Graduate Management Admission Council®

Computerized Adaptive Testing for Cognitive Diagnosis

Ying Cheng

University of Notre Dame

Presented at the Diagnostic Testing Paper Session, June 3, 2009

Abstract

Computerized adaptive testing (CAT) has become a popular testing mode in recent decades, because it improves testing efficiency over paper-and-pencil testing while maintaining accuracy. Usually CAT programs are built upon latent trait item response theory models. Cognitive diagnostic models, on the other hand, are constrained latent class models. It is of interest to both researchers and practitioners to build CAT based on cognitive diagnostic models, so that testing efficiency and formative feedback can be offered simultaneously. This study explored several item selection algorithms that enable adaptive testing based on cognitive diagnostic models. Performance of these item selection algorithms is evaluated by attribute recovery rate and cognitive profile recovery rate.

Acknowledgment

Presentation of this paper at the 2009 Conference on Computerized Adaptive Testing was supported in part with funds from GMAC®.

Citation

Cheng, Y. (2009). Computerized adaptive testing for cognitive diagnosis. In D. J. Weiss (Ed.), *Proceedings of the 2009 GMAC Conference on Computerized Adaptive Testing.* Retrieved [date] from www.psych.umn.edu/psylabs/CATCentral/

Author Contact

Ying Cheng, 118 Haggar Hall, Notre Dame, IN 46556, U.S.A.
Email: ycheng4@nd.edu

Computerized Adaptive Testing for Cognitive Diagnosis

Cognitive diagnosis has received much attention recently, especially since the No Child Left Behind Act (2001) mandated that diagnostic feedback should be provided to teachers, students and parents. The test, therefore, not only serves evaluative purposes, but also offers valuable information regarding each examinee's educational needs. Instead of receiving a summative score, students receive a profile, specifying which concepts and skills they have mastered, and which areas require remedial instruction.

Various models have been proposed for cognitive diagnosis, including the rule space model (Tatsuoka, 1983); the binary skills model (Haertel, 1984; Haertel & Wiley, 1993); the Bayesian inference network (Mislevy, Almond, Yan, & Steinberg, 1999); and the conjunctive latent class models, such as the DINA model (Junker & Sijtsma, 2001), the NIDA model (Maris, 1999) and the Fusion model (Hartz, 2002; Hartz, Roussos, & Stout, 2002).

An interesting research question is: How can computerized adaptive testing (CAT) be used to help perform cognitive diagnosis more efficiently? This problem was originally addressed by Xu, Chang & Douglas (2005). They developed an item selection algorithm that estimates an examinee's cognitive profile based on his or her responses to previous items and selects the next item that maximizes the Kullback-Leibler (KL) index calculated at the latest profile estimate.

This paper provides several modifications of the algorithm by Xu et al. (2005): (1) a likelihood-weighted KL information method, (2) a posterior-weighted KL information method, and (3) a hybrid algorithm that incorporates both a posterior and a distance between latent states.

The DINA Model

The cognitive diagnostic model used in this study is the "Deterministic Input; Noisy 'And' Gate" (DINA) model (Junker & Sijstma, 2001), which relates item responses to a set of latent attributes. An attribute is a task, subtask, cognitive process, or skill involved in answering an item. The purpose of cognitive diagnosis is to identify which attributes are mastered by an examinee and which are not. 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 a Q matrix (Tatsuoka, 1995).

$\mathbf{Q}$ is a J × K matrix: $q_{jk} = 1$ if the correct response to item j requires the mastery of attribute k when there is no guessing and $q_{jk} = 0$ otherwise. The Q matrix is usually constructed by content experts and psychometricians. An example of a Q matrix is

$$\begin{bmatrix} 010 \\ 101 \\ 110 \end{bmatrix} \qquad (1)$$

This Q matrix indicates that the first item requires the second attribute, the second item requires the first and the last attributes 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/her mastery status, α, primarily accounts for the pattern of $\mathbf{U}i$. To make it more practical, the DINA model also allows "slipping" and "guessing." Here, slips and guesses are modeled at the item level. The parameter s_j indicates the probability of slipping on the jth item

when an examinee has mastered all the attributes required by it. The 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 that all the required attributes are mastered and $\eta_{ij} = 0$, otherwise. η_{ij} can be calculated as

$$\eta_{ij} = \prod_{k=1}^{K} \alpha_{ik}^{q_{jk}} \tag{2}$$

The item response function, therefore, can be written as:

$$P(U_{ij} = 1|\alpha) = (1 - s_j)^{\eta_{ij}} g_j^{1-\eta_{ij}} \tag{3}$$

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

$$Like(s, g; \alpha) = \prod_{i=1}^{N} \prod_{j=1}^{J} \left[(1 - s_j)^{y_{ij}} s_j^{1-y_{ij}}\right]^{\eta_{ij}} \left[g_j^{y_{ij}} (1 - g_j)^{1-y_{ij}}\right]^{1-\eta_{ij}} \tag{4}$$

We also would like to emphasize here that the cognitive model was not the focus of this study. We will limit the discussion to the item selection algorithms, and the algorithms discussed are applicable to different cognitive models. The concern was how the algorithms compared with each other.

