# Cognitive Diagnostic Attribute-Level Discrimination Indices

Robert Henson, Louis Roussos, Jeff Douglas and Xuming He

*Applied Psychological Measurement* 2008 32: 275
DOI: 10.1177/0146621607302478

# Cognitive Diagnostic Attribute-Level Discrimination Indices

**Robert Henson, The University of North Carolina at Greensboro**
**Louis Roussos, Measured Progress**
**Jeff Douglas and Xuming He, University of Illinois at Urbana-Champaign**

Cognitive diagnostic models (CDMs) model the probability of correctly answering an item as a function of an examinee's attribute mastery pattern. Because estimation of the mastery pattern involves more than a continuous measure of ability, reliability concepts introduced by classical test theory and item response theory do not apply. The cognitive diagnostic index (CDI) measures an item's overall discrimination power, which indicates an item's usefulness in examinee attribute pattern estimation. Because of its relationship with correct classification rates, the CDI was shown to be instrumental in cognitively diagnostic test assembly. This article generalizes the CDI to attribute-level discrimination indices for an item. Two different attribute-level discrimination indices are defined; their relationship with correct classification rates is explored using Monte Carlo simulations. There are strong relationships between the defined attribute indices and correct classification rates. Thus, one important potential application of these indices is test assembly from a CDM-calibrated item bank. *Index terms: cognitive diagnosis, cognitive diagnostic index, item discrimination index, Kullback-Leibler information*

## Introduction

If the goal of a test is to measure an examinee's general ability accurately, indices exist that can provide an indication of the value of each item. However, modern methods for skills diagnosis are interested in determining mastery on $K$ dichotomous skills rather than assessing general ability; therefore, typical indices of a "good" item cannot directly apply. A new set of indices used to indicate a good item must be developed. First, a brief description of cognitive diagnosis models (CDMs) is provided, followed by the problems of defining a good item when using skills diagnosis models. Then, two indices are defined that can be used to indicate the value of an item and show its relationship to estimation of the examinees' mastery on a set of skills using a Monte Carlo simulation study.

As opposed to estimating an examinee's ability along a continuum, skills diagnostic models (CDMs) typically estimate a student's mastery profile and therefore provide those skills that an examinee has mastered or has not mastered. In doing this, the probability of a correct response is modeled as a function of mastery of the set of $K$ skills (the mastery profile, $\alpha$). The general basis underlying such models is that in order to correctly respond to an item, one must have mastered a basic set of skills; if any of these skills have not been mastered, the chance of a correct response is lowered.

*Applied Psychological Measurement,* Vol. 32 No. 4, June 2008, 275–288
DOI: 10.1177/0146621607302478

Because examinees are characterized by a set of dichotomous skills, most CDMs can be directly compared to latent class models. Specifically, most CDMs are constrained latent class models where each class is defined by the mastery profile (these have also been called multiple latent class models). All examinees with the same mastery profile (i.e., examinees in the same latent class) are assumed to have the same expected response pattern. Macready and Dayton (1977) were among the first to discuss such mastery latent class models using only one dichotomous trait to measure mastery of a test domain. In addition, Rindskopf (1983) suggests using latent class analysis with particular constraints placed on the item probabilities. Later, Haertel (1989) parameterized a simple model called the binary skills model, which was later called the DINA model (Deterministic Input, Noisy "And" gate model) by Junker and Sijtsma (2001). Other CDMs include the NIDA (Noisy Input, Deterministic "And" gate; Maris, 1999) and the reparameterized unified model (RUM; DiBello, Stout, & Roussos, 1995; Hartz, 2002; Stout, 2002). The remainder of this article will use a reduced version of the RUM as an example; however, the concepts and computations apply to any cognitive diagnostic model with a discrete latent examinee space.

The RUM includes three different item parameters, $\pi_j^*$, $r_{jk}^*$, and $P_{c_j}$, for $j = 1, \ldots, J$ (number of items) and $k = 1, \ldots, K$ (number of attributes). The probability of a correct response for the $i$th examinee given $\alpha$ and $\eta_i$ is

$$P(\mathbf{X}_{ij} = 1 | \alpha_i, \eta_i) = \pi_j^* \prod_{k=1}^{K} r_{jk}^{*(1-\alpha_{ik})q_{jk}} P_{c_j}(\eta_i). \tag{1}$$

Here, $\pi_j^*$ is the probability of correctly applying all required attributes for the $j$th item given that the examinee has mastered all required attributes for that item, $r_{jk}^*$ represents the discrimination of the $j$th item for the $k$th attribute (notice $r^*$ is an $I \times K$ matrix), $q_{jk}$ is an indicator for whether the $j$th item requires mastery of the $k$th attribute (the matrix, $\mathbf{Q}$, is an $I \times K$ matrix describing which attributes are required for each item), $P_{c_j}$ is a Rasch model with difficulty parameter $-c_j$, and $\eta_i$ is a general measure of the $i$th examinee's knowledge not otherwise specified by the Q-matrix. In the current research, $c_j$ is assumed to equal $\infty$; therefore, $P_{c_j} = 1$ and is excluded from the model.[1] Finally, the latent variables in the RUM indicating an examinee's ability include the mastery profile, $\alpha$, which is a $(1 \times K)$ vector indicating each examinee's pattern of mastery for a set of $K$ attributes, and $\eta$, which is a continuous variable containing all other abilities required to answer the items that have not been defined in the Q-matrix.

