PSYCHOMETRIKA—VOL. 74, NO. 4, 619–632
DECEMBER 2009
DOI: 10.1007/S11336-009-9123-2

WHEN COGNITIVE DIAGNOSIS MEETS COMPUTERIZED ADAPTIVE TESTING: CD-CAT

YING CHENG

UNIVERSITY OF NOTRE DAME

Computerized adaptive testing (CAT) is a mode of testing which enables more efficient and accurate recovery of one or more latent traits. Traditionally, CAT is built upon Item Response Theory (IRT) models that assume unidimensionality. However, the problem of how to build CAT upon latent class models (LCM) has not been investigated until recently, when Tatsuoka (J. R. Stat. Soc., Ser. C, Appl. Stat. 51:337–350, 2002) and Tatsuoka and Ferguson (J. R. Stat., Ser. B 65:143–157, 2003) established a general theorem on the asymptotically optimal sequential selection of experiments to classify finite, partially ordered sets. Xu, Chang, and Douglas (Paper presented at the annual meeting of National Council on Measurement in Education, Montreal, Canada, 2003) then tested two heuristics in a simulation study based on Tatsuoka's theoretical work in the context of computerized adaptive testing. One of the heuristics was developed based on Kullback–Leibler information, and the other based on Shannon entropy. In this paper, we showcase the application of the optimal sequential selection methodology in item selection of CAT that is built upon cognitive diagnostic models. Two new heuristics are proposed, and are compared against the randomized item selection method and the two heuristics investigated in Xu et al. (Paper presented at the annual meeting of National Council on Measurement in Education, Montreal, Canada, 2003). Finally, we show the connection between the Kullback–Leibler-information-based approaches and the Shannon-entropy-based approach, as well as the connection between algorithms built upon LCM and those built upon IRT models.

Key words: optimal sequential selection, latent class model, computerized adaptive testing, cognitive diagnosis, item response theory.

1. Introduction

Cognitive diagnosis in psychological and educational measurement features a combination of model-based measurement and formative assessment (Embretson, 2001). On the one hand, it is completely model-based. In the past two decades, various models have been proposed for cognitive diagnosis, such as the rule space model (Tatsuoka, 1983), the binary skills model (Haertel, 1984; Haertel & Wiley, 1993), the Bayesian inference network model (Mislevy, Almond, Yan, & Steinberg, 1999), the "Deterministic Input, Noisy 'And' Gate" (DINA) model (Haertel, 1989; Junker & Sijtsma, 2001), the "Deterministic Input, Noisy 'Or' Gate" (DINO) model (Templin & Henson, 2006), and the Fusion model (Hartz, 2002; Hartz, Roussos, & Stout, 2002), just to name a few. On the other hand, cognitive diagnosis attempts to offer feedback to each student regarding his or her strengths and weaknesses. Typically, instead of receiving a summative score, students receive a profile, specifying which concepts and skills they have mastered, and which ones they have not.

Another line of measurement research that has received increasing attention is computerized adaptive testing. Computerized adaptive testing is a mode of testing which tries to find the "best fitting" items for each individual. As a result, it can improve the efficiency and accuracy of testing over the traditional paper-and-pencil mode. An interesting research question that naturally arises

The author would like to thank the editors, anonymous reviewers, and Drs. Hua-Hua Chang and Jeff Douglas for their constructive suggestions.

Requests for reprints should be sent to Ying Cheng, 118 Haggar Hall, Notre Dame, IN, 46556, USA. E-mail: ycheng4@nd.edu

then is how can computerized adaptive testing be used to help perform cognitive diagnosis more efficiently, i.e., how to develop "cognitive diagnostic computerized adaptive testing" or CD-CAT. The purpose is to provide individualized diagnostic feedback using the tailored mode of testing.

From a modeling perspective, many of the cognitive diagnostic models are partially-ordered latent class models (von Davier, 2005). Most of the current computerized adaptive testing programs purport to locate examinees on a latent continuum, and thus utilize latent trait models. By contrast, cognitive diagnostic computerized adaptive testing (CD-CAT) aims at classifying examinees based on their latent states, and thus employs latent class models. It therefore poses challenges to researchers and practitioners to deliver a cognitive diagnostic test using the computerized adaptive platform.

Tatsuoka (2002) and Tatsuoka and Ferguson (2003) first addressed this problem by establishing a general theorem on the asymptotically optimal sequential selection of experiments (e.g., in the context of item response theory (IRT), items) to classify finite, partially ordered sets. Xu, Chang, and Douglas (2003) then examined specifically two item selection heuristics for cognitive diagnostic computerized adaptive testing (CD-CAT), one based on the Kullback–Leibler information and hereon called the "KL" algorithm, and the other based on the Shannon entropy and hereon called the "SHE" algorithm.

In this paper, we readdress the problem of developing item selection algorithms for CD-CAT by proposing two new algorithms following the theoretical work of Tatsuoka (2002) and Tatsuoka and Ferguson (2003): (a) Posterior-weighted KL information or the PWKL method. The former updates the posterior distribution of latent classes every time another item is answered by the test taker, and selects the next item accordingly. The latter differs from the form in that it takes into account the distance between latent classes. Further, we point out the connection between the KL-based approaches and the entropy-based approach, as well as the connection between the CD-CAT algorithms and those based on latent trait IRT models. Two simulation studies, one using simulated item parameters, the other using parameter estimates from real data, show that the PWKL and HKL algorithms outperform the KL and SHE algorithms uniformly. In addition, the comparison of the performance of the PWKL and the HKL algorithms against the randomized item selection highlights the advantages of applying optimal sequential experiment selection in the context of CD-CAT.

1.1. Computerized Adaptive Testing (CAT)

Computerized adaptive testing (CAT) is a desirable mode for testing because it can tailor items to the latent trait of an examinee. The key to a CAT program is the item selection algorithm, which picks the optimal items sequentially. One of the most popular item selection methods is the maximum Fisher information (MFI) method (Lord, 1980; Thissen & Mislevy, 2000).

