Modified Cognitive Diagnostic Index and Modified Attribute-Level Discrimination Index for Test Construction

Applied Psychological Measurement
2016, Vol. 40(5) 315–330
© The Author(s) 2016
Reprints and permissions:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/0146621616638643
apm.sagepub.com

Bor-Chen Kuo¹, Hsiao-Shan Pai¹, and Jimmy de la Torre²

Abstract

At present, there are only a limited number of studies examining how to optimally construct cognitive diagnostic tests. The cognitive diagnostic index (CDI) and attribute-level discrimination index (ADI) have been proposed to assemble such tests. The CDI and ADI have been shown to be instrumental in constructing cognitive diagnostic tests when the attribute relationships are assumed to be nonhierarchical. For greater generality when designing cognitive diagnostic assessment, attribute hierarchy and the ratio of test length to the number of attributes (RTA) are two important factors to be considered. This article proposes modified indices that take into account attribute hierarchy and RTA. Simulation studies show that, under the deterministic input, noisy, “and” gate model (DINA) and the reduced version of the reparameterized unified model (rRUM), the proposed indices provide higher attribute and attribute pattern correct classification rates than the original indices.

Keywords

cognitive diagnosis, test construction, attribute hierarchy, cognitive diagnosis index, attribute-level discrimination index

Introduction

Cognitive diagnosis models (CDMs) belong to a class of statistical models that can be used to assess multiple latent traits. In CDMs, examinees are typically characterized by a set of dichotomous skills or attributes. Each attribute pattern represents the attributes that an examinee has mastered or has not mastered. At present, several studies exist providing guidance on how tests can be constructed from a CDM perspective. First, the cognitive diagnostic index (CDI; Henson & Douglas, 2005) and the attribute-level discrimination index (ADI; Henson, Roussos, Douglas, & He, 2008) can be used to accomplish such a purpose. CDI measures an item’s overall discrimination power by using Kullback–Leibler information and serves as a measure of how

¹National Taichung University of Education, Taiwan

²Rutgers, The State University of New Jersey, New Brunswick, USA

Corresponding Author:

Hsiao-Shan Pai, Graduate Institute of Educational Information and Measurement, National Taichung University of Education, No. 140, Minsheng Road, West Dist., Taichung City 40306, Taiwan.

Email: mishare1127@yahoo.com.tw

informative an item is in correctly classifying the examinees' true status. However, CDI does not provide any information about an item's discrimination power for a specific attribute (Henson et al., 2008). To address this issue, ADI was developed to indicate the discrimination power of an item with respect to each of the attributes. Second, genetic algorithm (GA; Finkelman, Kim, & Roussos, 2009) and binary programming (BP; Finkelman, Kim, Roussos, & Verschoor, 2010) have been proposed for test construction purposes in the context of CDMs. Finkelman et al. (2010) provided guidelines on how to choose from different test construction methods, namely, CDI, BP, and GA. In particular, as the only feasible method, GA is recommended when specific correct classification rates (CCRs) are of interest. When targeting specific CCRs that are not specified or may not be feasible (e.g., when test length is fixed and relatively short), GA can still be used to minimize the average error rate or the maximum attribute-level error rate so long as the procedure remains computationally viable. However, when GA becomes computationally burdensome, CDI and BP can be used to accomplish the former and latter goals, respectively. It can be noted that these guidelines do not cover the use of ADI. In addition, although the usefulness of CDI and ADI has been demonstrated in Henson and Douglas (2005) and Henson et al. (2008), respectively, the performance of these two indices has yet to be compared. More importantly, the different test construction procedures discussed above have been studied only in the context of independent and correlated attribute structures, but not in the context of hierarchically structured attributes. To address this gap in the literature, this study will focus on comparing and addressing the limitations of CDI and ADI as indices for test construction when the attributes follow a hierarchical structure.

With most CDMs, the Q-matrix (Tatsuoka, 1983) is used to identify the skills or attributes required to answer each item correctly. Studies show that a subset of the attribute patterns will not occur when attributes have hierarchical structures (de la Torre, Hong, & Deng, 2010; Leighton, Gierl, & Hunka, 2004). The design of CDI and ADI implicitly assumes that attributes have a nonhierarchical relationship. For this reason, it is not clear that constructing a test of fixed length using the CDI method can provide optimal attribute classification rates across different attribute structures (de la Torre et al., 2010). In this study, some drawbacks of CDI and ADI are noted when attributes have hierarchical structures.

Cheng (2010) underlined the importance of ensuring that each attribute is measured by an adequate number of items. This implies that the ratio of test length to the number of attributes (RTA) needs to be considered when constructing a cognitive diagnostic test. However, CDI and ADI do not take RTA into consideration. Consequently, tests based on these indices may not measure all the attributes with the same level of accuracy.

The goal of this study is to address the shortcomings of CDI and ADI by modifying these indices to explicitly account for attribute hierarchy and RTA. Four simulation studies based on the setting of previous research were conducted to evaluate and compare the performance of the original and proposed indices.