As was mentioned, CDMs model the probability of correctly answering an item as a function of an examinee's attribute mastery pattern. Henson and Douglas (2005) mention that "because estimation of the mastery pattern no longer involves a continuous measure of ability, concepts initially introduced by CTT [classical test theory] and IRT [item response theory], such as reliability and information, do not apply" (Henson & Douglas, 2005, p. 263). For example, reliability in CTT can be defined as the proportion of the variance of the observed score that can be accounted for by the variance of the continuous latent true score (Lord & Novick, 1968). Although the concept of reliability may be applicable because an individual is or is not correctly classified, the interpretation of a reliability coefficient such as Cronbach's $\alpha$ should not be the same as the interpretation when using CTT. Also, the concept of Fisher information is no longer applicable. Mathematically, Fisher information is defined as the negative expectation of the second derivative of the log likelihood for a specified ability (Lord, 1980). Because attribute patterns are in a discrete space, it is not possible to compute the Fisher information at a specific attribute pattern by computing the second derivative with respect to the latent examinee variable. In summary, there is not

a clear choice of index that measures the effectiveness of a skills diagnostic item, or test, such as CTT's reliability or IRT's Fisher information.

Instead of the indices, or measures, that are traditionally used in CTT or IRT as indicators of good items, Henson and Douglas (2005) suggest using the cognitive diagnostic index (CDI) as a measure of an item's (or test's) discrimination power. The CDI is a Kullback-Leibler–based index that is related to the distances between the item response probability distributions for each attribute pattern. They show that the CDI strongly relates to the average correct classification rates across attributes and examinees for a test. Because of this, Henson and Douglas show that the CDI can be a useful index for item selection in test assembly. Specifically, to assemble a test from an item bank, those items with the largest CDIs should be selected first. Given the relationship between the CDI and average correct classification rates across attributes, this test will have a high correct classification rate when compared to all other tests that could be constructed from the same item bank.

Because the CDI can be computed for any CDM as a summary of the item's overall discriminating power, it does not indicate an item's discrimination power for a specific attribute. In addition, by its definition, the CDI ignores which attributes are required by which items (that is, $\mathbf{Q}$). Therefore, it is necessary to expand the CDI to a set of indices that measure the discrimination power of an item for each attribute, which incorporates $\mathbf{Q}$. If the attribute-level indices are constructed similarly to the CDI, one would expect an attribute discrimination index to have a strong association with correct classification rates for that attribute. Those tests with only items where attribute indices are large will also have high correct classification rates. In addition, item selection for test assembly based on attribute discrimination indices will not suffer from the same limitations as the CDI. It should be noted that by focusing on a Kullback-Leibler index, as opposed to specific indices based on model-specific item parameters, a general method is provided that will apply to all CDMs. For any given model, one could then define an item parameter index that would relate to these indices.

This report proposes two attribute discrimination indices. Because the discrimination indices are based on the Kullback-Leibler information, as is the CDI, a brief description of the Kullback-Leibler information is provided. Then, the two attribute discrimination indices are discussed, and a Monte Carlo simulation study is used to demonstrate the strong relationship between each index and correct classification rates.

## Kullback-Leibler Information

The Kullback-Leibler information is a measure of the difference between two probability distributions $f(\mathbf{X})$ and $g(\mathbf{X})$. Formally, the Kullback-Leibler information is

$$K[f, g] = E_f \left[ \log \left[ \frac{f(\mathbf{X})}{g(\mathbf{X})} \right] \right], \tag{2}$$

where the measure $K[f, g]$ is equal to the expectation, assuming $f(\mathbf{X})$ is the true distribution, of the log-likelihood ratio of any two probability density functions $f(\mathbf{X})$ and $g(\mathbf{X})$. $\mathbf{X}$ denotes the random data and can be a scalar or vector (Henson & Douglas, 2005). Although this value is typically thought of as a measure of distance because it ranges from 0 to $\infty$, it is not symmetric and it does not satisfy the triangle inequality; however, as the value increases, the distributions are easier to discriminate.

Henson and Douglas (2005) note that because the Kullback-Leibler does not require a continuous space, it can easily be applied to CDMs by setting $f(\mathbf{X}) = P_{\alpha_u}(\mathbf{X}_j)$ and $g(\mathbf{X}) = P_{\alpha_v}(\mathbf{X}_j)$,

where, generally speaking, $P_{\alpha}(\mathbf{X}_j)$ is the probability of the response $\mathbf{X}_j$ given $\alpha$. Therefore, the Kullback-Leibler information between these two distributions for item $j$ is

$$K_j[\alpha_u, \alpha_u] = \sum_{\mathbf{X}_j=0}^{1} P_{\alpha_u}(\mathbf{X}_j)\log\left[\frac{P_{\alpha_u}(\mathbf{X}_j)}{P_{\alpha_v}(\mathbf{X}_j)}\right], \tag{3}$$

namely,

$$P_{\alpha_u}(1)\log\left[\frac{P_{\alpha_u}(1)}{P_{\alpha_v}(1)}\right] + P_{\alpha}(0)\log\left[\frac{P_{\alpha}(0)}{P_{\alpha'}(0)}\right]. \tag{4}$$

$P_{\alpha_u}(1)$ and $P_{\alpha_v}(1)$ are defined as the probability of a correct response, and $P_{\alpha_u}(0)$ and $P_{\alpha_v}(0)$ are defined as the probability of an incorrect response given $\alpha_u$ and $\alpha_v$, respectively.

