI_j$ is discussed.

## Method

Matlab (MathWorks, 2002) is used to conduct a simulation study that allows the application of a heuristic for test construction on a given item bank. The heuristic discussed is simple in design, yet allows for possible modifications to incorporate design constraints. First, the construction of the item bank and the general design of the heuristic are discussed. Then, the methods used to determine the effectiveness of the heuristic are given.

### Test Construction

#### *Item Bank Construction*

First, an item bank containing $N$ items that measure $K$ attributes is generated. Both the DINA model and RUM require an $(N \times K)$ $Q$-matrix, indicating which attributes are measured by each item, so a simulated $Q$-matrix is generated by randomly selecting $N$ patterns (each a list of the required attributes for its item), with replacement from all possible attribute patterns measuring at least one attribute. Second, the parameters required for each item are randomly generated depending on the model used. The required item parameters for the DINA model are $s_j$ and $g_j$. Both $s_j$ and $g_j$ are randomly generated from a $U(.05, .40)$ distribution. The required parameters for the RUM are $\pi_j^*$, $r_{jk}^*$, and $c_j$. The $\pi_j^*$s are generated from a $U(.75, .95)$ distribution, and each $r_{jk}^*$ is

generated from a $U(.2, .95)$ distribution. For simplicity, the $c_j$s are dropped from the model, setting $P_i(\theta_c) = 1$, and therefore the RUM is mathematically equivalent to Maris's (1999) NIDA model, as defined by equations (4) and (5).

### *A Heuristic for Test Construction*

The goal is to obtain a test of $I$ items from the generated item bank of $N$ items while considering possible constraints. As was previously described, the objective function used in this article is the $CDI_\bullet$. The heuristic uses the following steps:

Step 0 Select the first item with the largest $CDI_j$ that satisfies all constraints.

Step 1 The remaining items in the item bank are evaluated to determine whether, if chosen, they would conform to the specified constraints.

Step 2 Select the next item such that $CDI_j$ is the maximum of all items satisfying the specified constraints.

Step 3 Steps 1 and 2 are repeated until the desired test length is achieved.

Although a number of possible constraints can be implemented in this way, two examples of possible constraints are to control for the number of times each attribute is measured and the number of items allowed to measure a specific number of attributes. For example, to construct a 10-item exam such that no more than 4 items can measure only a single attribute, the algorithm would first select the item with the largest $CDI_j$. Then, for each additional item, the algorithm would first count the number of items already selected that measure only one attribute. If 4 items have already been selected that measure only one attribute, all items measuring only one attribute will no longer be considered as possible items to be selected.

Notice that the procedure will select items such that the final test information ($CDI_\bullet$) is large. If no constraints are supplied, this value is also guaranteed to be optimal. In addition, the algorithm only requires $CDI_j$ and any variables necessary for the constraints. Therefore, it is not necessary to change the heuristic for the DINA model or RUM. It should be noted that the suggested heuristic is one of many possible indices, and each possible index could be implemented by many possible heuristics. However, the Kullback-Leibler matrix (equation (11)) can serve as an excellent basis for CDM test construction.

## Effectiveness of the Proposed Heuristic

Once a method is applied to construct a test, it is important to determine how well the test performs. One natural way is to compare its performance to many randomly constructed tests from the same item bank. The primary goal of CDMs is to classify mastery versus nonmastery. The correct classification rate is used as a measure of performance. The proposed heuristic's correct classification rate is compared to that of 2,000 randomly generated tests.

### *Examinee Generation*

Given a set of item parameters comprising a test, 10,000 examinees are simulated. When generating the examinees, it is important to consider two criteria. These are the proportion of examinees who have mastered the $k$th attribute, $p_k$, and the relationship between the attributes. In the simulation, all $p_k = .5$—namely, each attribute is moderately difficult to master. In addition, it is reasonable to believe that the attributes are correlated in the population. If an individual has mastered one attribute, he or she is more likely to have mastered a second attribute measured by the

exam. Therefore, an appropriate simulation of examinees would incorporate both a specific $p_k$ and a relationship between the attributes.