Background

CDM

Many CDMs have been proposed in the literature. The common CDMs include deterministic input, noisy, "and" gate model (DINA; Haertel, 1989; Junker & Sijtsma, 2001); noisy input, deterministic, "and" gate model (NIDA; Junker & Sijtsma, 2001); multiple classification latent class model (MCLCM; Maris, 1999); the reparameterized unified model (RUM; Roussos et al., 2007); and the log-linear cognitive diagnosis model (LCDM; Henson, Templin, & Willse, 2009). In their original study, Henson and Douglas (2005) used the reduced version of the reparameterized unified model (rRUM) and the DINA model. For comparative purposes, this study also

focuses only on the DINA model and rRUM. The following is a brief introduction on the DINA model and rRUM.

The DINA model assumes that to answer an item correctly, each attribute measured by the item must be successfully applied. The DINA model includes two different item parameters, s_j and g_j . In the DINA model, the probability of a correct response can be written as

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

where $\eta_{ij} = \prod_{k=1}^K \alpha_{ik}^{q_{jk}}$ is an indicator of whether examinee i has mastered all of the required attributes for item j ; s_j , the slip parameter, is the probability that an examinee who has all the required attribute misses the item j ; and g_j , the guessing parameter, is the probability that an examinee who lacks at least one of the required attributes answers item j correctly.

The rRUM also assumes that to answer an item correctly, each attribute measured by the item must be successfully applied albeit using a different process. The rRUM includes two different types of item parameters, π_j^* and r_{jk}^* . The probability of a correct response for the examinee i given α_i is

$$P(X_{ij} = 1 | \alpha_i, \eta_i) = \pi_j^* \prod_{k=1}^K r_{jk}^{*(1 - \alpha_{ik} \eta_{jk})}. \quad (2)$$

where π_j^* represents the probability of correctly applying all required attributes for item j by examinees who have all the required attributes, and r_{jk}^* is the penalty for not mastering attribute k , and its inverse is related to the discrimination power of the k th attribute of item j . Hence, the probability of a correct response for an examinee who has mastered all the required attributes will be higher than that of a correct response for an examinee who lacks one or more required attribute.

CDI and ADI

Henson and Douglas (2005) proposed the CDI for test construction in the context of CDMs. CDI, an alternative to Fisher information, utilizing the concept of Kullback–Leibler information (Chang & Ying, 1996), is defined as follows:

$$CDI_j = \frac{\sum_{u \neq v} [h(\alpha_u, \alpha_v)^{-1} D_{juv}]}{\sum_{u \neq v} h(\alpha_u, \alpha_v)^{-1}}, \quad (3)$$

where

$$h(\alpha_u, \alpha_v) = \sum_{k=1}^K (\alpha_{uk} - \alpha_{vk})^2, \quad (4)$$

and

$$D_{juv} = E_{\alpha_u} \left[\log \left[\frac{P_{\alpha_u}(X_j)}{P_{\alpha_v}(X_j)} \right] \right] = P_{\alpha_u}(1) \log \left[\frac{P_{\alpha_u}(1)}{P_{\alpha_v}(1)} \right] + P_{\alpha_u}(0) \log \left[\frac{P_{\alpha_u}(0)}{P_{\alpha_v}(0)} \right]. \quad (5)$$

In Equation 5, α_u and α_v are $1 \times K$ attribute vectors (i.e., $\alpha_u = [\alpha_{u1}, \dots, \alpha_{uK}]$ and $\alpha_v = [\alpha_{v1}, \dots, \alpha_{vK}]$), $P_{\alpha_u}(1)$ and $P_{\alpha_v}(1)$ are the probabilities of a correct response given α_u and α_v , respectively, and $P_{\alpha_u}(0)$ and $P_{\alpha_v}(0)$ are the corresponding probabilities of an incorrect response. Computed from the Kullback–Leibler matrix, $\mathbf{D}_{juv}$ functions as an indicator of how well α_u is measured when compared with α_v . Item j is more useful in discriminating between the attribute patterns u and v if $\mathbf{D}_{juv}$ is larger (Henson & Douglas, 2005). From Equation 3, CDI_j can be viewed as a weighted average of the elements in the $\mathbf{D}_j$ matrix. Because it is more difficult to distinguish patterns that are more similar to each other, Henson and Douglas (2005) defined the weight as the inverse of the squared Euclidean distance, where attribute patterns that are more similar are weighted higher than those that are easily distinguishable from each other. CDI_j can be summed across the I items to form the test-level CDI, as in, $\text{CDI} = \sum_{j=1}^I \text{CDI}_j$. To construct a test with a good discrimination between mastery and nonmastery, items with large CDI_j should be selected first.