The Kullback-Leibler, as defined above, describes only the ability to discriminate between two attribute patterns, but there are $2^K(2^K - 1)$ possible comparisons. For simplification, all values are contained in a $(2^K \times 2^K)$ matrix, $\mathbf{D}_j$ such that the $u, v$ element for the $j$th item is

$$D_{juv} = E_{\alpha_u}\left[\log\left[\frac{P_{\alpha_u}(x_j)}{P_{\alpha_v}(x_j)}\right]\right]. \tag{5}$$

Last, it should be noted that this matrix is additive across items. Specifically, if a test were to contain $J$ items, the matrix pertaining to the test, $\mathbf{D}_\bullet$, is simply the sum of the item matrices, $\mathbf{D}_j$. The CDI is a weighted average of the elements in this matrix. For a more thorough description of Kullback-Leibler information in test construction see Chang and Ying (1996), and for its use in cognitive diagnosis see Henson and Douglas (2005).

## Discrimination

Theoretically, each element of $\mathbf{D}_\bullet$ could function as an indicator of how well $\alpha_u$ is measured when compared to $\alpha_v$. However, in applications, it is not reasonable to consider the $2^K(2^K - 1)$ discrimination indices of an exam simultaneously. Therefore, to produce a useful indication of test discrimination, it is imperative that a single index, such as the CDI, or $2K$ attribute-level indices of discrimination, based on the entries of $\mathbf{D}_\bullet$, be defined. Specifically, the $k$th discrimination index for a test should be associated with the correct classification rate for the $k$th attribute for a given test.

It should be noted that although correct classification of an attribute has been discussed generically, there are two important components to correct classification: the correct classification rate of the masters, $p(\hat{\alpha}_k = 1|\alpha_k = 1)$, and the correct classification rate of the nonmasters, $p(\hat{\alpha}_k = 0|\alpha_k = 0)$. The purpose of the test (e.g., including differing costs and differing benefits of correct classification) can often determine whether the correct classification of masters or the correct classification of nonmasters is more important. Therefore, instead of defining only a single discrimination index for the $k$th attribute, a discrimination index will be defined that is related to the correct classification rate of the masters for the $k$th attribute and a discrimination index that is related to the correct classification rate of nonmasters.

The intent is to show that effective indices at the attribute level can be computed from the Kullback-Leibler matrix, $\mathbf{D}_\bullet$. In addition, by using only linear combinations of the elements of $\mathbf{D}_\bullet$, the defined attribute discrimination index for a test is simply the sum of each corresponding item attribute discrimination index (that is, additivity holds across items), which will allow attribute-specific test construction in a similar manner as described by Henson and Douglas (2005). The

following subsections provide definitions of two promising indices of attribute discrimination, each with their benefits and limitations.

### Attribute Discrimination Index A ($\mathbf{d}_{(A)j}$)

Some elements within $\mathbf{D}_j$ represent attribute pattern comparisons that are more difficult to discriminate than others. For example, an examinee who has not mastered any attributes is easily discriminated from an examinee who has mastered all attribute patterns. On the other hand, attribute patterns that differ by only one component (i.e., the $k$th component) are the most difficult to discriminate. Therefore, only the $D_{juv}$s for those comparisons such that the attribute patterns of mastery differ by only one component will be used to compute the discrimination. Notice that by using the attribute patterns that differ only on the $k$th attribute, the corresponding $D_{juv}$s describe the extent to which a master can be discriminated from a nonmaster, or a nonmaster from a master, on the $k$th attribute while holding attribute mastery constant on the remaining $(K-1)$ attributes.

Of the attribute comparisons that differ only by the $k$th attribute, there are $2^{(K-1)}$ comparisons describing the discrimination power of masters from nonmasters on the $k$th attribute (i.e., comparing attribute patterns such that $\alpha_{uk}=1$ and $\alpha_{vk}=0$), and there are $2^{(K-1)}$ comparisons describing the discrimination power of nonmasters from masters on the $k$th attribute (i.e., attribute patterns such that $\alpha_{uk}=0$ and $\alpha_{vk}=1$). The first index will compute the mean of the elements in $\mathbf{D}_j$ that satisfy these constraints. Specifically, equations (6) and (7) provide formal definitions of $d_{(A)jk1}$ and $d_{(A)jk0}$, for $j=1,\ldots,J$ and $k=1,\ldots,K$, in terms of the comparisons made in $\mathbf{D}_j$, for the masters and nonmasters, respectively:

$$d_{(A)jk1}=\frac{1}{2^{(K-1)}}\sum_{\Omega_{k1}}D_{juv}, \tag{6}$$

$$d_{(A)jk0}=\frac{1}{2^{(K-1)}}\sum_{\Omega_{k0}}D_{juv}, \tag{7}$$

where

$$\Omega_{k1}\equiv\{\alpha_{uk}=1 \text{ and } \alpha_{vk}=0 \text{ and } \alpha_{um}=\alpha_{vm}\ \forall\ m\neq k\}$$

and

$$\Omega_{k0}\equiv\{\alpha_{uk}=0 \text{ and } \alpha_{vk}=1 \text{ and } \alpha_{um}=\alpha_{vm}\ \forall\ m\neq k\}.$$

Indices $d_{(A)jk1}$ and $d_{(A)jk0}$ provide a simple measure of the average discrimination that item $j$ contains about attribute $k$ while controlling for the remaining attributes. Such a measure does not incorporate prior knowledge about the testing population and therefore assumes that all attribute patterns are equally likely. If the $j$th item does not measure the $k$th attribute (i.e., the $j,k$ element of the Q-matrix is 0), then that item contains no information about attribute mastery for the $k$th attribute; therefore, $d_{(A)jk1}$ and $d_{(A)jk0}$ are zero. Although the index has been defined at the item level, the matrix of test discrimination indices for an item, $\mathbf{d}_{(A)\bullet}$, is the sum across each item discrimination as given in equation (8).