A total of $n$ multivariate normal $k$-dimensional vectors ($\tilde{\alpha} \sim MVN(\mathbf{0}, \boldsymbol{\rho})$) are generated to simulate attributes that have a positive relationship, where $\rho$ represents a correlation matrix with equal off-diagonal elements. In the discussed experiments, the off-diagonal elements will be either all 0.0 or all 0.5. Using the $p_k$ s, a cutoff $z_c$ is computed for each attribute such that $P(\tilde{\alpha} \leq z_c) = p_k$. The $i$th individual's mastery for attribute $k$ is such that

$$\alpha_{ik} = \begin{cases} 1 & \text{if } \tilde{\alpha} \leq z_c \\ 0 & \text{otherwise} \end{cases}.$$

Notice that if the $\boldsymbol{\rho}$s are positive, then an individual with a specific attribute is more likely to have mastered a second attribute.

For reasons that will become clear in the following subsections, a population, composed of 10,000 examinees, is generated separate from the simulated sample of 10,000 examinees. In further sections, the population will be used to compute posterior distributions. This population will be used for a Monte Carlo approximation of the prior distribution of attribute patterns.

*Generation of Item Scores*

Given each subject's attribute pattern, scores for each item are generated based on the chosen model. Given the probability of a correct response, $P(Y_{ij} = 1|\alpha)$, a random $U(0, 1)$ variable $u$ is generated, and the score $Y_{ij}$ is

$$Y_{ij} = \begin{cases} 1 & \text{if } u \leq P(Y_{ij} = 1|\alpha) \\ 0 & \text{otherwise} \end{cases}.$$

*Estimation of the Attribute Patterns*

Because the item parameters are known, classification is accomplished by computing the likelihood for all possible attribute patterns given the examinees' scores and multiplying by the prior probabilities (estimated from the population of 10,000 subjects). The posterior mode is then used to classify the attribute pattern for that individual.

Given the estimated attribute patterns, the proportion of attributes that were correctly identified is recorded. It should be noted that the correct classification rate of attribute mastery is computed marginally (i.e., for each attribute) in addition to the correct classification rate for the entire attribute pattern.

## Procedure

It is important that a test construction method is effective even when the number of specified attributes is varied, constraints are implemented, and the distribution of the attribute patterns varies. In addition, the intent is to show that the proposed method is a general method for test construction that can be used for both the DINA model and RUM and hence for other closely related models. Therefore, all comparisons are made using both models with $\boldsymbol{\rho} = 0.0$ and $\boldsymbol{\rho} = 0.5$.

To explore the effectiveness of the heuristic when the number of attributes varies, item banks (i.e., one with the DINA model's parameterization and one with the RUM's parameterization) containing 300 items are constructed for both four and eight attributes. From the four-attribute item banks for the DINA model and RUM, 20 item exams are constructed using the discussed heuristic, assuming no constraints, and compared to 2,000 randomly generated exams from the same item bank. The method of comparison uses the simulation discussed previously such that the proportion

of correctly classified attributes and the proportion of correctly individually classified attribute patterns are reported. In addition, an empirical $p$ value is computed using the proportion of random exams that performed better than the exam constructed by the proposed heuristic.

Next, to implement possible constraints, two examples are supplied. The first set of constraints controls the number of items that measure a specific number of attributes. Therefore, of the 20-item exams measuring four attributes, 9 items must measure three attributes, 7 items must measure two attributes, and the remaining 4 items must measure only one attribute. The constraints for the 20-item exams measuring eight attributes are such that 9 items must measure four attributes, 7 items must measure three attributes, and the remaining 4 items must measure only two attributes. Notice that it is not necessary for the constraint to focus only on the number of attributes measured by an item. It is also possible to constrain items based on content or other aspects of the test setting. For example, in a 20-item math test, it may be necessary to require 10 items to measure algebra skills and 10 items to measure geometry skills.

As an alternative, it may be necessary to control the number of items that measure each attribute. Because $CDI_{\bullet}$ summarizes the elements of a matrix, it is possible that, under certain circumstances, $CDI_{\bullet}$ may be large without effectively measuring all attributes. By implementing a constraint controlling for the minimum number of times each attribute must be required by an item, such situations are eliminated. Using the item banks of 300 items measuring four attributes, exams for the DINA model and RUM were constructed requiring each attribute to be measured a minimum of seven times. For the eight-attribute test, it is not reasonable to believe that a 20-item exam will measure each attribute seven times. Therefore, the constraint was lowered such that each attribute must be measured at least three times. Again, 2,000 random tests, which also satisfy the respective constraints, were created for comparison. The correct classification rates for attributes and attribute patterns are reported.