The DINA model was chosen to be studied here because it requires only two easily interpretable parameters for each item, i.e. slipping and guessing. Based on the likelihood function and item responses, both item parameters and examinees' cognitive profiles can be estimated conveniently using MLE. Therefore, it is a good candidate for building a real-time system needed for a computerized adaptive testing program.

Previous Work on CAT for Cognitive Diagnosis

Cheng (2008) discussed a number of item selection methods that are based on Fisher information. However, Fisher information does not naturally lend itself to cognitive diagnosis because it is defined on a continuous variable. In a usual CAT, this condition is satisfied because the latent trait θ is a unidimensional continuum and the joint likelihood function is continuous and differentiable on θ. On the other hand, the latent structure underlying cognitive diagnosis involves multidimensional discrete latent classes $\hat{\boldsymbol{\alpha}}$. This leads us to another information measure, Kullback-Leibler (KL) information.

KL information is not a foreign concept to adaptive testing. Chang & Ying (1996) proposed the global information method which utilized the KL distance or information instead of Fisher information in item selection. They showed that global information is more robust and could be used to combat the instability of ability estimation in the early stage of CAT.

KL information is also used in CAT for cognitive diagnosis (e.g., Xu, et al., 2005) and cognitive diagnostic test construction (Henson & Douglas, 2005). They defined an index for item

selection based on the KL distance between the item responses conditional on the estimated latent profile, i.e. the estimated latent cognitive state or class $\hat{\boldsymbol{\alpha}}$ and the item responses conditional on all the other possible latent states.

McGlohen (2004) proposed a two-step item selection scheme to provide both cognitive diagnosis and the usual ability estimate. In her study, first a shadow test (van der Linden, 2000; van der Linden & Chang, 2003) is constructed so that all the non-statistical constraints are met and the test reaches optimality with respect to θ. Then one item is selected from this shadow test to optimize with respect to $\hat{\boldsymbol{\alpha}}$ based on the KL index (McGlohen, 2004). She showed that fairly good estimation of both θ and $\boldsymbol{\alpha}$ can be achieved simultaneously.

Method

Xu et al.'s (2005) KL Algorithm

KL information is a measure of the "distance" between two probability distributions (Cover & Thomas, 1991):

$$d[f,g] = E_f\left[\log\frac{f(x)}{g(x)}\right] \tag{5}$$

Here $f(x)$ and $g(x)$ are two probability distributions. Note that KL information is not strictly a distance measure because it is not symmetric: $d[f, g] \neq d[g, f]$. The reason why it is sometimes referred to as KL distance is that the larger $d[f, g]$ is, the easier it is to statistically discriminate between the two probability distributions $f(x)$ and $g(x)$ (Henson & Douglas, 2005).

In the usual CAT context, the probability distribution we are concerned with is the distribution of the item responses, X, given the latent trait level θ. When X is dichotomous, the KL information defined as follows quantifies how well the jth item discriminates between $f\left(X|\theta\right)$ and $f\left(x|\hat{\theta}\right)$:

$$KL_j(\theta||\hat{\theta}) = P_j(\hat{\theta})\log[\frac{P_j(\hat{\theta})}{P_j(\theta)}] + [1-P_j(\hat{\theta})]\log[\frac{1-P_j(\hat{\theta})}{1-P_j(\theta)}] \tag{6}$$

Chang & Ying (1996) proposed a global information measure to capture the power of the jth item to discriminate between the distribution of X conditional on $\hat{\theta}$ and the distribution of X conditioning on the neighboring points of $\hat{\theta}$, which is essentially the expected KL information $\mathrm{KL}\left(\theta\middle\|\hat{\theta}\right)$ over an interval around $\hat{\theta}$:

$$G_j(\hat{\theta}) = \int_{\hat{\theta}-\delta}^{\hat{\theta}+\delta} KL_j(\theta||\hat{\theta})d_\theta \tag{7}$$

where δ can take on the value of $c/\sqrt{m}$ where c can be selected according to a specified coverage probability. Following Chang & Ying (1996), c is set to 3 in this study, and m is the number of items that have already been administered in a test. This means that as the test progresses, the integration is done over a smaller interval around $\hat{\theta}$, so the global information measure becomes more focused. In fact as m $\rightarrow\infty$, the global information becomes just local information. See

Chang & Ying (1996) for more a detailed discussion.