$$\mathbf{d}_{(A)\bullet}=\sum_{j=1}^{J}\mathbf{d}_{(A)j}. \tag{8}$$

### Attribute Discrimination Index B ($\mathbf{d}_{(B)j}$)

Because it is essential to use prior testing to calibrate the items, there is usually some knowledge of the population characteristics. For example, Hartz (2002) estimates attribute associations

and the population probability of mastery using the fusion model to fit the RUM. If the fusion model is fitted, joint probabilities of attribute patterns are estimated. In addition, it can be argued, in general, that there are not many cases in which all attribute patterns are equally likely. Therefore, a second set of indices, $d_{(B)jk1}$ and $d_{(B)jk0}$, are defined, as in equations (9) and (10), in which the expectation given the distribution of $\alpha$ is used (i.e., the joint probabilities, or estimates of the joint probabilities, of the attribute patterns are used to weight the appropriate elements of $\mathbf{D}_j$):

$$d_{(B)jk1} = E_{\alpha}[D_{juv}|\Omega_{k1}], \tag{9}$$

$$d_{(B)jk0} = E_{\alpha}[D_{juv}|\Omega_{k0}], \tag{10}$$

where $\Omega_{k1}$ and $\Omega_{k0}$ are defined previously for Index A.

Provided that the distribution of $\alpha$ is known, or can be estimated, equation (9) can be rewritten as

$$d_{(B)jk1} = \sum_{\Omega_{k1}} w_{k1} D_{juv}, \tag{11}$$

where

$$w_{k1} = P(\alpha|\alpha_k = 1),$$

and equation (10) can be rewritten as

$$d_{(B)jk0} = \sum_{\Omega_{k0}} w_{k0} D_{juv}, \tag{12}$$

where

$$w_{k0} = P(\alpha|\alpha_k = 0).$$

Like $d_{(A)jk1}$ and $d_{(A)jk0}$, $d_{(B)jk1}$ and $d_{(B)jk0}$ provide a simple measure of discrimination, but population information is used to weight the elements of $\mathbf{D}_j$, giving those values for which $\alpha$ is more likely higher weights than less likely attribute patterns. $d_{(B)jk1}$ and $d_{(B)jk0}$ are interpreted as the amount of information provided by an item about attribute $k$. It should be noticed that if all $P(\alpha|\alpha_k = 1)$ are equal, then $d_{(B)jk1} = d_{(A)jk1}$, and if all $P(\alpha|\alpha_k = 0)$ are equal, then $d_{(B)jk0} = d_{(A)jk0}$. Therefore, $d_{(A)jk1}$ and $d_{(A)jk0}$ are special cases of $d_{(B)jk1}$ and $d_{(B)jk0}$, respectively. Again, as in the discrimination index, $\mathbf{d}_{Aj}$,

$$\mathbf{d}_{(B)\bullet} = \sum_{j=1}^{J} \mathbf{d}_{(B)j}. \tag{13}$$

## Examples

A simple example using an item with RUM model parameters will illustrate the calculations of the two indices (i.e., $\mathbf{d}_{(A)jk1}$ and $\mathbf{d}_{(B)jk1}$). The item has an $r^*_{j1}$ equal to .125, a $\pi^*_j = .8$, and a Q-matrix entry equal to (1 0). Notice that the Q-matrix entry indicates that the first of only two attributes is required to answer the item correctly.

To compute $\mathbf{d}_{(A)j}$ and $\mathbf{d}_{(B)j}$, the matrix $\mathbf{D}_j$ must be calculated using the probability of a correct response as defined by the RUM. Recall that the matrix $\mathbf{D}_j$ is a $(2^K \times 2^K)$ matrix containing the Kullback-Leibler information for all pairs of attributes as defined in equation (3). In addition, as shown in equation (4), for any given two attribute patterns, this value is only a function of the probability of a correct response and the probability of an incorrect response (i.e., one minus the probability of a correct response). Therefore, to compute $\mathbf{D}_j$ for the example, the probability of a correct response for all possible attribute patterns must first be computed. In this case, there are $2^2$ attribute

patterns. Using the RUM where $P_c(\eta) = 1$, the probability of a correct response for the four attribute patterns are as follows, based on the function given in equation (1):

$$\begin{aligned}
P_{\{0,0\}}(1) &= \pi_j^* r_{j1}^* = (.8)(.125) = .1,\\
P_{\{0,1\}}(1) &= \pi_j^* r_{j1}^* = (.8)(.125) = .1,\\
P_{\{1,0\}}(1) &= \pi_j^* = (.8) = .8,\\
P_{\{1,1\}}(1) &= \pi_j^* = (.8) = .8.
\end{aligned}$$

As can be seen in this example, the probability of a correct response is equal to $\pi_j^*$ in those cases where the required attribute, Attribute 1, has been mastered and equal to $\pi_j^* r_{j1}^*$ when the required attribute has not been mastered. In this case, because the second attribute is not required, mastery of Attribute 2 does not play a role in determining the probability of a correct response.