The Fisher information measures the amount of information that an observable random variable X carries about an unknown parameter θ, upon which the likelihood function of θ depends. Mathematically, it can be expressed as

$$I(\theta) = E\left\{\left[\frac{\partial}{\partial\theta}\ln f(X;\theta)\right]^2 \middle| \theta\right\}. \tag{1}$$

Here, $f(X;\theta)$ represents the likelihood function computed based on the item response functions (e.g., the 1PL, 2PL, or 3PL model; see Hambleton & Swaminathan, 1985) and the θ is the latent trait of interest. In the following discussion, we will refer to θ as "ability," which is used here as a generic term.

Suppose t items have been administered. The MFI algorithm selects the next item which has the highest Fisher information at $\hat{\theta}_i^{(t)}$, i.e., the latest ability estimate. In other words, the $(t+1)$th item for the ith examinee is

$$\arg\max_{h}\left\{I_h\left(\hat{\theta}_i^{(t)}\right) : h \in R^{(t)}\right\} \tag{2}$$

where $R^{(t)}$ represents the set of available items at stage t. Because the variance of the ability estimate, i.e., $\hat{\theta}$, is inversely related to the Fisher information, by selecting items that maximize Fisher information at the interim ability estimates, the MFI algorithm can locate the true θ quickly.

Though MFI remains very popular in CAT research and operations, algorithms based on the Kullback–Leibler information and Shannon entropy are receiving more and more attention. An important reason is that the ability estimate $\hat{\theta}^{(t)}$ may not be in the vicinity of the true θ, when t is small. Consequently, the MFI algorithm is prone to capitalization on chance. Therefore, in the early stage of CAT, a measure that is more global might be helpful. An example is Chang and Ying (1996) global information method developed on the basis of the Kullback–Leibler information.

On the other hand, when the underlying latent structure involves discrete latent classes, or when nonparametric IRT models are used, the Fisher information does not apply. But neither Kullback–Leibler information or Shannon entropy has this limitation. For example, they were used in CAT built upon nonparametric IRT models (Xu & Douglas, 2006) and cognitive diagnosis models (McGlohen & Chang, 2008; Xu et al., 2003).

Next, we will review in detail the two algorithms examined in Xu et al. (2003), which demonstrated the optimal sequential experiment selection methodology (Tatsuoka, 2002; Tatsuoka & Ferguson, 2003) in the context of computerized adaptive testing.

1.1.1. Applications of the Kullback–Leibler Information and Shannon Entropy in CD-CAT The Kullback–Leibler (or KL) information is a measure of divergence or "distance" between two probability distributions (Cover & Thomas, 1991):

$$D[f,g] = E_f\left[\log\frac{f(\mathbf{x})}{g(\mathbf{x})}\right]. \tag{3}$$

Here, $f(\mathbf{x})$ and $g(\mathbf{x})$ are two probability distributions. Usually, $f(\mathbf{x})$ represents the "true" distribution of data, observations, or a precise theoretical distribution. The measure $g(\mathbf{x})$ typically represents a theory, a model, a description, or an approximation of $f(\mathbf{x})$. Note that the KL information is not strictly a distance measure because it is not symmetric, i.e., $D[f,g] \neq D[g,f]$. It is sometimes referred to as the KL "distance" because it reflects how divergent, or how far apart, two probability distributions are. In other words, the larger $D[f,g]$ is the easier it is to statistically tell apart the two probability distributions (Henson & Douglas, 2005). Nevertheless, in the following discussion, the terms "KL information," "KL divergence" and "KL distance" will be used interchangeably.

1.1.2. The KL Algorithm Based on Kullback–Leibler Information Suppose t items are selected, and at this stage the available items in the pool form a set $R^{(t)}$. Consider item h in $R^{(t)}$. In cognitive diagnosis, we are interested in the conditional distribution of person i's item responses U_{ih} given his or her latent state, or cognitive profile, $\boldsymbol{\alpha}_i$. Following the notation of McGlohen and Chang (2008), $\boldsymbol{\alpha_i} = (\alpha_{i1}, \alpha_{i2}, \ldots, \alpha_{ik}, \ldots, \alpha_{iK})'$, where $\alpha_{ik} = 1$ indicates that the ith examinee masters the kth attribute and $\alpha_{ik} = 0$ otherwise. An attribute is a task, cognitive process, or skill involved in answering an item.

Since the true state is unknown, a global measure of discrimination can be constructed on the basis of the KL distance between the distribution of U_{ih} given the current estimate of person i's latent cognitive state (i.e., $f(U_{ih} \mid \hat{\boldsymbol{\alpha}}_i^{(t)})$) and the distribution of U_{ih} given other states.

The KL distance between $f(U_{ih}\,|\,\hat{\boldsymbol{\alpha}}_i^{(t)})$ and the conditional distribution of U_{ih} given another latent state $\boldsymbol{\alpha_c}$, i.e., $f(U_{ih}\,|\,\boldsymbol{\alpha_c})$, can be computed as follows:

$$D_h\big(\hat{\boldsymbol{\alpha}}_i^{(t)} \parallel \boldsymbol{\alpha_c}\big) = \sum_{q=0}^{1} \log\left(\frac{P(U_{ih}=q\,|\,\hat{\boldsymbol{\alpha}}_i^{(t)})}{P(U_{ih}=q\,|\,\boldsymbol{\alpha_c})}\right) P\big(U_{ih}=q\,|\,\hat{\boldsymbol{\alpha}}_i^{(t)}\big). \tag{4}$$

Xu et al. (2003) proposed using the straight sum of the KL distances between $f(U_{ih}\,|\,\hat{\boldsymbol{\alpha}}_i^{(t)})$ and all the $f(U_{ih}\,|\,\boldsymbol{\alpha}_c)$s, $c = 1, 2, \ldots, 2^K$ (when there are K attributes, there are 2^K possible latent cognitive states):

$$KL_h\big(\hat{\boldsymbol{\alpha}}_i^{(t)}\big) = \sum_{c=1}^{2^K} D_h\big(\hat{\boldsymbol{\alpha}}_i^{(t)} \parallel \boldsymbol{\alpha}_c\big). \tag{5}$$

Then the $(t+1)$th item for the ith examinee is the item in $R^{(t)}$ that maximizes $KL_h(\hat{\boldsymbol{\alpha}}_i^{(t)})$. This is referred to as the KL algorithm. The items selected using this algorithm are the most powerful ones on average in distinguishing the current latent class estimate from all other possible latent classes.