In cognitive diagnosis, we are interested in the conditional distribution of person *i*'s item responses U_{ij} given $\boldsymbol{\alpha}$. The KL distance between the distribution of U_{ij} conditioning on the current estimated latent state, i.e., $f\left(U_{ij}\left|\hat{\boldsymbol{\alpha}}\right.\right)$ and the conditional distribution of U_{ij} given another latent state $\boldsymbol{\alpha}_c$, i.e., $f\left(U_{ij}\left|\boldsymbol{\alpha}_c\right.\right)$, can be computed as follows:

$$kl_{ij}(\hat{\boldsymbol{\alpha}}||\boldsymbol{\alpha}_c)=\sum_{u=0}^{1}\log(\frac{P(U_{ij}=u|\hat{\boldsymbol{\alpha}})}{P(U_{ij}=u|\boldsymbol{\alpha}_c)})P(U_{ij}=u|\hat{\boldsymbol{\alpha}}) \tag{8}$$

To distinguish it from the final KL index, here lower case *kl* is used.

In order to obtain a measure of the global discrimination power of item *j* between $f\left(U_{ij}\left|\hat{\boldsymbol{\alpha}}\right.\right)$ and all possible latent classes or states, Xu et al. (2005) proposed the following index (i.e., the KL index), which is basically the sum of the KL distances between $f\left(U_{ij}\left|\hat{\boldsymbol{\alpha}}\right.\right)$ and all $f\left(U_{ij}\left|\boldsymbol{\alpha}_c\right.\right)$s

$$KL_{ij}(\hat{\boldsymbol{\alpha}})=\sum_{c=1}^{2^K}[\sum_{u=0}^{1}\log(\frac{P(U_{ij}=u|\hat{\boldsymbol{\alpha}})}{P(U_{ij}=u|\boldsymbol{\alpha}_c)})P(U_{ij}=u|\hat{\boldsymbol{\alpha}})] \tag{9}$$

Note that when there are K attributes, there will be 2^K possible latent cognitive states. Assuming local independence, the KL index is additive, i.e., the discriminating power of a test can be defined as the summation of the item KLs.

New Indices

There is an implicit assumption made in the computation of the KL index—, that that is all the latent states $\boldsymbol{\alpha}_c$ (c =1, 2, … , 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 indices in which the condition is relaxed.

Likelihood-Weighted KL (LWKL) index. As the test progresses, we learn more about an examinee and some latent states $\boldsymbol{\alpha}_c$ can be quickly ruled out, or at least we know that they are less likely than others to be the possible true state. Therefore, it is plausible to weight the $\boldsymbol{\alpha}_c$s by their corresponding likelihoods. Suppose the examinee *i* has answered *t* items and his or her item response vector is $U_{i1} = u_{i1}, U_{i2} = u_{i2}, \ldots, U_{it} = u_{it}$. Denote the vector by $\mathbf{U}_i^{(t)} = \mathbf{u}_i^{(t)}$. The joint likelihood of a latent state $\boldsymbol{\alpha}_c$ and the item response vector $\mathbf{U}_i^{(t)} = \mathbf{u}_i^{(t)}$:

$$Like\left(\boldsymbol{\alpha}_c;\mathbf{u}_i^{(t)}\right)=\prod_{l=1}^{t}p_{cil}^{u_{il}}\left(1-p_{cil}\right)^{1-u_{il}} \tag{10}$$

where

$$p_{cil}=\text{Prob}\left(u_{il}=1|\boldsymbol{\alpha}_c\right) \tag{11}$$

can be computed from Equation 3. The resulting new LWKL index can be written as:

$$LWKL_{ij}(\hat{\alpha}) = \sum_{c=1}^{2^K} \left\{ \left[\sum_{u=0}^{1} \log \left(\frac{P(U_{ij} = u|\hat{\alpha})}{P(U_{ij} = u|\boldsymbol{\alpha}_c)} \right) P(U_{ij} = u|\hat{\alpha}) \right] Like\left(\boldsymbol{\alpha}_c; \mathbf{u}_i^{(t)}\right) \right\} \tag{12}$$

The (t + 1)th item to be selected for the ith examinee is therefore:

$$\max_{j \in \Omega^t} \{LWKL_{ij}\} \tag{13}$$

where Ω^t is the set of eligible items at stage t.