Although the CDI measures an item's overall discrimination power, it does not indicate an item's discrimination power for a specific attribute. To address this limitation, Henson et al. (2008) proposed an index, $\mathbf{D}_j$, to denote the discrimination power of an item for a specific attribute. Attribute patterns that differ only by one attribute are the most difficult to distinguish. Therefore, only such attribute patterns will be used to compute the discrimination index $\mathbf{D}_j$ (Henson et al., 2008). To combine the indices of mastery and nonmastery, define ADI_j as

$$\text{ADI}_j = \frac{d_{j1} + d_{j0}}{2} = \frac{\sum_{k=1}^K d_{jk1} + \sum_{k=1}^K d_{jk0}}{2K_j^*}, \quad (6)$$

where $d_{jk1} = \sum_{\Omega_{k1}} w_{k1} \mathbf{D}_{juv}$, $d_{jk0} = \sum_{\Omega_{k0}} w_{k0} \mathbf{D}_{juv}$, $\Omega_{k1} \equiv \{(\alpha_u, \alpha_v) | \alpha_{uk} = 1 \text{ and } \alpha_{vk} = 0 \text{ and } \alpha_{um} = \alpha_{vm} \forall m \neq k\}$, and $\Omega_{k0} \equiv \{(\alpha_u, \alpha_v) | \alpha_{uk} = 0 \text{ and } \alpha_{vk} = 1 \text{ and } \alpha_{um} = \alpha_{vm} \forall m \neq k\}$.

For Item j , d_{jk1} is the power to discriminate masters from nonmasters on the k th attribute, whereas d_{jk0} is the power to discriminate nonmasters from masters on the k th attribute, Ω_{k1} and Ω_{k0} are the sets of attribute pattern pairs (α_u, α_v) that differ only on the k th attribute, and w_{k1} is defined as the joint probability of attribute patterns given that the examinee has mastered the k th attribute, $P(\alpha | \alpha_k = 1)$, and w_{k0} is defined as the joint probability of attribute patterns given that the examinee has not mastered the k th attribute, $P(\alpha | \alpha_k = 0)$. Notice that if Item j does not measure the k th attribute, then the item does not contain any information about attribute mastery or nonmastery of the k th attribute. As with CDI_j , ADI_j can be added across the I items to form the test-level ADI, as in $\text{ADI} = \sum_{j=1}^I \text{ADI}_j$.

The CDI and ADI have been evaluated separately in Henson and Douglas (2005) and Henson et al. (2008), respectively. The results of simulation studies show that CDI-based tests outperform randomly constructed tests. Specifically, the former provides higher attribute and attribute pattern CCRs compared with the latter. A strong relationship between ADI and CCRs has also been observed.

As currently defined, CDI and ADI have a few important limitations. One limitation of the indices is that both indices assume that the attributes have no hierarchical structure. However, in some domains, some hierarchies between attributes may exist (Leighton et al., 2004). The following example shows that if attributes have a linear hierarchy, then CDI and ADI may perform poorly under some conditions. Suppose there are three attributes A_1 , A_2 , and A_3 (labeled 1, 2, and 3) with a linear hierarchy as shown in Figure 1. The hierarchy means that A_1 is a prerequisite to A_2 , which in turn is a prerequisite to A_3 .

The set of potential items is described by the incidence matrix $\mathbf{Q}$ of order (K, i) , where K is the number of attributes and i is the number of potential items equal to $2^K - 1$. When K is equal to 3, the $\mathbf{Q}$ -matrix is shown below:

$$\mathbf{Q}^T = \begin{pmatrix} 1 & 0 & 0 & 1 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 & 1 & 1 & 1 \end{pmatrix}. \quad (7)$$

For some CDMs (e.g., DINA model), the set of potential items can be reduced when the attributes are related hierarchically. For the example above, the binary representation of Item 2 is (010) indicating the item probes Attribute 2. However, the hierarchy indicates that Attribute 2 requires Attribute 1 as its prerequisite. Thus, Item 2 must be represented by (110). For this reason, Item 2 can be removed. The removal of items in this manner produces a reduced $\mathbf{Q}$ -matrix $\mathbf{Q}_r$. The reduced $\mathbf{Q}$ -matrix contains only three attribute specifications, $\mathbf{q}_1$, $\mathbf{q}_2$, and $\mathbf{q}_3$. The linear hierarchy also implies a reduced number of possible attribute patterns of examinees α ; $\mathbf{Q}_r$ and α are shown in the following.

$$\mathbf{Q}_r = \begin{pmatrix} \mathbf{q}_1 \\ \mathbf{q}_2 \\ \mathbf{q}_3 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \end{pmatrix} \text{ and } \alpha = \begin{pmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \end{pmatrix}. \quad (8)$$

For illustration purposes, assume $s_j = g_j = .05$ for all items. $\mathbf{D}_j$ can be computed using Equation 5, and the resulting $\mathbf{D}_{juv}$ s for the three $\mathbf{q}$ -vectors of item j are

$$D_{1uv} = \begin{pmatrix} 0 & 2.65 & 2.65 & 2.65 \\ 2.65 & 0 & 0 & 0 \\ 2.65 & 0 & 0 & 0 \\ 2.65 & 0 & 0 & 0 \end{pmatrix}, \quad D_{2uv} = \begin{pmatrix} 0 & 0 & 2.65 & 2.65 \\ 0 & 0 & 2.65 & 2.65 \\ 2.65 & 2.65 & 0 & 0 \\ 2.65 & 2.65 & 0 & 0 \end{pmatrix},$$

$$\text{and } D_{3uv} = \begin{pmatrix} 0 & 0 & 0 & 2.65 \\ 0 & 0 & 0 & 2.65 \\ 0 & 0 & 0 & 2.65 \\ 2.65 & 2.65 & 2.65 & 0 \end{pmatrix}. \quad (9)$$