1.2. The SHE Algorithm Based on Shannon Entropy

Xu et al. (2003) examined an algorithm which was first discussed in Tatsuoka (2002), which tries to minimize the expected Shannon entropy of the posterior distribution of the $\hat{\boldsymbol{\alpha}}$. We will refer to this algorithm as the "SHE" algorithm.

The Shannon entropy is a measure of uncertainty associated with a probability distribution, first proposed by Shannon (1948). Shannon entropy of a discrete probability distribution $\mathbf{P}$ is defined as

$$H(\mathbf{P}) = -\sum_{i=1}^{n} p_i \log_b p_i, \tag{6}$$

where $\mathbf{P} = (p_1, p_2, \ldots, p_n)$ and $p_i = \text{Prob}(X = x_i)$. The $H(\mathbf{P})$ is a nonnegative, concave function, which reaches 0 when $\mathbf{P}$ is most concentrated, i.e., when there is a $p_i = 1$, and every $p_j = 0$ when $j \neq i$. It is maximized when all the outcomes are equally likely, i.e., when $p_i = \frac{1}{n}$ for $\forall i = 1, 2, \ldots, n$.

Suppose the prior is chosen as follows:

$$\Pr(\boldsymbol{\alpha}_c) = \pi_{0c}, \tag{7}$$

subject to $\sum_{c=1}^{2^K} \pi_{0c} = 1$ and $\pi_{0c} \geq 0$ for $\forall c = 1, 2, \ldots, 2^K$. The posterior distribution after t responses are observed can then be written as

$$\pi_{i,t}(\boldsymbol{\alpha}_c) \propto \pi_{0c} \cdot L\big(\mathbf{u}_i^{(t)}\,|\,\boldsymbol{\alpha}_c\big) \tag{8}$$

where $\mathbf{u}_i^{(t)}$ is the realization of the random response vector for person i at stage t, and $L(\mathbf{u}_i^{(t)}\,|\,\boldsymbol{\alpha}_c)$ is the likelihood function, and it is simply the product of each item response function when local independence is assumed.

The Shannon entropy of the posterior distribution $\pi_{i,t}$ can then be expressed as

$$H(\pi_{i,t}) = -\sum_{c=1}^{2^K} \pi_{i,t}(\boldsymbol{\alpha}_c) \log_b\big(\pi_{i,t}(\boldsymbol{\alpha}_c)\big). \tag{9}$$

Therefore, considering item h in $R^{(t)}$, we obtain the *expected* Shannon entropy as follows:

$$SHE_h(\pi_{i,t+1}) = \sum_{q=0}^{1} H\big(\pi_{i,t+1} \mid \mathbf{u}_i^{(t)}, U_{(i,h)} = q\big) \cdot P\big(U_{(i,h)} = q \mid \mathbf{u}_i^{(t)}\big). \tag{10}$$

Finally, the $(t+1)$th item to be selected for the ith examinee is the one in $R^{(t)}$ that minimizes $SHE_h(\pi_{i,t+1})$. Items picked for a given person by the SHE algorithm are the ones that lead to the minimum uncertainty of the posterior distribution of his or her latent class estimate.

The base of the logarithm, b, actually does not influence the item selection, because it only changes the unit in which the Shannon entropy is measured (Cover & Thomas, 1991). In this study, we use natural logarithm.

1.3. New Algorithms

Xu et al. (2003) showed that the SHE algorithm works better than the KL algorithm across all conditions. However, there is an implicit assumption made in the KL algorithm, that is all the latent states $\boldsymbol{\alpha}_c$ $(c = 1, 2, \ldots, 2^K)$ are equally likely to be the true state for each examinee at each step of item selection. This assumption is unnecessary and might cause inefficiency. In the following, we offer several new algorithms in which the condition is relaxed.

1.3.1. Posterior-Weighted KL (PWKL) Algorithm We might be able to infer the distribution of the latent states in the current sample by analyzing old samples. For example, it is often reasonable to assume that this year's test-taker population does not differ much from last year's. Therefore, we can impose informative priors on the latent states and obtain posterior distributions at each step as shown above in (7) and (8).

The posterior-weighted KL (PWKL) index can thus be defined as follows:

$$\begin{aligned} PWKL_h\big(\hat{\boldsymbol{\alpha}}_i^{(t)}\big) &= \sum_{c=1}^{2^K} D_h\big(\hat{\boldsymbol{\alpha}}_i^{(t)} \parallel \boldsymbol{\alpha}_c\big)\pi_{i,t}(\boldsymbol{\alpha}_c) \\ &= \sum_{c=1}^{2^K} D_h\big(\hat{\boldsymbol{\alpha}}_i^{(t)} \parallel \boldsymbol{\alpha}_c\big)\pi_{i,0}(\boldsymbol{\alpha}_c) L\big(\mathbf{u}_i^{(t)} \mid \boldsymbol{\alpha}_c\big). \end{aligned} \tag{11}$$

Assuming local independence, we can write the likelihood function $L(\mathbf{u}_i^{(t)}; \boldsymbol{\alpha}_c)$ as

$$L\big(\mathbf{u}_i^{(t)}; \boldsymbol{\alpha}_c\big) = \prod_{k=1}^{t} \big[P_i(\boldsymbol{\alpha}_c)\big]^{u_{ik}} \big[1 - P_i(\boldsymbol{\alpha}_c)\big]^{1-u_{ik}}, \tag{12}$$

where $P_i(\boldsymbol{\alpha}_c)$ is the item response function defined by a cognitive diagnosis model.