Posterior-Weighted KL (PWKL) index. We might be able to infer the probabilistic distribution of the latent states in the current sample by analyzing old samples. For example, it is reasonable to assume that this year's examinee population does not differ much from last year's. Therefore, we can impose an informative prior on the latent states and obtain the posterior distribution at each step. Suppose the prior follows

$$\Pr(\boldsymbol{\alpha}_c) = p_{0c} \tag{14}$$

subject to $\sum_{c=1}^{2^K} p_{0c} = 1$ and $p_{0c} \geq 0$ $\forall c = 1, 2, \ldots, 2^K$.

The posterior of the latent states given t item responses is therefore:

$$\text{Prob}\left(\boldsymbol{\alpha} = \boldsymbol{\alpha}_c | \mathbf{u}_i^{(t)}\right) = \frac{p_{0c} \cdot f\left(\mathbf{u}_i^{(t)} | \boldsymbol{\alpha}_c\right)}{\sum_{c=1}^{2^K} p_{0c} \cdot f\left(\mathbf{u}_i^{(t)} | \boldsymbol{\alpha}_c\right)} \tag{15}$$

for $\forall c = 1, 2, \ldots, 2^K$. Denote the posterior as g_t. The PWKL index can thus be defined as

$$PWKL_{ij}(\hat{\alpha}) = \sum_{c=1}^{2^K} \left\{ \left[\sum_{u=0}^{1} \log \left(\frac{P(U_{ij} = u|\hat{\alpha})}{P(U_{ij} = u|\boldsymbol{\alpha}_c)} \right) P(U_{ij} = u|\hat{\alpha}) \right] g_t(\boldsymbol{\alpha}_c) \right\} \tag{16}$$

Note that the denominator is a constant for a given examinee i and a fixed stage t. Consequently, it does not affect item selection and can be dropped. In other words, the following index will work the same as PWKL does:

$$\sum_{c=1}^{2^K} \left\{ \left[\sum_{u=0}^{1} \log \left(\frac{P(U_{ij} = u|\hat{\alpha})}{P(U_{ij} = u|\boldsymbol{\alpha}_c)} \right) P(U_{ij} = u|\hat{\alpha}) \right] p_{0c} \cdot f\left(\mathbf{u}_i^{(t)} | \boldsymbol{\alpha}_c\right) \right\} \tag{17}$$

The (t + 1)th item to be selected for the ith examinee is therefore:

$$\max_{j \in \Omega^t} \{PWKL_{ij}\} \tag{18}$$

where Ω^t is the set of eligible items at stage t.

Relationship between LWKL and PWKL indices. If a discrete uniform prior is imposed, i.e., $p_{0c} = p_{0c'}$ holds for ∀ pairs of c and c', PWKL further degenerates into

$$\sum_{c=1}^{2^K}\left\{\left[\sum_{u=0}^{1}\log\left(\frac{P\left(U_{ij}=u|\hat{\boldsymbol{\alpha}}\right)}{P\left(U_{ij}=u|\boldsymbol{\alpha}_c\right)}\right)P\left(U_{ij}=u|\hat{\boldsymbol{\alpha}}\right)\right]f\left(\mathbf{u}_i^{(t)}|\boldsymbol{\alpha}_c\right)\right\} \tag{19}$$

In fact,

$$f\left(\mathbf{u}_i^{(t)}|\boldsymbol{\alpha}_c\right)=Like\left(\boldsymbol{\alpha}_c;\mathbf{u}_i^{(t)}\right) \tag{20}$$

So when a discrete uniform prior is imposed, PWKL becomes:

$$\sum_{c=1}^{2^K}\left\{\left[\sum_{u=0}^{1}\log\left(\frac{P\left(U_{ij}=u|\hat{\boldsymbol{\alpha}}\right)}{P\left(U_{ij}=u|\boldsymbol{\alpha}_c\right)}\right)P\left(U_{ij}=u|\hat{\boldsymbol{\alpha}}\right)\right]Like\left(\boldsymbol{\alpha}_c;\mathbf{u}_i^{(t)}\right)\right\} \tag{21}$$

which is nothing but the LWKL index.

Hybrid KL (HKL) index incorporating distance between latent states. The modification done to the original KL index in the LWKL and PWKL considers the fact that not all the latent states are equally likely. These indices, however, overlook the fact that the distances between different latent states and the current estimate are not all of equal importance. Henson (2005) noted that if an item "discriminates well between attribute patterns which are similar, it will discriminate well between those that are dissimilar."