## Results

### Test Construction for the DINA Model

The heuristic is first applied to the item banks with DINA model parameters with $\rho$ equal to 0, and all results are summarized in Tables 1 and 2. Table 1 addresses attribute patterns, and Table 2 addresses attributes marginally. In general, the tests constructed using the heuristic significantly outperformed randomly constructed tests both from the joint attribute pattern and the marginal attribute perspective. When constructing a test measuring four attributes, on average, the heuristic exam's proportion of correct classification of attribute patterns is 32.5% higher when compared to the average proportion of correct classification for exams randomly constructed from the same item bank. In the marginal case, the heuristic is, on average, 13.5% higher that the randomly constructed tests when measuring four attributes.

In addition, the heuristic is used to construct a 20-item exam to measure eight attributes. Although the percentages of correct classification of attribute mastery patterns are much lower, due to the difficult task of measuring eight attributes with only 20 items, the heuristic still shows significant improvement over randomly selected tests. When constructing a test measuring eight attributes, the heuristic's correct classification rate was, on average, 18.3% higher when compared to exams randomly constructed from the same item bank. Similarly, in the marginal case, the heuristic is, on average, 16.0% higher than the randomly constructed tests when measuring four attributes.

Similar results are obtained for the DINA model when $\rho$ equals 0.5. The results are similar to those with $\rho$ equal to 0.0. Specifically, the average difference of the correct classification rate between the heuristically and randomly constructed four-attribute tests is 33.7% for attribute patterns and 14.5% marginally. The average difference of the correct classification rate between

**Table 1**

Correct Classification Rates of Attribute Patterns Using the Deterministic Input Noisy "and" Gate Model (DINA) When $\rho = 0.0$

| Number of Attributes | Condition | Heuristically Constructed | Randomly Constructed | Random Outperforms |
|---|---|---|---|---|
| 4 | No constraints | .923 | .530 | 0 |
| 4 | Item constraints | .695 | .519 | .069 |
| 4 | Attribute constraints | .900 | .531 | 0 |
| 8 | No constraints | .232 | .080 | 0 |
| 8 | Item constraints | .247 | .111 | .002 |
| 8 | Attribute constraints | .370 | .085 | 0 |

**Table 2**

Marginal Correct Classification Rates of Attributes Using the Deterministic Input Noisy "and" Gate Model (DINA) When $\rho = 0.0$

| Number of Attributes | Condition | Heuristically Constructed | Randomly Constructed | Random Outperforms |
|---|---|---|---|---|
| 4 | No constraints | .980 | .816 | 0 |
| 4 | Item constraints | .891 | .806 | .069 |
| 4 | Attribute constraints | .972 | .815 | 0 |
| 8 | No constraints | .772 | .607 | 0 |
| 8 | Item constraints | .715 | .628 | .002 |
| 8 | Attribute constraints | .832 | .603 | 0 |

the heuristically and randomly constructed eight-attribute tests is 13.6% for attribute patterns and 13.3% marginally. The results are provided in Tables 3 and 4.

Last, it should be noted that there is only one condition for which a positive percentage of the 2,000 randomly selected exams performed better than the heuristic. The result suggests that only a small proportion of all possible combinations is better than those selected by the heuristic. One possible reason that the heuristic is occasionally outperformed by randomly constructed exams

**Table 3**

Correct Classification Rates of Attribute Patterns Using the Deterministic Input Noisy "and" Gate Model (DINA) When $\rho = 0.5$

| Number of Attributes | Condition | Heuristically Constructed | Randomly Constructed | Random Outperforms |
|---|---|---|---|---|
| 4 | No constraints | .913 | .551 | 0 |
| 4 | Item constraints | .790 | .535 | .001 |
| 4 | Attribute constraints | .947 | .552 | 0 |
| 8 | No constraints | .368 | .210 | 0 |
| 8 | Item constraints | .309 | .221 | .082 |
| 8 | Attribute constraints | .374 | .213 | 0 |

**Table 4**

Marginal Correct Classification Rates of Attributes Using the Deterministic Input Noisy "and" Gate Model (DINA) When $\rho = 0.5$

| Number of Attributes | Condition | Heuristically Constructed | Randomly Constructed | Random Outperforms |
|---|---|---|---|---|
| 4 | No constraints | .977 | .825 | 0 |
| 4 | Item constraints | .938 | .815 | .001 |
| 4 | Attribute constraints | .986 | .826 | 0 |
| 8 | No constraints | .828 | .643 | 0 |
| 8 | Item constraints | .727 | .664 | .082 |
| 8 | Attribute constraints | .816 | .655 | 0 |