The $(t+1)$th item to be selected for the ith examinee is therefore the one in $R^{(t)}$ that can maximize $PWKL_h(\hat{\boldsymbol{\alpha}}_i^{(t)})$.

If the prior is discrete uniform, then the PWKL index is equivalent to one that weights the KL distance by the likelihood of each latent state:

$$LWKL_h\big(\hat{\boldsymbol{\alpha}}_i^{(t)}\big) = \sum_{c=1}^{2^K} D_h\big(\hat{\boldsymbol{\alpha}}_i^{(t)} \parallel \boldsymbol{\alpha}_c\big) L\big(\mathbf{u}_i^{(t)} \mid \boldsymbol{\alpha}_c\big). \tag{13}$$

This is referred to as the LWKL algorithm, standing for likelihood-weighted KL algorithm.

1.3.2. Hybrid KL (HKL) Index Incorporating Distance between Latent States Henson and Douglas (2005) noted that if an item "discriminates well between attribute patterns which are similar, it will discriminate well between those that are dissimilar." Therefore, we can define another index which assigns more weight to those latent states that are closer to the current estimate.

A common measure of distance is Euclidean distance:

$$d(\boldsymbol{\alpha}_c, \boldsymbol{\alpha}_{c'}) = \sqrt{\sum_{k=1}^{K} (\alpha_{ck} - \alpha_{c'k})^2}. \tag{14}$$

So, every element in the PWKL index can be further weighted by the inverse of the distance between the $\hat{\boldsymbol{\alpha}}_i^{(t)}$ and any other possible latent state. By doing that we get the hybrid index (a.k.a. the HKL):

$$HKL_h\bigl(\hat{\boldsymbol{\alpha}}_i^{(t)}\bigr) = \sum_{c=1}^{2^K} D_h\bigl(\hat{\boldsymbol{\alpha}}_i^{(t)} \parallel \boldsymbol{\alpha}_c\bigr)\pi_{i,t}(\boldsymbol{\alpha}_c)\frac{1}{d(\boldsymbol{\alpha}_c, \hat{\boldsymbol{\alpha}}_i^{(t)})}. \tag{15}$$

The difference between the HKL and the PWKL index is that other things being equal, the former favors items that better discriminates $\hat{\boldsymbol{\alpha}}$ and those latent states that are close to it.

1.4. The Relations Among Algorithms

1.4.1. The SHE Algorithm and the KL-Based Indices It can be shown that

$$H(\mathbf{P}) = \log_b n - KL(\mathbf{P} \parallel \mathbf{V}) \tag{16}$$

where $\mathbf{V}$ represents the discrete uniform distribution, i.e., every outcome has equal probability. Therefore, minimizing the expected Shannon entropy of the predicted posterior is equivalent to maximizing the expected KL distance between the predicted posterior and the discrete uniform distribution. As a matter of fact, due to its connection to the Shannon entropy, the Kullback–Leibler information is also known as the "relative entropy."

1.4.2. The SHE Algorithm and the MEPV Algorithm The SHE algorithm is similar to an item-selection algorithm proposed by van der Linden based on IRT models (1998), known as the minimum expected posterior variance (MEPV) method, which selects the $(t+1)$th item as follows:

$$\min_{h \in R^t} \{EPV_{ih}\}, \tag{17}$$

where the expected posterior variance (EPV) can be written as

$$EPV_{ih} = \left\{\sum_{x=0}^{1} \Pr\bigl(X_h = x \mid \mathbf{u}_i^{(t)}\bigr)\operatorname{Var}\bigl(\mathbf{u}_i^{(t)}, X_h = x\bigr)\right\}, \tag{18}$$

where $\mathbf{u}_i^{(t)}$ is the posterior distribution of Θ, the unidimensional latent trait.

Comparing (10) and (18), one can easily see that the SHE algorithm essentially follows the same logic as the MEPV algorithm. The only difference is that the SHE algorithm uses the Shannon entropy of the posterior distribution, whereas MEPV uses the posterior variance. However, as pointed out by one reviewer, the two methods are distinct in that the Shannon entropy is restricted to discrete random variables. For instance, it is defined here in CD-CAT on discrete latent classes. Variance, on the other hand, cannot be defined on latent classes, but works well with the latent trait model which assumes a continuous random variable. The continuous entropy, also known as differential entropy, can be considered an extension of the Shannon entropy to the domain of real numbers:

$$h[f] = -\int_{-\infty}^{\infty} f(x) \log f(x)\, d_x. \tag{19}$$

Nevertheless, unlike Shannon entropy, the differential entropy is not a good measure of uncertainty or information (see Cover & Thomas, 1991), and in the continuous realm, the relative entropy of a distribution, i.e., the Kullback–Leibler divergence is used more extensively. Therefore, even though carrying the same logic, the MEPV and SHE algorithms cannot be used interchangeably.

1.4.3. The KL, PWKL and HKL Algorithms and the Global Information Approach Chang and Ying (1996) proposed a global information measure to capture the discriminating power of an item, i.e., how powerful an item is in telling apart two conditional distributions: the distribution of item response given $\hat{\theta}_i^{(t)}$, i.e., $f(U_{ih} \mid \hat{\theta}_i^{(t)})$, and the distribution of item response given neighboring points of $\hat{\theta}_i^{(t)}$:

$$G_h\big(\hat{\theta}_i^{(t)}\big) = \int_{\hat{\theta}_i^{(t)}-\delta}^{\hat{\theta}_i^{(t)}+\delta} D_h\big(\theta \parallel \hat{\theta}_i^{(t)}\big)\, d_\theta, \tag{20}$$

where δ is inversely related to $\sqrt{t}$.