One common measure of distance is the Euclidean distance:

$$d(\boldsymbol{\alpha}_c,\boldsymbol{\alpha}_{c'})=\sqrt{\sum_{k=1}^{K}\left(\alpha_{ck}-\alpha_{c'k}\right)^2} \tag{22}$$

The LWKL/ PWKL index, therefore, can be weighted by the inverse of the distance between the $\hat{\boldsymbol{\alpha}}$ and any possible latent state. Since LWKL can be considered a special case of PWKL, we will use PWKL as an example. The HKL can be defined as

$$HKL_{ij}(\hat{\boldsymbol{\alpha}})=\sum_{c=1}^{2^K}\left\{\left[\sum_{u=0}^{1}\log\left(\frac{P\left(U_{ij}=u|\hat{\boldsymbol{\alpha}}\right)}{P\left(U_{ij}=u|\boldsymbol{\alpha}_c\right)}\right)P\left(U_{ij}=u|\hat{\boldsymbol{\alpha}}\right)\right]\frac{g_t\left(\boldsymbol{\alpha}_c\right)}{d(\boldsymbol{\alpha}_c,\hat{\boldsymbol{\alpha}})}\right\} \tag{23}$$

Note that one of the $\hat{\boldsymbol{\alpha}}$s must be the same as the current $\hat{\boldsymbol{\alpha}}$. The corresponding d($\alpha_c$, $\hat{\alpha}$) will be 0 and cannot be the denominator. In that case, the $d\left(\boldsymbol{\alpha}_c,\hat{\boldsymbol{\alpha}}\right)$ is set to be 0.01.

Are these indices generalizable? The indices introduced above (including the original KL index) can all be generalized to other models than the DINA model. The only thing that needs to be changed is the probability defined in Equations 3 and 4, 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 might need to be acquired before others (Karelitz & de la Torre, 2008). In this case, the total number of possible latent states should be less than 2^K. But this does not affect the generalizability of the indices introduced above. The only change that

needs to be made in Equations 12, 16 and 23 is that the summation should be done only on the set of possible states. Or with PWKL, the only thing we need to do is to assign 0 probabilities to those "impossible" states as the prior.

Xu et al.'s (2005) Algorithm Based on Shannon's Entropy

Xu et al. (2005) also proposed using Shannon's Entropy (SHE) as an item-selection criterion. Their study showed that the SHE algorithm worked better than the KL algorithm across all conditions.

Shannon's entropy is a measure of uncertainty associated with a random variable, first proposed by Claude E. Shannon (Shannon, 1948). An example is a fair coin, which has an entropy of one unit. An unfair coin would have a lower entropy, because when we bet on the outcome, we would be more than 50% in favor of either head or tail, meaning that the uncertainty is reduced. In modern communication theory, Shannon's entropy quantifies the information contained in a message, usually in bits or bits/symbol.

The Shannon's entropy of a discrete random variable X, that can take on possible values x_1, x_2, … ,x_n is formally defined as:

$$H(X) = -\sum_{i=1}^{n} p(x_i) \log_b (x_i) \tag{24}$$

where $p(x_i) = Pr(X = x_i)$. We will use P for (p_1, p_2, … ,p_n). Then $H(X)$ can also be written as $H(p_1, p_2, \ldots, p_n)$ or $H(P)$, suggesting that the Shannon's entropy is a measure of uncertainty associated with a distribution. Obviously, the p_is need to satisfy $p_i \geq 0$ for $\forall i = 1, 2, \ldots, n$ and $\sum_{1}^{n} p_i = 1$.

The logarithm is used so that independent uncertainties are additive, and b is the base of the logarithm. Possible values of b are 2, e and 10. They only change is the unit in which the uncertainty is measured. With $b = 2$, the unit is bit; with $b = e$, the unit is nat and with $b = 10$, the unit is dit or digit. In the following, b will be omitted because it does not affect item selection, where we are basically finding the smallest expected $H(P)$ regardless of the unit in which $H(P)$ is measured.

The $H(P)$ defined above satisfy the following properties:

1. Nonnegativity:

$$H(p_1, p_2, \cdots, p_n) \geq 0 \tag{25}$$

   with equality holding iff $\exists p_i = 1$ and $p_j = 0$ when $j = i$, which represents the most "concentrated", or most "certain" distribution.

2. Maximality. $H(P)$ reaches maximum when all the outcomes are equally likely:

$$H(p_1, p_2, \cdots, p_n) \leq H\left(\frac{1}{n}, \frac{1}{n}, \cdots, \frac{1}{n}\right) \tag{26}$$

   with equality holding iff $p_i = \frac{1}{n} \forall i = 1, 2, \ldots, n$. For equiprobable events, the entropy

should increase with the number of outcomes, i.e.,

$$H\left(\frac{1}{n},\frac{1}{n},\cdots,\frac{1}{n}\right) < H\left(\frac{1}{n+1},\frac{1}{n+1},\cdots,\frac{1}{n+1}\right) \tag{27}$$

3. Concavity. $H(P)$ is a concave function of P.
4. Continuity. $H(P)$ is a continuous function of P, so that changing the value of probabilities by a very small amount should only change the entropy by a small amount.
5. Symmetry. $H(P)$ is a symmetric function of its arguments, i.e.,

$$H\left(p_{\tau(1)},p_{\tau(2)},\cdots,p_{\tau(n)}\right) \tag{28}$$

   where τ defines any permutation from 1 to n.