The CDIs of $\mathbf{q}_1$, $\mathbf{q}_2$, and $\mathbf{q}_3$ are 1, 1.43, and 1.12, respectively. For this example, assume further that there is a large enough item bank from which to construct the test. If CDI is used to assemble a test, then only items with $\mathbf{q}_2$ will be selected because these items have the largest CDI. As a result, the third attribute will not be measured. Note that the ADIs of $\mathbf{q}_1$, $\mathbf{q}_2$, and $\mathbf{q}_3$ are 2.65, 1.33, and 0.88, respectively. If the test construction procedure is solely based on ADI, the second and third attributes will not be measured because only items measuring the first attributes, which corresponds to items with the largest ADI, will be selected.

Another shortcoming of CDI and ADI is that if discriminations of the items are similar, CDI and ADI tend to select single-attribute items regardless of test length. A previous study (Pai, Kuo, & Chen, 2012) shows that when $I = 10$ and $K = 5$, the accuracy of a test with both one- and two-attribute items is better than that of a test with one-attribute items only.

Although Henson and Douglas (2005) also employed item constraints and attribute constraints for test construction, no general rules and explanation about the setting of these constraints are stated in their study. The goal of this article is to address these two shortcomings.

Figure 1. An example of linear hierarchy.

Modified CDI and Modified ADI

In this section, two more comprehensive and general indices, the modified cognitive diagnostic index (MCDI) and the modified attribute-level discrimination index (MADI), that consider RTA and attribute hierarchy are proposed. The MCDI and MADI are defined as follows:

$$\text{MCDI}_j = w_j^L w_j^H \text{CDI}_j \quad \text{and} \quad \text{MADI}_j = w_j^L w_j^H \text{ADI}_j, \quad (10)$$

where CDI_j and ADI_j are as defined in Equations 3 and 6,

$$w_j^L = \left(1 + I(r_L < 3) \sum_{v=1}^V I(q_j^* = s_v) \right)^{-1}, \quad \text{and} \quad w_j^H = \left(1 + r_H \sum_{v=1}^V I(q_j^* = s_v) \right)^{-1}. \quad (11)$$

In these equations, w_j^L is the weight for the RTA, and w_j^H is the weight for attribute hierarchy, and q_j^* be the attribute specification of Item j , which has not been selected. Let $S = \{s_1, \dots, s_V, \dots, s_V\}$ as the set of attribute patterns for items that have been selected. When $q_j^* = s_v$, CDI_j and ADI_j will be multiplied by the weights of the RTA and the attribute hierarchy; when $q_j^* \neq s_v$, both weights will be equal to 1. Items in the bank which have the same attribute specifications as the selected items would be given a smaller weight; in contrast, items in the bank, which have different attribute specifications, would have the full weight (i.e., 1).

Let the RTA be $r_L = I/K$. Kuo, Wu, and Shih (2012) and Liu (2013) indicated that each attribute in the test must be measured at least 3 times to attain better correct attribute classification. When test length is sufficiently long to let each attribute be measured at least 3 times, then $r_L \geq 3$. In this situation, it is not necessary to weight the indices. As such, w_j^L will be equal to 1. But when the Item j has the same attribute specification as some of the items already selected, and test length is relatively short to let each attribute be measured at least 3 times, w_j^L will be equal to .5.

Attribute hierarchy is another important factor that affects test construction. Depending on the attribute hierarchy, a reachability matrix $\mathbf{R}$ can be obtained (Leighton et al., 2004). $\mathbf{R} = [r_{mn}]$ is a $K \times K$ matrix, where the element r_{mn} is equal to 1 when attribute m is a prerequisite of attribute n . Define

$$r_H = \frac{\sum_{m \leq n} r_{mn}}{K(K+1)/2}, \quad 1 \leq m, n \leq K \quad (12)$$

as a measure of the hierarchical relationship between the attribute. This index is $0 < r_H \leq 1$ and is larger with more layers in the attribute hierarchy. When the attributes have a linear hierarchy, $r_H = 1$; when the attributes have a nonhierarchical relationship, $r_H = 2K / [K(K+1)]$.

Given in Figure 2 is a flowchart of how the MCDI is used in the test construction process. To construct a test using MCDI, the item with the highest CDI was chose first. After which, the MCDIs of the remaining items in the pool are (re)calculated. Of the remaining items, the item

Figure 2. A flowchart of the test construction algorithm based on MCDI.
 Note. MCDI = modified cognitive diagnostic index; CDI = cognitive diagnostic index.

with the highest MCDI was chosen. The process of recalculating the MCDI and adding an item to the test was continued until the target test length has been reached.