Chang and Ying (1996) also proposed a global Bayesian information index which takes into consideration the posterior distribution of θ:

$$GB_h\big(\hat{\theta}_i^{(t)}\big) = \int_{-\infty}^{+\infty} D_h\big(\theta \parallel \hat{\theta}_i^{(t)}\big) \cdot p\big(\theta \mid \mathbf{U}_i^{(t)} = \mathbf{u}_i^{(t)}\big)\, d\theta, \tag{21}$$

where $p(\theta \mid \mathbf{u}_i^{(t)})$ is the posterior distribution of θ given all the available responses from person i, i.e., $\mathbf{u}_i^{(t)}$.

The KL algorithm examined in Xu et al. (2003) is analogous to the global information approach when $\delta = \infty$. The PWKL algorithm, on the other hand, is analogous to the global Bayesian information approach when $\delta = \infty$. Finally, the HKL algorithm is analogous to the global information approaches in that the latent states that are further away the current estimate are given less weight—the global information approaches are simply assigning zero weights to those θ that are not in the vicinity of the current θ estimate.

2. Data and Simulation Design

Before getting into details of the simulation design, we would like to introduce the cognitive diagnostic model used in this study.

2.1. Cognitive Diagnostic Model

The cognitive diagnostic model used here is the “Deterministic Input; Noisy ‘And’ Gate” (DINA) model, which was originally proposed in Haertel (1989), and later discussed extensively in Junker and Sijtsma (2001). It relates item responses to a set of latent attributes. The purpose of cognitive diagnosis is to identify which attributes are mastered by a test taker and which ones are not. As mentioned before, for each examinee, the mastery status translates into a vector:

$\boldsymbol{\alpha_i} = (\alpha_{i1}, \alpha_{i2}, \ldots, \alpha_{ik}, \ldots, \alpha_{iK})'$, where $\alpha_{ik} = 1$ indicates that the ith examinee masters the kth attribute, and $\alpha_{ik} = 0$ otherwise. Items are related to attributes by an incidence matrix, denoted by **Q** (Tatsuoka, 1995).

Q is a $J \times K$ matrix: $q_{jk} = 1$ if getting a correct response to item j requires the mastery of attribute k when there is no guessing, and $q_{jk} = 0$ otherwise. **Q** matrix is usually identified by content experts and psychometricians. An example of **Q** matrix is given as follows:

$$\begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix}. \tag{22}$$

This **Q** matrix indicates that the first item requires the second attribute, the second item requires the first and the last attribute, and the third item requires the first two attributes.

Let $\mathbf{U}_i$ denote a vector of dichotomous item response for the ith examinee: $\mathbf{U}_i = (U_{i1}, U_{i2}, \ldots, U_{iJ})$. His or her mastery status $\boldsymbol{\alpha}$ accounts for the pattern of $\mathbf{U}_i$ mostly. In addition, the DINA model also allows for "slipping" and "guessing." Here, slips and guesses are modeled at the item level. Parameter s_j represents the probability of slipping on the jth item when an examinee has mastered all the attributes it requires. Parameter g_j denotes the probability of correctly answering the jth item when an examinee does not master all the required attributes.

Let η_{ij} denote whether the ith examinee possesses all the required attributes of item j. $\eta_{ij} = 1$ indicates all the required attributes are mastered, and $\eta_{ij} = 0$ otherwise. It can be calculated as follows:

$$\eta_{ij} = \prod_{k=1}^{K} \alpha_{ik}^{q_{jk}}. \tag{23}$$

The item response function therefore can be written as

$$P(U_{ij} = 1 \,|\, \boldsymbol{\alpha}) = (1 - s_j)^{\eta_{ij}} g_j^{1-\eta_{ij}}. \tag{24}$$

With the assumptions of local independence and independence among all the subjects, the joint likelihood function of the DINA model can be written as

$$Like(s, g; \boldsymbol{\alpha}) = \prod_{i=1}^{N} \prod_{j=1}^{J} \left[(1 - s_j)^{y_{ij}} s_j^{1-y_{ij}}\right]^{\eta_{ij}} \left[g_j^{y_{ij}} (1 - g_j)^{1-y_{ij}}\right]^{1-\eta_{ij}}. \tag{25}$$

We would like to emphasize here that the cognitive model itself is not the focus of this study. Our concern is to develop an item selection algorithm which works well for cognitive diagnostic purpose and the algorithm should be applicable to various cognitive diagnostic models. The DINA model is chosen for this study due to its computational ease. For a real-time system such as CAT, computational efficiency is a very desirable property.

2.2. *Details of Simulation*

Two simulation studies are carried out, one using a simulated item bank, and the other based on items calibrated from real data. The following are the details of the first simulation study.

1. Test takers. The test takers are generated assuming that every examinee has a 50% chance of mastering each attribute. In other words, for a 6-attribute test, the 64 cognitive states are equally likely in the population. In total, 2,000 $\boldsymbol{\alpha}$s are generated to mirror the sample size of the real data in the second simulation study. These are the true $\boldsymbol{\alpha}$s used to generate

item responses. It is expected that the distribution of latent classes in a real test-taker population will not follow the discrete uniform distribution as simulated here, but if that is the case, we can employ informative prior in item selection.