6. Expansible, i.e., $H(p_1, p_2, \ldots, p_n) = H(p_1, p_2, \ldots, p_n, 0)$.

There are many other properties of $H(P)$ that are not listed here. The above properties suffice for our discussion here. Theories around Shannon's entropy have been developed over the last several decades, and researchers have tried to relate Shannon's Entropy to the concept of "entropy" raised in the field of thermodynamics .

The SHE algorithm. Xu et al. (2005) proposed an algorithm which tries to minimize the expected Shannon entropy of the posterior distribution of the $\boldsymbol{\alpha}$, which we will refer to as the SHE algorithm. Following the notation in Equations 10, 15 and 20, the posterior after t responses can be written as:

$$g_t(\boldsymbol{\alpha}_c) = \frac{p_{0c}\prod_{l=1}^{t} p_{cil}^{u_{il}}(1-p_{cil})^{u_{il}}}{\sum_{c=1}^{2^K} p_{0c}\prod_{l=1}^{t} p_{cil}^{u_{il}}(1-p_{cil})^{u_{il}}} = \frac{p_{0c}\prod_{l=1}^{t} p_{cil}^{u_{il}}(1-p_{cil})^{u_{il}}}{\varphi_{it}} \tag{29}$$

where

$$\varphi_{it} = \sum_{c=1}^{2^K} p_c \prod_{l=1}^{t} p_{cil}^{U_{il}}\left(1-p_{cil}^{U_{il}}\right), \tag{30}$$

which is a normalizing constant for given i and t, so that the posterior is proper. Then the Shannon entropy of the posterior distribution g_t can be written as:

$$H(g_t) = -\sum_{c=1}^{2^K} g_t(\boldsymbol{\alpha}_c)\log\left(g_t(\boldsymbol{\alpha}_c)\right) \tag{31}$$

Consider item j in the item bank. Denote the answer to it as X_j which can take a value of either 0 or 1. Suppose item j is selected and administered, and a response to it is recorded, i.e., X_j is realized as a value x_j. The item response vector is now of length $t + 1$ and it becomes $\mathbf{u}_i^{(t+1)} = \left(u_{i1}, u_{i2}, \ldots, u_{it}, x_j\right)$. Then following Equation 29, we can update the posterior of the latent states based on u_i, i.e., the $g_{t+1}(\boldsymbol{\alpha}_c)$s. Consequently, the Shannon entropy associated with g_{t+1}, i.e., $H(g_{t+1})$ can be updated as well, according to Equation 30.

However, since X_j is a random variable, both g_{t+1} and $H(g_{t+1})$ are unknown. But the

distribution of X_j can be defined as the posterior predictive distribution given the responses to earlier items, i.e.,

$$\Pr\left(X_j = x|\mathbf{u}_i^{(t)}\right) = \sum_{c=1}^{2^K}\left[\Pr\left(X_j = x|\boldsymbol{\alpha}_c\right) g_t\left(\boldsymbol{\alpha}_c\right)\right] \tag{32}$$

Therefore, we can obtain the expected predictive Shannon Entropy (SHE) index for item *j* as

$$SHE_{ij}\left(g_{t+1}\right) = \sum_{x=0}^{1} H\left(g_{t+1}|X_j = x\right)\Pr\left(X_j = x|\mathbf{u}_i^{(t)}\right) \tag{33}$$

where

$$g_{t+1}|X_j = x\left(\boldsymbol{\alpha}_c\right) = \frac{p_{0c}\prod_{l=1}^{t} p_{cil}^{u_{il}}\left(1 - p_{cil}^{u_{il}}\right)}{\varphi_{i(t+1)}} p_{cj}^{x}\left(1 - p_{cj}\right)^{1-x} \tag{34}$$

and:

$$H\left(g_{t+1}|X_j = x\right) = -\sum_{c=1}^{2^K} g_{t+1|X_j=x}\left(\boldsymbol{\alpha}_c\right)\log\left(g_{t+1|X_j=x}\left(\boldsymbol{\alpha}_c\right)\right) \tag{35}$$

Finally, the (t + 1)th item to be selected for the ith examinee is

$$\min_{j\in\Omega^t}\left\{SHE_{ij}\right\} \tag{36}$$

where Ω^t is the set of eligible items at stage t.