**Table 5**

Correct Classification Rates of Attribute Patterns Using the Reparameterized Unified Model (RUM) When $\rho = 0.0$

| Number of Attributes | Condition | Heuristically Constructed | Randomly Constructed | Random Outperforms |
|---|---|---|---|---|
| 4 | No constraints | .832 | .550 | 0 |
| 4 | Item constraints | .806 | .555 | 0 |
| 4 | Attribute constraints | .831 | .550 | 0 |
| 8 | No constraints | .256 | .127 | 0 |
| 8 | Item constraints | .249 | .140 | 0 |
| 8 | Attribute constraints | .263 | .128 | 0 |

when implementing Constraint 1 is due to the specific attributes measured by the items selected. Notice that the $CDI_j$ does not incorporate the $Q$-matrix and therefore ignores possible situations for which one or more of the attributes are not effectively measured in a constructed exam. Therefore, under certain situations, it may be more appropriate to select items with lower discriminating power if they measure attributes not otherwise measured in the items already selected. Such a situation is likely to occur when constructing a 20-item exam to measure eight attributes, which can be seen by the heuristic performing most effectively when controlling for the minimum number of times each attribute is measured.

## Test Construction for the RUM

The identical simulations are run using items that were parameterized for the RUM. All results are provided in Tables 5 and 6. Again, as in the results using the DINA model for $\rho$ equal to 0.0, there are marked improvements in the correct classification of attribute patterns using the RUM when the heuristic is used to construct a test when compared to randomly constructed exams from the same item bank. When constructing a 20-item exam measuring four attributes, the correct classification based on the heuristic is 26% higher than the randomly constructed exams for the entire attribute patterns. In addition, for the 20-item exams measuring the eight-attribute $\rho = 0$ case, the correct classification rate obtained by the heuristic for attribute mastery patterns is 11.6% higher. In the simulation, there were no cases in which a randomly constructed exam performed better than the

**Table 6**

Marginal Correct Classification Rates of Attributes Using the Reparameterized Unified Model (RUM) When $\rho = 0.0$

| Number of Attributes | Condition | Heuristically Constructed | Randomly Constructed | Random Outperforms |
|---|---|---|---|---|
| 4 | No constraints | .951 | .850 | 0 |
| 4 | Item constraints | .941 | .850 | 0 |
| 4 | Attribute constraints | .951 | .849 | 0 |
| 8 | No constraints | .833 | .745 | 0 |
| 8 | Item constraints | .820 | .760 | 0 |
| 8 | Attribute constraints | .838 | .745 | 0 |

heuristic. In the marginal case, the average difference of classification rate between the heuristically and randomly constructed tests is 9.8% for the four-attribute case and 7.7% for the eight-attribute case.

Again, as in the simulations using the DINA model parameterization, similar results are obtained for the RUM when $\rho$ equals 0.5. For example, on average, the attribute pattern classification rate is 21.1% higher than randomly constructed tests, and the marginal classification rate for a four-attribute test is 7.2% higher. All results are given in Tables 7 and 8.

**Table 7**

Correct Classification Rates of Attribute Patterns Using the Reparameterized Unified Model (RUM) When $\rho = 0.5$

| Number of Attributes | Condition | Heuristically Constructed | Randomly Constructed | Random Outperforms |
|---|---|---|---|---|
| 4 | No constraints | .855 | .631 | 0 |
| 4 | Item constraints | .826 | .638 | 0 |
| 4 | Attribute constraints | .852 | .633 | 0 |
| 8 | No constraints | .422 | .330 | 0 |
| 8 | Item constraints | .435 | .321 | 0 |
| 8 | Attribute constraints | .438 | .325 | 0 |

**Table 8**

Marginal Correct Classification Rates of Attributes Using the Reparameterized Unified Model (RUM) When $\rho = 0.5$

| Number of Attributes | Condition | Heuristically Constructed | Randomly Constructed | Random Outperforms |
|---|---|---|---|---|
| 4 | No constraints | .960 | .883 | 0 |
| 4 | Item constraints | .951 | .886 | 0 |
| 4 | Attribute constraints | .959 | .884 | 0 |
| 8 | No constraints | .882 | .827 | 0 |
| 8 | Item constraints | .876 | .829 | 0 |
| 8 | Attribute constraints | .884 | .825 | 0 |