Given the probability of a correct response for all attribute patterns, $\mathbf{D}_j$ can be computed using equation (4). For example, to compute the Kullback-Leibler information comparing $P_{\{0,1\}}(\mathbf{X})$ to the assumed distribution $P_{\{1,1\}}(\mathbf{X})$, the following equation is used:

$$P_{\{1,1\}}(1)\log\left[\frac{P_{\{1,1\}}(1)}{P_{\{0,1\}}(1)}\right] + P_{\{1,1\}}(0)\log\left[\frac{P_{\{1,1\}}(0)}{P_{\{0,1\}}(0)}\right].$$

Through substitution, the Kullback-Leibler information is

$$(.80)\log\left[\frac{.80}{.10}\right] + (.20)\log\left[\frac{.20}{.90}\right] = 1.36.$$

In computing the Kullback-Leibler information for all other comparisons, the final $\mathbf{D}_j$ is

$$\mathbf{D}_j = \begin{pmatrix} 0 & 0 & 1.14 & 1.14\\ 0 & 0 & 1.14 & 1.14\\ 1.36 & 1.36 & 0 & 0\\ 1.36 & 1.36 & 0 & 0 \end{pmatrix}.$$

In $\mathbf{D}_j$, rows (and columns) 1-4 represent examinees who have not mastered either attribute, (0 0); examinees who have mastered only the second attribute, (0 1); examinees who have mastered only the first attribute, (1 0); and examinees who have mastered both attributes (1 1), respectively. The $i$, $j$ element of $\mathbf{D}_j$ is the Kullback-Leibler information of the $i$th attribute pattern versus the $j$th attribute pattern.

To compute $d_{(A)j11}$, only the elements that correspond to comparisons of examinee patterns (1 $x$) to (0 $x$) are considered, where $x$ is either a 1 or 0, as defined previously by $\Omega_{11}$. Specifically, only the bold elements in equation (14) are considered.

$$\mathbf{D}_j = \begin{pmatrix} 0 & 0 & 1.14 & 1.14\\ \underline{0} & 0 & 1.14 & 1.14\\ \mathbf{1.36} & 1.36 & 0 & 0\\ 1.36 & \mathbf{1.36} & \underline{0} & 0 \end{pmatrix}. \tag{14}$$

For example, $D_{31}$ represents the comparison of examinee pattern (1 0) to examinee pattern (0 0). Because $d_{(A)j11}$ is the average of the bold numbers,

$$\begin{aligned}
d_{(A)j11} &= \frac{1.36 + 1.36}{2}\\
&= 2.72/2\\
&= 1.36.
\end{aligned}$$

The discrimination index can also be computed for Attribute 2, $d_{(A)j21}$, using the underlined values in equation (14). Because the item does not require Attribute 2, $d_{(A)j21} = 0$. Using similar equations, $d_{(A)j10}$ and $d_{(A)j20}$ can be computed.

Next, to compute $d_{(B)j11}$, the same bold elements in equation (14) are used, only now it is assumed that information about the population is known or has been estimated. The index, $d_{(B)j11}$, is the weighted mean of the elements used for the index $d_{(A)j11}$. For this example, assume that a random examinee has attribute pattern (0 0) with probability .27, attribute pattern (0 1) with probability .43, attribute pattern (1 0) with probability .03, and attribute pattern (1 1) with probability .27. Therefore,

$$d_{(B)j11} = \frac{.03(1.36) + .27(1.36)}{.3} = 1.36.$$

Also, $d_{(B)j21} = 0$ and the indices $d_{(B)j10}$ and $d_{(B)j20}$ can be computed using similar equations.

## A Simulation Study of the Performance of the Two Indices

Using Monte Carlo simulation, random tests are generated as described below where items are calibrated using the RUM. For each test, $\mathbf{d}_{(A)j}$, $\mathbf{d}_{(B)j}$, and the responses of 10,000 simulated examinees are computed. In addition, attribute patterns of each simulated examinee are estimated, and correct classification rates for both the masters and nonmasters are computed. Next, correlations are computed between correct classification rates and the two discrimination indices in order to assess the performance of the indices.

As a device to simulate attributes where there are associated attributes with a fixed proportion of masters for each attribute, 10,000 multivariate normal $K$-dimensional vectors ($\tilde{\alpha} \sim MVN(0, \rho)$) are randomly generated (Henson & Douglas, 2005). Here $\rho$ represents a correlation matrix where all off-diagonal elements are equal. The $p_k$s were then used to define a cutoff, $\kappa_k = 0$, for each attribute so that $P(\tilde{\alpha} \leq \kappa_k) = p_k = .5$. The $i$th individual's mastery for attribute $k$ thus

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

These simulated attribute patterns then used to compute the probability of a correct response (based on the RUM) for each item, which then used to simulate the item responses.

In addition, a second sample of 10,000 attribute patterns generated and used to compute a Monte Carlo approximation of the prior distribution of attribute patterns, which is needed for Index B.

Because the item parameters will also be simulated and therefore are known, instead of using a Markov chain Monte Carlo approach to carry out a Bayesian analysis, Bayesian-based classification is accomplished by computing the likelihood for all possible attribute patterns given the examinees' scores and multiplying by the prior probabilities of attribute patterns (estimated from the second sample of 10,000 subjects). The Bayesian posterior mode is then used to classify the attribute pattern for that individual. Given the estimated attribute patterns, for each attribute the proportion of examinees for which the attribute was correctly classified is recorded.