A flowchart for updating the MCDI of Item j is given in Figure 3. Weights are applied to CDI to derive the MCDI depending on the q -vector of Item j and the length of the test relative to the number of attributes. The same steps can be followed when using MADI in place of MCDI.

Simulation Studies

Four simulation studies were carried out to examine how MCDI and MADI perform under various conditions. In addition, the performance of the proposed indices was also compared against that of CDI and ADI.

Design

In this section, six attribute hierarchies $H_0, H_1, \dots, H_5$ were considered to examine how the proposed indices perform when attributes exhibit different hierarchies (refer to Figure 4 for the six hierarchies used in this study).

For this study, the number of attributes was fixed to $K = 7$. The reduced Q -matrix described above was used with the DINA model; in contrast, the full Q -matrix (i.e., all possible q -vectors) was used in the rRUM. The linear hierarchy was used (H_5) to illustrate how the items banks

Figure 3. A flowchart on how MCDI of item j is recalculated.

Note. MCDI = modified cognitive diagnostic index; CDI = cognitive diagnostic index.

were constructed in the DINA model. If the attributes have a nonhierarchical relationship (i.e., H_0), the number potential item types (i.e., items with unique attribute specifications) is equal to $2^K - 1$, which is 127 when $K = 7$. However, as mentioned earlier, the number of attribute specifications can be reduced when the attributes are related in a hierarchy. If the attribute hierarchy follows H_5 , the resulting reduced Q-matrix $\mathbf{Q}_r$ has only seven unique item types. For the remaining hierarchies, the numbers of unique item types were 64 for H_1 and H_2 , and 26 for H_3 and H_4 .

Studies 1 and 3 involved fixed item parameters, whereas Studies 2 and 4 involved randomly generated item parameters. For each item type, 30 items were generated resulting in the following item bank sizes: 3,810 (H_0); 1,920 (H_1 and H_2); 750 (H_3 and H_4); and 210 (H_5) in the DINA model. Only one item bank size (3,810) was used in rRUM. To minimize the impact of the Monte Carlo error, 100 item banks were constructed for each study. However, due to the nature of the item parameters, only in Studies 2 and 4 were the 100 item banks different from each other.

To generate the item responses, two of the studies (1 and 2) used the DINA model and the other two studies (3 and 4) used the rRUM. For Study 1, the slip and the guessing parameters in the DINA model were set to be $s_j = g_j = .05$; for Study 2, the slip and guessing parameters were randomly generated from the uniform distribution, $U(.05, .4)$; for Study 3, the item parameters were set to $\pi_j^* = 0.95$ and $r_{jk}^* = 0.2$; and for Study 4, the π_j^* and r_{jk}^* were randomly generated from $U(.85, .95)$ and $U(.1 + .6(i - 1)/29, .3 + .6(i - 1)/29)$, $i = \{1, \dots, 30\}$, respectively. It should be noted that the item parameter settings in Studies 2 and 4 are based on Henson and Douglas (2005) and Henson et al. (2008).

In each study, tests of length I from an item bank of size J were constructed based on different indices, and test lengths $I = 10, 20$, and 30 were used. Items were selected sequentially from the item bank across different attribute hierarchies and CDMs until I items were selected. For all algorithms, items with the largest CDI, ADI, MCDI, and MADI were selected first; for the proposed method, MCDI and MADI of the items left in the bank were recalculated as the test construction progressed.

In the simulation studies, the total number of examinees was set to be 16,640. To achieve this, responses of 130 (H_0), 256 (H_1 and H_2), 640 (H_3 and H_4), and 2,080 (H_5) for each permissible

Figure 4. Attribute hierarchies in the simulation studies.

Table 1. Simulation Design.

Simulation conditions	Setting
Attribute hierarchy	$H_0, H_1, H_2, H_3, H_4, H_5$
Number of items in each item bank (J)	DINA: H_0 :3,810; H_1 and H_2 :1,920; H_3 and H_4 :750; H_5 :210 rRUM: H_0 : 3,810
Model	DINA, rRUM
Number of examinees per attribute pattern	H_0 : 130, H_1, H_2 : 256, H_3, H_4 : 640, H_5 : 2,080
Test construction method	Random, CDI, ADI, MCDI, MADI
Test length (I)	10, 20, 30
Estimation method	EAP
Replications	100

Note. DINA = deterministic input, noisy, “and” gate model; rRUM = reduced version of the reparameterized unified model; CDI = cognitive diagnostic index; ADI = attribute-level discrimination index; MCDI = modified cognitive diagnostic index; MADI = modified attribute-level discrimination index; EAP = expected a posteriori.

attribute pattern were generated. In these studies, the item parameters were assumed to be known, therefore, not estimated. Examinee classification was based on expected a posteriori (EAP). Finally, based on estimated attribute patterns, the attribute correct classification rates (ACCR) and the attribute pattern correct classification rates (PCCR) were computed and compared. This process from item response generation to computing the CCR was replicated 100 times.