2. Item bank generation, including the generation of **Q**-matrix, slipping and guessing parameters. In total, 300 items are generated. A rule of thumb of item banking is that the pool needs to have at least 12 times as many items as the test length (Stocking, 1994), and some researchers recommended even larger ratios (Chang & Zhang, 2002). Therefore, we are simulating a 300-item bank to make sure that we have a large enough pool. In addition, an item pool of several hundred items is not atypical in applications. The **Q**-matrix used in this study is generated item by item and attribute by attribute. Each item has 20% chance of measuring each attribute. This mechanism is employed to make sure that every attribute is adequately and equally represented in the item pool. The expected number of attributes measured per item is 1.2, and the expected number of items per attribute in the pool is 60. The guessing and slipping level are set at 5%. These parameter values are chosen to reflect a minimal level of noise.
3. Estimation. The initial $\hat{\boldsymbol{\alpha}}$, i.e., $\hat{\boldsymbol{\alpha}}^{(\mathbf{0})}$, is randomly generated, with each attribute independently generated with equal probability of being 0 or 1. Then the maximum likelihood estimation (MLE) method is used to update $\hat{\boldsymbol{\alpha}}^{(t)}$s. The final estimates are $\hat{\boldsymbol{\alpha}}^{(T)}$s, where T is the test length. In this study, $T = 12$.
4. Item selection rule. The original KL index, SHE index and the new indices are used to select items. Since the test takers are generated in such a way that all latent states are equally likely, we adopt a flat prior, i.e., discrete uniform. Consequently, the likelihood-weighted KL (LWKL) index and the posterior-weighted KL (PWKL) index are the same in this simulation. Randomized item selection is also included in the simulation study to provide the baseline. In summary, five item selection methods are considered: KL, SHE, PWKL (in this case identical to LWKL), HKL, and randomized (RANDOM) selection.
5. The dependent variables are the recovery rates of the cognitive profiles, including recovery of the whole pattern and of each attribute. This can be accomplished by comparing each $\boldsymbol{\alpha}$ with $\hat{\boldsymbol{\alpha}}^{(T)}$s. For instance, if the true $\boldsymbol{\alpha}$ is [0 1 1 1 0 1] and the final $\hat{\boldsymbol{\alpha}}^{T}$ is also [0 1 1 1 0 1], we say the whole pattern is recovered. If the $\hat{\boldsymbol{\alpha}}^{T}$ turns out to be [1 0 0 0 1 1], only the last attribute is recovered. There are 2,000 pairs of $\boldsymbol{\alpha}$ and $\hat{\boldsymbol{\alpha}}^{(T)}$, so the recovery rates for the entire pattern and each attribute can be calculated by comparing every single pair of $\boldsymbol{\alpha}$ and $\hat{\boldsymbol{\alpha}}^{(T)}$.

The second simulation study differ from the first one in the following aspects:

1. Test takers and item bank. A random sample of 2,000 third-grade students are selected on a state reading assessment in spring 2002. The reading test is a criterion-referenced test closely aligned with the state mandated curriculum. It consists of 36 multiple choice items. Each item is scored dichotomously. The same dataset was also analyzed in McGlohen (2004), where two Q-matrices were developed for this test. In our study, we adopted the complex Q-matrix, meaning each item may measure more than one attribute (see Table 1). For the substantive meaning of each attribute, please see McGlohen and Chang (2008).

   The 2,000 × 36 item response matrix serves as the calibration sample. The program *CDM* (Templin, 2006) is used to calibrate test takers' cognitive profiles and item guessing and slipping parameters. The profiles and item parameters then serve as true profiles and true parameters in our second simulation study. The mean and standard deviation of the guessing and slipping parameter estimates are provided in Table 2. Obviously, the guessing values are quite high. The implication of having such high guessing values will be discussed in the next section. Table 3 summarizes the proportion of items in the pool

TABLE 1.
The Q-matrix (simulation 2).

Item	Attributes
1	0 1 0 0 0 0
2	0 0 0 1 1 0
3	1 0 0 0 0 0
4	0 1 0 1 0 0
5	0 0 1 0 1 0
6	0 1 0 0 0 0
7	0 0 0 0 0 1
8	0 0 0 1 1 0
9	1 0 0 0 0 0
10	0 1 0 1 0 0
11	0 1 0 1 0 0
12	0 0 0 0 1 0
13	0 0 0 0 0 1
14	0 0 0 0 0 1
15	1 0 0 0 0 0
16	0 1 0 0 0 0
17	0 1 0 1 0 0
18	0 0 0 1 1 0
19	0 1 0 0 0 0
20	0 1 0 0 0 0
21	0 0 0 0 1 0
22	0 0 0 0 0 1
23	1 0 0 0 0 0
24	0 1 0 1 0 0
25	0 0 1 0 1 0
26	0 0 1 0 1 0
27	1 0 0 0 0 0
28	0 1 0 0 0 0
29	1 0 0 0 0 0
30	0 0 1 0 1 0
31	0 1 0 0 0 0
32	0 0 0 0 1 0
33	1 0 0 0 0 0
34	0 0 1 0 1 0
35	0 1 0 0 1 0
36	0 0 1 0 1 0

TABLE 2.
Descriptive statistics of the thirty-six items (simulation 2).

	Mean	SD
Guessing (g)	0.58	0.14
Slipping s	0.07	0.06

TABLE 3.
Descriptive statistics of the item pool and the examinees (%) (simulation 2).

	Attr. 1	Attr. 2	Attr. 3	Attr. 4	Attr. 5	Attr. 6
Proportion of items	19	36	17	22	36	11
Proportion of examinees	78	81	95	95	81	78

measuring each attribute, as well as the proportion of examinees mastering each attribute in the sample.

2. Test length. Test length is doubled in the second simulation study, i.e., $T = 24$.

Other aspects are the same for both simulation studies, including the way to update profile estimates, item selection rules, the cognitive diagnosis model being used, and how the results are analyzed.

3. Results and Educational Implications

Table 4 compares the five item selection methods using the simulated item bank. The first three rows replicate the findings in Xu et al. (2003), though the two studies use a different cognitive diagnostic models.[1] The SHE algorithm does a fairly good job and outperforms the KL algorithm and the randomization item selection method.