## Discussion

The results provide solid evidence that a heuristic index ($CDI_\bullet$) based on Kullback-Leibler information is effective in constructing exams that yield considerably higher attribute classification rates than randomly constructed tests. It should be noted that the $CDI_\bullet$ does not make explicit use of the $Q$-matrix. Under some circumstances, using only the $CDI_\bullet$ may be suboptimal. Because the $CDI_\bullet$ is a summary of many comparisons, it is possible to construct two tests with equal $CDI_\bullet$s, and yet one test classifies better than the other. It may be necessary to define constraints such as Constraint 2, which control for aspects of the $Q$-matrix (e.g., constraints that specify the minimum number of times any attribute must be measured in the constructed test) to ensure that unacceptable situations do not occur. In addition, effects of population characteristics are only briefly explored (e.g., different relationships among attributes). Different population characteristics, such as proportion of mastery for each attribute, $p_k$, or the attribute correlations may influence the effectiveness of the proposed heuristic.

Last, the proposed heuristic is a simple algorithm used to construct a test such that $CDI_\bullet$ is large. The algorithm has been shown to be effective under situations with simple constraints. However, as the number of constraints is increased, it is no longer guaranteed that the heuristic is effective. Due to the linear nature of the $CDI_\bullet$, a more appropriate method of item selection when the number of constraints is large would be to implement integer programming. As was briefly discussed for test construction in IRT, integer programming is used to select items that optimize the $CDI_\bullet$ given a set of constraints. Future research will explore the applications and advantages of using more complicated algorithms such as integer programming for test construction using cognitively diagnostic assessment while examining the effects of population characteristics such as attribute correlations.

## References

Chang, H., & Ying, Z. (1996). A global information approach to computerized adaptive testing. *Applied Psychological Measurement*, *20*, 213-229.

de la Torre, J., & Douglas, J. (in press). Higher-order latent trait models for cognitive diagnosis. *Psychometrika*.

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

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

Hambleton, R., & Swamiinathan, H. (2000). *Item response theory*. Boston: Kluwer Nijhoff.

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

Hartz, S. (2002). *A Bayesian framework for the unified model for assessing cognitive abilities: Blending theory with practicality*. Unpublished doctoral dissertation, University of Illinois, Urbana-Champaign.

Junker, B. W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement*, *12*, 55-73.

Lehmann, E., & Casella, G. (1998). *Theory of point estimation: Second Edition*. New York: Springer-Verlag.

Lord, F. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum.

Lord, F., & Novick, M. (1968). *Statistical theories of mental test scores with contributions from Alan Birnbaum*. Reading, MA: Addison-Wesley.

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

Madigan, D., & Almond, R. (1995). Test selection strategies for belief networks. In D. Fisher & H.-J. Lenz (Eds.), *Learning from data: AI and statistics IV* (pp. 89-98). New York: Springer-Verlag.

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

MathWorks. (2002). *Getting started with Matlab*. Natrick, MA: Author.

McDonald, R. (1999). *Test theory: A unified treatment*. Mahwah, NJ: Lawrence Erlbaum.

Rao, C. R. (1962). Efficient estimates and optimum inference procedures in large samples. *Journal of the Royal Statistics Society Series B*, *24*, 46-72.

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.

Tatsouka, C., & Ferguson, T. (2003). Sequential classification on partially ordered sets. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *65*(1), 143-158.

Tatsuoka, K. (1985). A probabilistic model for diagnosing misconceptions in the pattern classification approach. *Journal of Educational Statistics*, *12*, 55-73.

van der Linden, W. (1998). Optimal assembly of psychological and educational tests. *Applied Psychological Measurement*, *22*, 195-211.

Veldkamp, B. P., & van der Linden, W. J. (2002). Multidimensional adaptive testing with constraints on test content. *Psychometrika*, *67*(4), 575-588.

Yamamoto, K. (1987). *A model that combines IRT and latent class models*. Unpublished doctoral dissertation, University of Illinois, Champaign-Urbana.

## Acknowledgments

*The authors thank Dr. Daniel Eignor and Dr. Russell Almond for their constructive comments about an earlier draft. They would also like to thank ETS for funding this project.*

## Author's Address

Address correspondence to Robert Henson, 604 Haines Blvd., Champaign, IL 61820; e-mail: rahenson@uiuc.edu.