The design of the current simulation studies is summarized in Table 1.

Results

Test Construction Based on the DINA Model

The results of the test construction based on the DINA model are summarized in Figures 5 (Study 1) and 6 (Study 2). Figures 5a, 5c, and 5e show the ACCR when test lengths are equal

Figure 5. ACCR and PCCR based on the DINA Model in Study 1.
Note. ACCR = attribute correct classification rate; PCCR = pattern correct classification rate; DINA = deterministic input, noisy, “and” gate model; CDI = cognitive diagnostic index; ADI = attribute-level discrimination index; MCDI = modified cognitive diagnostic index; MADI = modified attribute-level discrimination index.

to 10, 20, and 30, respectively; Figures 5b, 5d, and 5f show the PCCR when test lengths are equal to 10, 20, and 30, respectively.

Overall, in Study 1, MCDI and MADI outperformed Random, CDI, and ADI with respect to ACCR and PCCR. When RTA was small, MCDI and MADI substantially outperformed Random, CDI, and ADI. For example, when test length was equal to 10 ($RTA < 3$) and H_0 was true, the ACCR of CDI, ADI, MCDI, and MADI were .87, .87, .95, and .95, respectively; the PCCR of CDI, ADI, MCDI, and MADI are .34, .35, .71, and .71.

When the attributes had hierarchical structures (H_1 - H_5), CDI and ADI performed poorly, except for the 30-item case and H_1 was true. In addition, the results of ADI were very unstable. In all conditions, MCDI and MADI performed well—when $I = 20$ or 30, the ACCR of MCDI and MADI were both higher than .98; when $I = 10$, the ACCR of MCDI and MADI were all

Figure 6. ACCR and PCCR based on the DINA Model in Study 2.
Note. ACCR = attribute correct classification rate; PCCR = pattern correct classification rate; DINA = deterministic input, noisy, “and” gate model; CDI = cognitive diagnostic index; ADI = attribute-level discrimination index; MCDI = modified cognitive diagnostic index; MADI = modified attribute-level discrimination index.

higher than .93. However, it should be noted that the performance of the random method was sometimes better than CDI and ADI when the attributes have hierarchical structures.

The results of Study 2 were similar to those of Study 1. MCDI and MADI outperformed the original indices and random method. When attributes had a nonhierarchical relationship (i.e., H_0), ADI performed poorly in all situations. With shorter tests, MCDI and MADI performed better than CDI, and the gaps between these methods were more obvious. When the attributes were hierarchically structured, the ACCR and PCCR of ADI were unstable and can be worse than those of other indices. The results of CDI were more stable and better than those of ADI, but still worse than those of MCDI and MADI.

Clearly, the result based on the DINA model showed that the proposed indices had substantial improvement when RTA was small or attributes had a hierarchical relationship. As test

Figure 7. ACCR and PCCR based on the rRUM in Study 4.

Note. ACCR = attribute correct classification rate; PCCR = pattern correct classification rate; rRUM = reduced version of the reparameterized unified model; CDI = cognitive diagnostic index; ADI = attribute-level discrimination index; MCDI = modified cognitive diagnostic index; MADI = modified attribute-level discrimination index.

length decreased or the number of layers in the attribute hierarchy increased, the limitations of the original indices became more evident. In most situations, MADI was a little better than MCDI, except for some cases, and when H_2 and H_5 were true.

Test Construction Based on the rRUM

The results for the two rRUM studies were very similar. Due to space limitation, only the results for Study 4 (see Figure 7) will be discussed. The results in their entirety can be requested from the first author. Figures 7a, 7c, and 7e show the ACCR when $I = 10, 20$, and 30 , respectively; Figures 7b, 7d, and 7f show PCCR when $I = 10, 20$, and 30 , respectively.

Table 2. Number of Times k -Attribute Items Are Used Under H_0 and DINA Model With $s_j = g_j = .05$.

I	Method	ACCR	PCCR	k						
				1	2	3	4	5	6	7
10	Random	0.73	0.18	0.7	1.8	2.5	2.5	1.8	0.6	0.1
	CDI	0.87	0.34	10	0	0	0	0	0	0
	ADI	0.87	0.35	10	0	0	0	0	0	0
	MCDI	0.95	0.71	7	3	0	0	0	0	0
	MADI	0.95	0.71	7	3	0	0	0	0	0
20	Random	0.81	0.34	1.1	3.4	5.4	5.5	3.4	1.1	0.2
	CDI	0.96	0.73	20	0	0	0	0	0	0
	ADI	0.96	0.72	20	0	0	0	0	0	0
	MCDI	0.98	0.85	7	13	0	0	0	0	0
	MADI	0.98	0.85	7	13	0	0	0	0	0
30	Random	0.87	0.48	1.8	5.0	8.7	8.2	4.7	1.5	0.2
	CDI	0.98	0.89	30	0	0	0	0	0	0
	ADI	0.98	0.90	30	0	0	0	0	0	0
	MCDI	0.99	0.95	14	16	0	0	0	0	0
	MADI	0.99	0.96	28.5	1.5	0	0	0	0	0

Note. DINA = deterministic input, noisy, “and” gate model; ACCR = attribute correct classification rate; PCCR = pattern correct classification rate; CDI = cognitive diagnostic index; ADI = attribute-level discrimination index; MCDI = modified cognitive diagnostic index; MADI = modified attribute-level discrimination index.