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

$$EPV_{ij} = \left\{\sum_{x=0}^{1}\Pr\left(X_j = x|\mathbf{u}_i^{(t)}\right) Var\left(\mathbf{u}_i^{(t)}, X_j = x\right)\right\} \tag{37}$$

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

$$EPV_{ij} = \left\{\sum_{x=0}^{1}\Pr\left(X_j = x|\mathbf{u}_i^{(t)}\right) Var\left(\mathbf{u}_i^{(t)}, X_j = x\right)\right\} \tag{38}$$

Comparing Equation 37 with 32, shows that the MEPV algorithm essentially follows the same logic as the SHE algorithm. The only difference is that the SHE algorithm uses the Shannon entropy of the posterior whereas MEPV uses the posterior variance. This suggests that the criterion of "minimum expected Shannon entropy of the posterior" can be applied to CAT.

Data and Simulation Design

A simulation study was designed to evaluate the performance of the proposed new indices. The following are details of the simulation:

1. *Examinees.* The examinees were generated assuming that every examinee had a 50% chance of mastering each attribute. In other words, for a 6-attribute test, the 64 cognitive states were equally likely in the population. In total 2,000 $\boldsymbol{\alpha}$s were generated. These were the true $\boldsymbol{\alpha}$s that are used to generated the item responses.
2. *Q matrix, slipping and guessing parameters.* The Q matrix used in this study was generated item-by-item and attribute-by-attribute. Each item had a 20% chance of measuring each attribute. Following De la Torre and Douglas (2004), *g* and *s* were generated from a 4-Beta(0, 0.9, 1.5, 2) distribution. Note that in total 300 items were generated.
3. *Cognitive diagnosis model.* The DINA model was used to generate item responses.
4. *Estimation.* The initial $\hat{\boldsymbol{\alpha}}$, i.e., $\hat{\boldsymbol{\alpha}}^0$ was randomly generated. Each attribute was generated independently following a Bernoulli distribution with $p = 0.5$. Then MLE was used to update the $\hat{\boldsymbol{\alpha}}^t$s. The final estimates were $\hat{\boldsymbol{\alpha}}^L$s, where L is the test length. In this study, $L = 12$.
5. *Item selection rule.* The original KL index, the SHE index, and the new indices were used to select items. Note that since the examinees were generated in such a way that all latent states were equally likely, the LWKL index and the PWKL index were the same in this simulation. In order to show the effect of adaptive item selection, fully randomized item selection was also studied as the baseline. Five item selection methods were considered: KL, SHE, PWKL (or LWKL), HWL and randomized (RANDOM) selection.

Results were analyzed according to the recovery rate of the cognitive profiles, including recovery of the whole pattern and of each attribute. This can be accomplished by comparing each $\boldsymbol{\alpha}$ with the $\hat{\boldsymbol{\alpha}}^L$s. For instance, if the true $\boldsymbol{\alpha}$ was [0 1 1 1 0 1] and the final $\hat{\boldsymbol{\alpha}}^L$ was also [0 1 1 1 0 1], then we say that the whole pattern was recovered. If the $\hat{\boldsymbol{\alpha}}$ yielded [0 0 0 0 1 0], then only the first attribute was recovered. There were 2,000 pairs of $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\alpha}}^L$, so the recovery rates for the entire pattern and each attribute could be calculated by comparing every pair of $\boldsymbol{\alpha}$ and $\hat{\boldsymbol{\alpha}}^L$.

One important aspect overlooked in the current literature on cognitive diagnosis is that the recovery rate computed in the way described above is inflated, because an attribute or pattern can be recovered by sheer chance. Therefore, chance correction procedures needed to be applied to the raw recovery rates. The procedure was as follows:

Step 1. Compute from the true cognitive profiles as the proportion of examinees mastering each attribute. This serves as the base rate for each attribute. Because in this simulation the true $\boldsymbol{\alpha}$s were generated so that each attribute had 50% chance to be 1 and the attributes were independent, the base rate for each attribute should be 0.5.

Step 2. The chance corrected recovery rate for the *k*th attribute (CRR_k) was computed based on the corresponding raw recovery rate (RR_k) and the base rate (BR_k) obtained in Step 1:

$$CRR_k = \frac{RR_k - BR_k}{1 - BR_k} \tag{39}$$

Step 3. The chance corrected pattern recovery rate (CPR) was computed as follows:

$$CPR = \frac{PR - \prod_k^K BR_k}{1 - \prod_k^K BR_k} \tag{40}$$

where PR represents the raw pattern recovery rate. Obviously it was assumed that all the attributes are independent, which was reasonable in the current simulation, but when attributes are correlated or ordered (i.e., one attribute must be mastered before another), this mechanism should be changed accordingly.