Next, to determine the characteristics of the simulated tests, one must remember that the purpose of this study is to examine the relationship between each discrimination index and correct classification rates. Tests are generated that are realistic while intentionally creating variability of

the measurement quality of the tests (i.e., some tests have high correct classification rates, whereas others do not perform as well); 1,000 randomly generated 40-item tests are constructed to measure five attributes. On average, each item requires two attributes in each test. In addition, item parameters are generated such that tests will range from low to high cognitive structure. In this context, low cognitive structure will be defined as situations for which the absence of one or more of the required attributes has a relatively small influence on examinees' probability of a correct item response and therefore the items are at best moderately informative, whereas in high cognitive structure the probability of a correct response is strongly influenced by the presence or absence of one or more of the required attributes. The characteristics of the randomly generated item parameters for each test are as follows:

- $\pi^*$s are randomly generated from a uniform distribution, $U(.85, .95)$, for all 1,000 tests.
- $r^*$s are used to modify the cognitive structure. Specifically, for the $i$th simulation, $i = \{1, \ldots, 1{,}000\}$, $r^*$s are randomly generated from a uniform distribution, $U(.1 + \frac{.6(i-1)}{999}, .3 + \frac{.6(i-1)}{999})$. Notice that for the first simulation, $r^*$s are generated to resemble high cognitive structure (i.e., $r^*$s range from .1 to .3). For each simulation, the range slowly shifts to resemble a lower cognitive structure until the last simulated test contains $r^*$s that range from .7 to .9, low cognitive structure indeed.
- $c$s are all set to $\infty$; therefore, $P_c(\eta) = 1$ for all items.[2]

It is important to remember that each index incorporates the dependence between the attributes to a different degree. Specifically, $\mathbf{d}_{(A)j}$ totally ignores the association between attributes, and $\mathbf{d}_{(B)j}$ incorporates the association in the form of multiplicative weights (based on the prior probabilities) of the Kullback-Leibler information. Three different simulations of 1,000 tests are run: a simulation with all the off-diagonal elements of ρ equal to .5, a simulation with all of the off-diagonal elements of ρ equal to .75, and a simulation with all of the off-diagonal elements of ρ equal to .95.

## Results

For each of the three simulation studies, the basic descriptive statistics of the discrimination indices and correct classification rates are provided. In addition, the correlations between the discrimination indices and the appropriate correct classification rates are computed. The following paragraphs summarize the results from the three simulations.

To begin, the minimum, maximum, and mean values of correct classification rates over the 1,000 simulation replications, $\mathbf{d}_{(A)j}$, and $\mathbf{d}_{(B)j}$ for both masters and nonmasters, respectively, are summarized in Tables 1 to 3. It should be noted that because tests are randomly generated with all $p_k = .5$, and all attribute correlations are the same within a study, the results of the five attributes are indistinguishable. Therefore, the basic descriptive statistics will be summarized across all attributes, which provide more efficient estimates of their true values.

In general, although tests were developed to allow for a large range of correct classification rates, it is clear that as the correlation between attributes increases, there is a restriction of range for the values of correct classification rates from the simulations. For example, the correct classification rates for the simulation study with attributes that have correlations of .5 range from .75 to 1.00 with an average of approximately .92, whereas they range from approximately .90 to 1.00 with an average of .97 in the simulation study with attributes that have correlations of .95. In addition, although the intent of this study is only to define indices that correlate with correct classification rates, it can be seen that both $\mathbf{d}_{(A)j}$ and $\mathbf{d}_{(B)j}$ appear to be on similar scales.

Next, Table 4 provides the means of the correlations between the correct classification rates and $\mathbf{d}_{(A)j}$ and $\mathbf{d}_{(B)j}$ for the masters and nonmasters.[3] The table shows that in general, correlations

**Table 1**
Basic Descriptive Statistics When the Correlation Between Attributes is .5

| | Index | *Min* | *M* | *Max* |
|---|---|---|---|---|
| Master | CC | 0.75 | 0.92 | 1.00 |
| | $\mathbf{d}_{(A)jk1}$ | 0.57 | 4.92 | 17.81 |
| | $\mathbf{d}_{(B)jk1}$ | 0.63 | 5.41 | 19.36 |
| Nonmaster | CC | 0.75 | 0.93 | 1.00 |
| | $\mathbf{d}_{(A)jk0}$ | 0.69 | 5.97 | 19.93 |
| | $\mathbf{d}_{(B)jk0}$ | 0.78 | 6.76 | 21.83 |

*Note.* CC = correct classification.

**Table 2**
Basic Descriptive Statistics When the Correlation Between Attributes is .75

| | Index | *Min* | *M* | *Max* |
|---|---|---|---|---|
| Master | CC | 0.80 | 0.93 | 1.00 |
| | $\mathbf{d}_{(A)jk1}$ | 0.60 | 4.92 | 15.73 |
| | $\mathbf{d}_{(B)jk1}$ | 0.67 | 5.57 | 18.24 |
| Nonmaster | CC | 0.84 | 0.95 | 1.00 |
| | $\mathbf{d}_{(A)jk0}$ | 0.72 | 5.96 | 16.82 |
| | $\mathbf{d}_{(B)jk0}$ | 0.83 | 7.01 | 20.25 |

*Note.* CC = correct classification.