The performances of original indices and proposed indices were close when attributes have a linear hierarchical relationship, and test length was equal to 20 or 30. MADI was better than the other indices in all situations. When the attributes were hierarchically structured, the results of ADI were relatively unstable and unsatisfactory, whereas MCDI and MADI had the best performances in most situations.

As with the DINA model, the results based on the rRUM showed that the proposed indices had substantial improvement over the original indices when RTA is small or the attributes had a hierarchical relationship. Regardless of the situation, MCDI and MADI always had better or similar results compared with the other indices. It should be noted that test length should not be too short if good performances were to be expected of all indices.

The Usage of k -Attribute Items

In Tables 2 and 3, two attribute structures, H_0 and H_5 , and the DINA model with $s_j = g_j = .05$ were used to illustrate the relationship between the usage of k -attribute items and attribute classification accuracy across different methods of test construction. A k -attribute item is an item that requires k attributes to be answered correctly.

Table 2 shows that, when the attributes had a nonhierarchical structure, CDI and ADI exclusively used one-attribute items, whereas MCDI and MADI used one- and two-attribute items. This table clearly shows that tests with one-attribute items can lead to poorer performances particularly when I was smaller. Although one-attribute items individually can better discriminate between the examinee attribute patterns when the attributes have a nonhierarchical relationship, these items collectively cannot measure the attributes with sufficient number of times when RTA was small leading to relatively poorer attribute classification accuracy. It should be noted that Random, which used all but the seven-attribute items, provided the worst results. Table 3 shows that, when the attributes were linearly structured, CDI and ADI exclusively used four-

Table 3. Number of Times *k*-Attribute Items Are Used Under *H*₅ and DINA Model With *s*_{*j*} = *g*_{*j*} = .05.

I	Method	ACCR	PCCR	k						
				I	I	I	I	I	I	I
10	Random	0.96	0.75	1.4	1.4	1.5	1.7	1.2	1.4	1.5
	CDI	0.86	0.25	0	0	0	10	0	0	0
	ADI	0.79	0.25	10	0	0	0	0	0	0
	MCDI	0.98	0.89	1	1	2	2	2	1	1
	MADI	0.98	0.89	3	2	1	1	1	1	1
20	Random	0.99	0.93	2.9	2.9	2.8	2.7	3.0	2.8	2.9
	CDI	0.86	0.25	0	0	0	20	0	0	0
	ADI	0.79	0.25	20	0	0	0	0	0	0
	MCDI	0.99	0.97	2	3	3	4	3	3	2
	MADI	0.99	0.94	5	4	3	2	2	2	2
30	Random	1.00	0.98	4.3	4.4	4.1	4.5	4.3	4.4	4.0
	CDI	0.86	0.25	0	0	0	30	0	0	0
	ADI	0.79	0.25	30	0	0	0	0	0	0
	MCDI	1.00	0.99	3	4	5	6	5	4	3
	MADI	0.99	0.96	12	6	4	3	2	2	1

Note. DINA = deterministic input, noisy, “and” gate model; ACCR = attribute correct classification rate; PCCR = pattern correct classification rate; CDI = cognitive diagnostic index; ADI = attribute-level discrimination index; MCDI = modified cognitive diagnostic index; MADI = modified attribute-level discrimination index.

and one-attribute items, respectively, whereas Random, MCDI, and MADI used all the item types, albeit with different distributions. The table shows that using a single item type for this particular attribute structure resulted in dramatically poorer performances. In addition, the distribution of item type usage can affect the attribute classification—poor results were associated with a more even use of the item types.

Discussion

Cognitively diagnostic test combined tests with CDMs can better inform instruction and learning. Previously, CDI and ADI were proposed for the test construction in the CDM context. This article proposed new modified indices, namely, MCDI and MADI, that account for RTA and attribute hierarchies. Results indicate that MCDI and MADI are more effective than the original CDI and ADI in constructing tests based on the DINA model and rRUM, particularly when RTA is less than 3 or attributes have hierarchies. Specifically, MCDI and MADI can yield considerably higher correct attribute classification rates.

The current work focuses on MCDI- and MADI-based procedures for test construction. However, as noted earlier, other test construction procedures for cognitive diagnosis exist. In the future, it would be instructive to examine how the MCDI and MADI procedures compare with GA and BP. Of particular interest would be the computational feasibility of the GA when the errors to be minimized are defined at the attribute pattern level, not at the individual attribute level. Finkelman et al. (2010) noted that, in situations where GA is not feasible, the original CDI can be used to minimize the average attribute-level error rate. It is not clear whether this statement remains true when the attributes follow a hierarchical structure, or when CDI is replaced by ADI, MCDI, and MADI. Additional work is needed to better understand the properties of GA, BP, and these indices when they are used to construct tests that measure hierarchically structured attributes.