What is of more concern to us are the performances of the new indices, as shown in the last two rows. Clearly, the likelihood or posterior-based KL indices improve the recovery rates substantially. Not surprisingly, they outperform the original KL algorithm and the randomization algorithm. The improvement is uniform, meaning that the recovery rates of every attribute and of the whole pattern are all better. The PWKL and HKL algorithms also outperform the SHE algorithm. The improvement of the PWKL algorithm over the SHE algorithm is uniform.

Between the PWKL and HKL method, it is hard to pinpoint which one is the winner. The PWKL method leads to higher recovery rates on attributes 5 and 6, whereas the HKL method performs better with respect to attribute 2 and the whole pattern. But, in general, the difference between their performances is very small.

Following Xu et al. (2003), all the whole-pattern recovery rates are divided by the whole-pattern recovery rate of the randomization method. The results are summarized in Table 5. Clearly, the SHE, PWKL, and HKL algorithms are more than twice as efficient as the randomization

TABLE 4.
Recovery rates of the examinees' cognitive profiles—simulation 1.

Methods	Attr. 1	Attr. 2	Attr. 3	Attr. 4	Attr. 5	Attr. 6	Entire pattern
Random	0.85	0.82	0.86	0.85	0.84	0.83	0.37
KL	0.96	0.66	0.90	0.79	0.95	0.95	0.42
SHE	0.99	0.95	0.96	0.95	0.96	0.96	0.81
PWKL	0.99	0.98	0.99	0.99	1.00	0.99	0.93
HKL	0.99	0.99	0.99	0.99	0.99	0.98	0.94

TABLE 5.
Relative efficiency of the item selection algorithms: simulation 1.

Methods	Relative efficiency
Random	1.00
KL	1.13
SHE	2.22
PWKL	2.54
HKL	2.56

[1] Xu et al. used a simplified version of the Fusion model whereas this study is based on the DINA model.

TABLE 6.
Recovery rates of the examinees' cognitive profiles—simulation 2.

Methods	Attr. 1	Attr. 2	Attr. 3	Attr. 4	Attr. 5	Attr. 6	Entire pattern
Random	0.84	0.91	0.70	0.77	0.89	0.80	0.38
KL	0.89	0.93	0.62	0.72	0.92	0.86	0.39
SHE	0.81	0.93	0.74	0.81	0.89	0.84	0.47
PWKL	0.88	0.94	0.71	0.81	0.93	0.86	0.48
HKL	0.88	0.94	0.71	0.82	0.93	0.86	0.48

TABLE 7.
Relative efficiency of the item selection algorithms: simulation 2.

Methods	Relative efficiency
Random	1.00
KL	1.02
SHE	1.23
PWKL	1.27
HKL	1.26

method. Actually, the PWKL and HKL algorithms are about 2.5 times as efficient as the baseline, and are about 1.14 times (i.e., 2.5/2.2) as efficient as the SHE algorithm.

Table 6 shows the recovery rates of each attribute and the entire pattern when we use items calibrated from real data. Table 7, like Table 5, presents the relative efficiency. Tables 6 and 7 give us the same pattern as Tables 4 and 5—the KL and SHE algorithms do better than the randomized item selection, the SHE algorithm does better than the KL algorithm, and the new algorithms, i.e., the PWKL and the HKL algorithms, outperform the KL and the SHE algorithms.

However, we also find that the numbers in Table 6 are less impressive than those in Table 4. This might be due to three reasons:

1. A much smaller item bank. The simulated bank has 300 items, whereas the real pool has only 36. Having a real pool consisting of hundreds of items implies constructing a huge **Q**-matrix. Though a very important task, it is very demanding in time and resources. Therefore in this study, we are limited by the size of the real item pool.
2. Complex **Q**-matrix. In the first simulation study, the **Q**-matrix is generated item by item, and attribute by attribute. The **Q**-matrix used in the second simulation study is of a much more complex structure (see McGlohen, 2004 for details). This makes the profile recovery more difficult.
3. Model fit. The calibrated guessing parameter (g) values are larger than ideal. A larger g suggests that the examinees' response are less governed by their latent cognitive state, and consequently, the estimation error gets bigger. It is analogous to having large c values in a three-parameter logistic IRT model, which is known to be a challenge to latent trait estimation. The high g values here suggest that the DINA model does not fit our data very well. It might be due to the fact that the exam is relatively easy: some attributes are mastered by almost every examinee (see Table 3).

Readers might wonder why the performances of the KL algorithm and the SHE algorithm differ so much when the KL information and Shannon entropy are closely connected. The reason is that (16) only holds when KL information and the Shannon entropy concern the same probability distribution **P**. The SHE algorithm looks at the posterior distribution of latent class estimate, while the KL algorithm assumes that every latent class is equally likely to be the true cognitive state. Therefore, it is not surprising that their performances differ. Moreover, the performances of the SHE algorithm and the PWKL algorithm are more comparable, because they both use the posterior distribution.

It is also important to note the difference in the computation time needed for each algorithm. The CPU times for all the five algorithms to select an item on a PC with Intel Core Duo CPU (E8335 2.66 GHz) are less than 0.01 second. However, among the four adaptive item selection algorithms, the KL algorithm works the fastest; PWKL and HKL are essentially equally fast; and the SHE algorithm is substantially slower.

4. Summary and Discussion

The two simulation studies suggest that the PWKL and the HKL algorithms are very promising as item selection methods for cognitive diagnosis. They improve considerably the efficiency of testing.

All the algorithms can be generalized to cognitive diagnostic models other than the DINA model. The only thing that needs to be changed is the item response function (IRF) defined in (24), which varies with the model.