**Table 3**
Basic Descriptive Statistics When the Correlation Between Attributes is .95

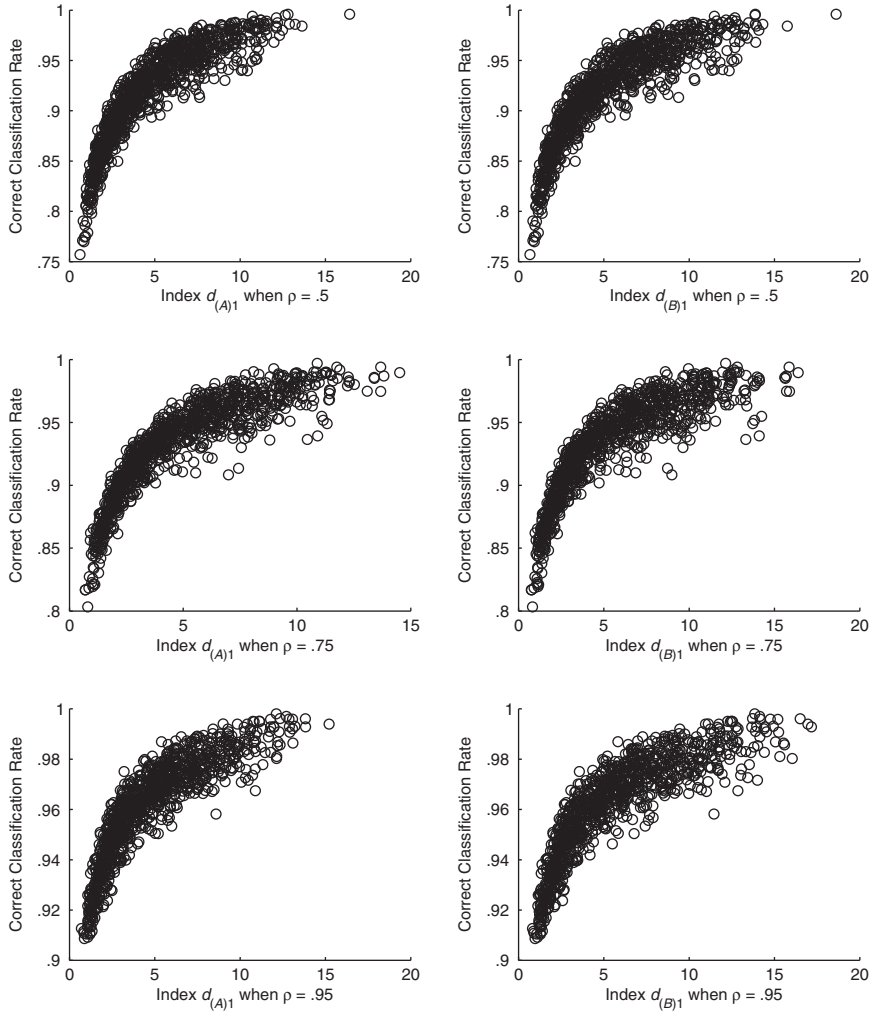
| | Index | *Min* | *M* | *Max* |
|---|---|---|---|---|
| Master | CC | 0.89 | 0.96 | 1.00 |
| | $\mathbf{d}_{(A)jk1}$ | 0.64 | 4.91 | 15.82 |
| | $\mathbf{d}_{(B)jk1}$ | 0.68 | 5.67 | 17.70 |
| Nonmaster | CC | 0.92 | 0.98 | 1.00 |
| | $\mathbf{d}_{(A)jk0}$ | 0.77 | 5.96 | 17.38 |
| | $\mathbf{d}_{(B)jk0}$ | 0.86 | 7.17 | 20.00 |

*Note.* CC = correct classification.

**Table 4**
Correlation Between the Discrimination Indices and Correct Classification

| | Study | $r_A$ | $r_B$ |
|---|---|---|---|
| Masters | Attribute correlation .5 | .87 | .86 |
| | Attribute correlation .75 | .87 | .85 |
| | Attribute correlation .95 | .85 | .83 |
| Nonmasters | Attribute correlation .5 | .90 | .90 |
| | Attribute correlation .75 | .88 | .88 |
| | Attribute correlation .95 | .83 | .83 |

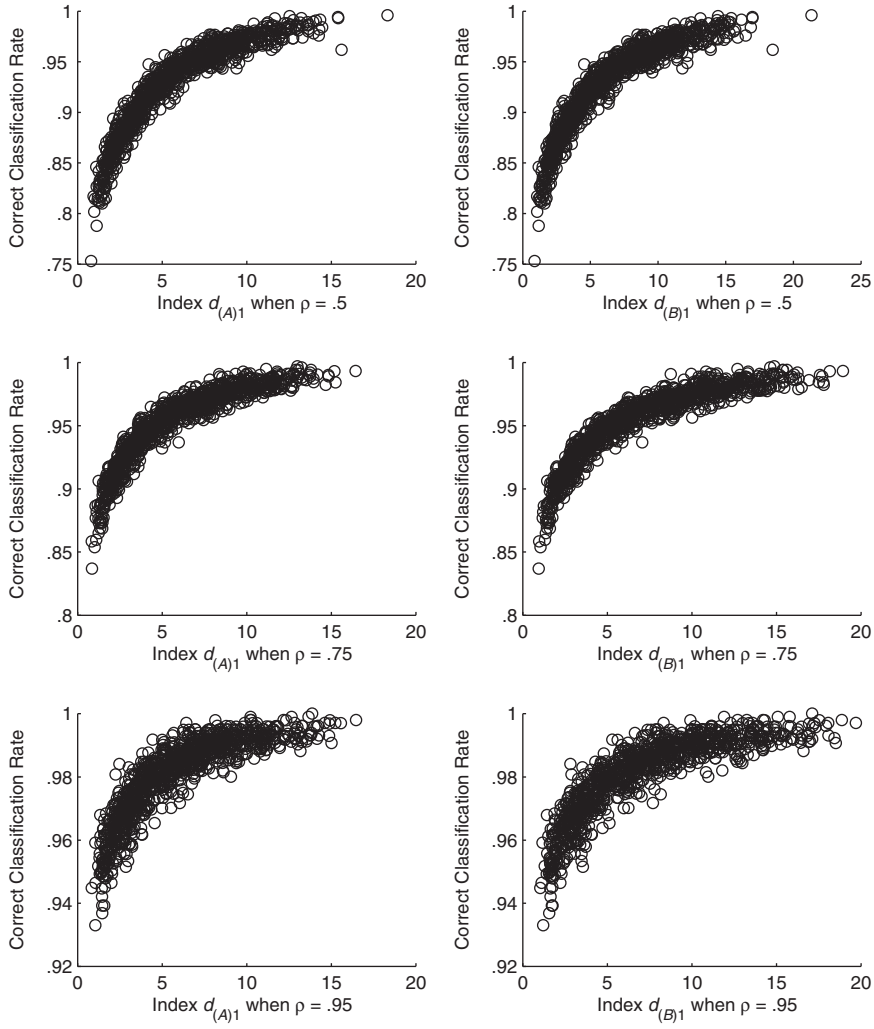
**Figure 1**
Scatterplots of the Discrimination Indices With Correct Classification for Masters