Although the modified indices allow each attribute to be measured adequate number of times, they still have poor performance if test length is very short (i.e., $RTA < 1$). In addition, for the proposed indices to work properly, attribute hierarchies have to be correctly identified. Follow-up studies are needed to investigate the ideal test lengths to achieve specific levels of attribute classification accuracy, and the consequences of using incorrect assumed attribute hierarchies.

For greater generality, future studies can also be carried out on constructing tests using MCDI or MADi with other CDMs such as G-DINA model (de la Torre, 2011), Higher-order-DINA model (de la Torre & Douglas, 2004), NIDA (Junker & Sijtsma, 2001), and GDM (von Davier, 2008). In addition, different strategies might be involved in solving a problem, and these strategies might require different attributes (de la Torre & Douglas, 2008). Therefore, test construction based on multiple-strategy CDM should also be considered.

Acknowledgments

The authors would like to thank the editor and the anonymous reviewers for insightful comments and valuable suggestions.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was supported by Ministry of Science and Technology, R.O.C. 102-2511-S-142-008-MY3.

References

- Chang, H., & Ying, Z. (1996). A global information approach to computerized adaptive testing. *Applied Psychological Measurement, 20*, 213-229.
- Cheng, Y. (2010). Improving cognitive diagnosis computerized adaptive testing by balancing attribute coverage: The modified maximum global discrimination index method. *Educational and Psychological Measurement, 70*, 902-913.
- de la Torre, J. (2011). The generalized DINA model framework. *Psychometrika, 76*, 179-199.
- de la Torre, J., & Douglas, J. (2004). Higher-order latent trait models for cognitive diagnosis. *Psychometrika, 69*, 333-353.
- de la Torre, J., & Douglas, J. (2008). Model evaluation and multiple strategies in cognitive diagnosis: An analysis of fraction subtraction data. *Psychometrika, 73*, 595-624.
- de la Torre, J., Hong, Y., & Deng, W. (2010). Factors affecting the item parameter estimation and classification accuracy of the DINA model. *Journal of Educational Measurement, 47*, 227-249.
- Finkelman, M., Kim, W., & Roussos, L. A. (2009). Automated test assembly for cognitive diagnosis models using a genetic algorithm. *Journal of Educational Measurement, 46*, 273-292.
- Finkelman, M., Kim, W., Roussos, L., & Verschoor, A. (2010). A binary programming approach to automated test assembly for cognitive diagnosis models. *Applied Psychological Measurement, 34*, 310-326.
- Haertel, E. H. (1989). Using restricted latent class models to map the skill structure of achievement items. *Journal of Educational Measurement, 26*, 333-352.
- Henson, R. A., & Douglas, J. (2005). Test construction for cognitive diagnostics. *Applied Psychological Measurement, 29*, 262-277.

- Henson, R. A., Roussos, L., Douglas, J., & He, X. (2008). Cognitive diagnostic attribute-level discrimination indices. *Applied Psychological Measurement, 32*, 275-288.
- Henson, R. A., Templin, J. L., & Willse, J. T. (2009). Defining a family of cognitive diagnosis models using log-linear models with latent variables. *Psychometrika, 74*, 191-210.
- Junker, B. W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement, 25*, 258-272.
- Kuo, B. C., Wu, H. M., & Shih, S. C. (2012, July). *The validity of Q-matrix design for DINA model: A practical perspective*. Paper presented at the 77th Annual Meeting of the Psychometric Society, Lincoln, NE.
- Leighton, J. P., Gierl, M. J., & Hunka, S. (2004). The attribute hierarchy model: An approach for integrating cognitive theory with assessment practice. *Journal of Educational Measurement, 41*, 205-236.
- Liu, J. C. (2013, July). *The self-learning Q-matrix—Theory and applications*. Paper presented at the 78th Annual Meeting of the Psychometric Society, Arnhem, The Netherlands.
- Maris, E. (1999). Estimating multiple classification latent class models. *Psychometrika, 64*, 87-212.
- Pai, H. S., Kuo, B. C., & Chen, C. H. (2012, July). *Cognitive Diagnostic Indices with pattern and attribute level adjustment*. Paper presented at the 77th Annual Meeting of the Psychometric Society, Lincoln, NE.
- Roussos, L. A., DiBello, L. V., Stout, W., Hartz, S. M., Henson, R. A., & Templin, J. L. (2007). The fusion model skills diagnosis system. In J. P. Leighton & M. J. Gierl (Eds.), *Cognitive diagnostic assessment for education: Theory and applications* (pp. 275-318). Cambridge, UK: Cambridge University Press.
- Tatsuoka, K. K. (1983). Rule space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement, 20*, 345-354.
- von Davier, M. (2008). A general diagnosis model applied to language testing data. *British Journal of Mathematical and Statistical Psychology, 61*, 287-307.