These indices can also be generalized to conditions where the structure of latent states is restricted. For instance, some attributes may need to be acquired before others (Karelitz & de la Torre, 2008), which means the total number of possible latent states is less than 2^K. But this does not affect the generalizability of the indices introduced above. In (5), (10), (11), (13), and (15), the summation could be done only on the subset of possible states. Or with PWKL, we can assign 0 probabilities to those "impossible" states as the prior.

The simulation studies, however, are still limited as we discussed earlier. Just like any CAT, CD-CAT's performance does not only depend on the item selection algorithm, but also to a great extent the quality of the item bank. We believe that with a better item bank, for example, a larger bank that covers more adequately each attribute, advantages of the new algorithms can be better explored. A challenge that is unique to CD-CAT is that a **Q**-matrix needs be identified for the item bank, and research has shown that a misspecified **Q**-matrix can seriously affect the calibration of item parameters (e.g., Rupp & Templin, 2008). Therefore, our next line of research is to improve item banking with **Q**-matrix validation techniques.

In addition, the algorithms examined in this paper are general enough to be applied to other situations, which we would like to consider in our future work, such as variable-length CAT. Furthermore, the algorithms discussed here have not considered exposure control. Usually, cognitive diagnostic tests are relatively low-stakes, and test security is not a big concern. But if the test is high-stakes, exposure control becomes necessary. In future studies, we would like to consider different exposure control techniques and incorporate them into our item selection algorithms.

References

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

Chang, H., & Zhang, J. (2002). Hypergeometric family and item overlap rates in computerized adaptive testing. *Psychometrika*, *67*, 387–398.

Cover, T.M., & Thomas, J.A. (1991). *Elements of information theory*. New York: Wiley.

Embretson, S.E. (2001). *The second century of ability testing: Some predictions and speculations*. Retrievable at http://www.ets.org/Media/Research/pdf/PICANG7.pdf.

Haertel, E.H. (1989). Using restricted latent class models to map the skill structure of achievement items. *Journal of Educational Measurement*, *26*, 301–321.

Haertel, E.H. (1984). An application of latent class models to assessment data. *Applied Psychological Measurement*, *8*, 333–346.

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

Hambleton, R. & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston: Kluwer-Nijhoff.

Hartz, S. (2002). *A Bayesian framework for the Unified Model for assessing cognitive abilities: blending theory with practice*. Unpublished doctoral thesis, University of Illinois at Urbana-Champaign.

Hartz, S., Roussos, L., & Stout, W. (2002). *Skill diagnosis: Theory and practice [Computer software user manual for Arpeggio software]*. Princeton: ETS.

Henson, R., & Douglas, J. (2005). Test construction for cognitive diagnosis. *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.

Karelitz, T.M., & de la Torre, J. (2008). *When do measurement models produce diagnostic information? An investigation of the assumptions of cognitive diagnosis modeling*. In National Council on Measurement in Education Annual Meeting in New York, NY.

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

McGlohen, M.K. (2004). *The application of cognitive diagnosis and computerized adaptive testing to a large-scale assessment*. Unpublished doctoral thesis, University of Texas at Austin.

McGlohen, M.K., & Chang, H. (2008). Combining computer adaptive testing technology with cognitively diagnostic assessment. *Behavioral Research Methods*, *40*, 808–821.

Mislevy, R.J., Almond, R.G., Yan, D., & Steinberg, L.S. (1999). Bayes nets in educational assessment: Where do the numbers come from? In K.B. Laskey & H. Prade (Eds.). *Proceedings of the fifteenth conference on uncertainty in artificial intelligence* (pp. 437–446). San Mateo: Morgan Kaufmann.

Rupp, A., & Templin, J. (2008). The effects of Q-matrix misspecification on parameter estimates and classification accuracy in the DINA model. *Educational and Psychological Measurement*, *68*, 78–96.

Shannon, C.E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, *27*, 379–423.

Stocking, M.L. (1994). *Three practical issues for modern adaptive testing item pools* (ETS Research Rep. No. 94-5). Princeton: Educational Testing Service.

Tatsuoka, C. (2002). Data analytic methods for latent partially ordered classification models. *Journal of the Royal Statistical Society, Series C (Applied Statistics)*, *51*, 337–350.

Tatsuoka, C., & Ferguson, T. (2003). Sequential classification on partially ordered sets. *Journal of Royal Statistics, Series B*, *65*, 143–157.

Tatsuoka, K.K. (1983). Rule space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement*, *20*, 345–354.

Tatsuoka, K.K. (1995). Architecture of knowledge structures and cognitive diagnosis: a statistical pattern classification approach. In P. Nichols, S. Chipman, & R. Brennan (Eds.). *Cognitively diagnostic assessments* (pp. 327–359). Hillsdale: Erlbaum.

Templin, J. (2006). *CDM: cognitive diagnosis modeling using Mplus, user guide*. Retrievable at: http://www.iqgrads.net/jtemplin/downloads/CDM_user_guide.pdf.

Templin, J., & Henson, R. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychological Methods*, *11*, 287–305.

Thissen, D., & Mislevy, R.J. (2000). Testing algorithms. In H. Wainer et al. (Eds.). *Computerized adaptive testing: A primer* (pp. 101–133). Hillsdale: Erlbaum.

van der Linden, W.J. (1998). Bayesian item-selection criteria for adaptive testing. *Psychometrika*, *63*, 201–216.

von Davier, M. (2005). *A general diagnostic model applied to language testing data* (ETS Research Rep. No. RR-05-16). Princeton: ETS.

Xu, X., Chang, H., & Douglas, J. (2003). *Computerized adaptive testing strategies for cognitive diagnosis*. Paper presented at the annual meeting of National Council on Measurement in Education, Montreal, Canada.

Xu, X., & Douglas, J. (2006). Computerized adaptive testing under nonparametric IRT models. *Psychometrika*, *71*, 121–137.

Manuscript Received: 6 OCT 2008
Final Version Received: 12 MAR 2009
Published Online Date: 22 APR 2009