are quite high; therefore, it is reasonable to use the discrimination indices as indicators of correct classification rates.

One assumption when using the correlation is that the relationship is approximately linear. Therefore, it is important that the scatterplots be explored for linearity. Figures 1 and 2 are examples of scatterplots used to explore the assumption of linearity between discrimination and correct classification rates for masters (Figure 1) and nonmasters (Figure 2). Columns 1 and 2 represent the plots for the discrimination indices $\mathbf{d}_{(A)j}$ and $\mathbf{d}_{(B)j}$, respectively, crossed with the rows, which represent the three simulations (i.e., when correlations between attributes are .5, .75, and .95).

**Figure 2**
Scatterplots of the Discrimination Indices With Correct Classification for Nonmasters




Clearly, the relationship is not linear due to the asymptotic effect of correct classification rates (i.e., they approach 1). Therefore, the true relationship is stronger than what is indicated by the correlation coefficients (i.e., if the values were transformed, such that the relationship is linear, correlations will be higher). It is also possible that some correlations are smaller due to a restricted range, which may explain the reduction of the correlations for the simulation where all attributes have a correlation of .95.

To explore the true strength of the relationship between the discrimination, or a monotonic function of discrimination, and correct classification, the log transformation of the discrimination indices can be used so that the relationship is linear. Table 5 shows the correlation between the

**Table 5**
Correlation Between the Transformed Discrimination Indices and Correct Classification

| | Study | $r_A$ | $r_B$ |
|---|---|---|---|
| Masters | Attribute correlation .5 | .97 | .97 |
| | Attribute correlation .75 | .95 | .94 |
| | Attribute correlation .95 | .93 | .92 |
| Nonmasters | Attribute correlation .5 | .95 | .94 |
| | Attribute correlation .75 | .96 | .96 |
| | Attribute correlation .95 | .93 | .92 |

logarithm of each discrimination index and correct classification rates, again providing strong evidence of a useful relationship.

## Discussion

The results provide strong evidence that both indices are good candidates as possible indicators for an attribute's correct classification rates for any given test. Given this relationship, it is now possible to define the discriminating power and, hence, the usefulness of each item for accurately estimating each attribute. Those items with a high discrimination index for an attribute contribute more to the estimation of that attribute than those with small values. In addition, as in the case of the Henson and Douglas (2005) CDI, the set of $2K$ attribute discrimination indices for each item can be used to construct an effective attribute diagnostic test from an item bank. Specifically, by selecting items where the test attribute discrimination index is large for all attributes, the test will have high correct classification rates for every attribute when compared to all possible tests that can be constructed from the same item bank. Although optimization with multiple objective functions has limitations, simple methods that have previously been implemented in IRT, where the goal is to develop a test that fits a specific information curve, can be implemented (Henson, 2004).

A future study will compare test construction using the two indices presented in this article to test construction using the CDI. The current study does not vary the proportion of masters for each of the five attributes. Because correct classification rates and their relationship to the indices may change as a function of the proportion of masters for an attribute, additional studies should generalize this study to include values other than .5.

## Notes

1. Notice that by excluding the $P_{c_j}$ term from the model the reparameterized unified model is also a constrained latent class model. However, if this term is included, then the model is a constrained finite mixture model, where responses conditional on the mastery profile (class membership) are dependent.
2. This specification is used to explore the usefulness of the Kullback-Leibler information when there are only classes. A future study will explore the possibility of integrating across $\eta$ in the computation of these indices when $c$ is in the model.
3. All standard errors are less than or equal to .01.

## References

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

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

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.

Henson, R. (2004). *Test discrimination and test construction for cognitive diagnostic models*. Unpublished doctoral dissertation, University of Illinois.

Henson, R., & Douglas, J. (2005). Test construction for cognitive diagnostics. *Applied Psychological Measurement*, *29*, 262-277.

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.

Lord, F. M. (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.

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

Rindskopf, D. (1983). A general framework for using latent class analysis to test hierarchical and nonhierarchical learning models. *Psychometrika*, *48*, 85-97.

Stout, W. (2002). Psychometrics: From practice to theory. *Applied Psychological Measurement*, *29*, 262-277.

## Acknowledgments

*The research reported here was completed under the auspices of the External Diagnostic Research Team, supported by the Educational Testing Service.*

## Author's Address

Address correspondence to Robert Henson, The University of North Carolina at Greensboro, Department of Educational Research Methodology, 207 Curry Building, PO Box 26170, Greensboro, NC 27402-6170; e-mail: rahenson@uncg.edu.