Results

Table 1 compares the five item selection methods. The first three rows can be viewed as a replication of Xu et al. (2005). It does show the same pattern: the SHE algorithm did a fairly good job and outperformed the KL algorithm and the randomization item selection method.

Table 1. Recovery Rates of the Examinees' Cognitive Profiles Before and After Chance Correction

	Attribute						Entire
Method	1	**2**	**3**	**4**	**5**	**6**	Pattern
Before Chance Correction							
Random	0.852	0.819	0.864	0.848	0.838	0.834	0.366
KL	0.964	0.655	0.899	0.792	0.949	0.949	0.415
SHE	0.990	0.950	0.960	0.948	0.957	0.964	0.811
LWKL	0.989	0.984	0.988	0.987	0.999	0.986	0.931
HWL	0.992	0.986	0.988	0.993	0.991	0.979	0.935
After Chance Correction							
Random	0.704	0.638	0.728	0.696	0.676	0.668	0.356
KL	0.928	0.310	0.798	0.584	0.898	0.898	0.406
SHE	0.980	0.900	0.920	0.896	0.914	0.928	0.808
LWKL	0.978	0.968	0.976	0.974	0.998	0.972	0.930
HWL	0.984	0.972	0.976	0.986	0.982	0.958	0.934

What is of more concern is the performance of the new indices, as shown in the last two rows of Table 1. Clearly, the likelihood or posterior-based KL indices (PWKL and HWL) improved the recovery rates substantially. Not surprisingly, they outperformed the original KL algorithm and the randomization algorithm. The improvement was uniform, meaning that the recovery rates of every attribute and of the whole pattern were all better.

The PWKL and HWL algorithms also outperformed the SHE algorithm. The improvement of the PWKL algorithm over the SHE algorithm was almost uniform, except for the first attribute. The HWL algorithm outperformed the SHE algorithm in all conditions.

Between the PWKL and HWL method, it was to determine which performed better. The PWKL method led to higher recovery rates on Attributes 5 and 6, and the HWL method performed better with respect to all the other attributes and the whole pattern. In general, the difference between the PWKL and the HWL methods were fairly small.

It is also important to note the difference in the computation time needed for each algorithm. The KL algorithm was the fastest. PWKL and HWL were essentially equally fast. However, the SHE algorithm was substantially slower.

Following Xu et al. (2005), all the whole pattern recovery rates were divided by the whole pattern recovery rate of the randomization method. The results are summarized in Table 2. Clearly, before chance correction the SHE, LWKL and HKL algorithms were more than twice as efficient as the randomization method. Actually the LWKL and HKL algorithms were about 2.5 times as efficient as the baseline, and were about 1.14 times (i.e., 2.5/2.2) as efficient as the SHE algorithm. After chance correction, the gain over the randomization method appeared to be even larger.

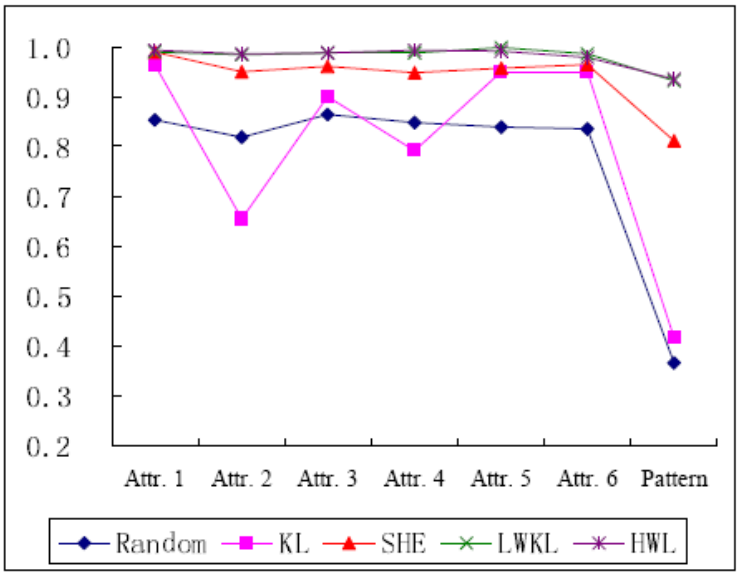
Table 2. Relative Efficiency of the Item Selection Algorithms Before and After Chance Correction

	Chance Correction	
Method	Before	After
Random	1.000	1.000
KL	1.134	1.140
SHE	2.216	2.270
PWKL	2.544	2.612
HKL	2.555	2.624

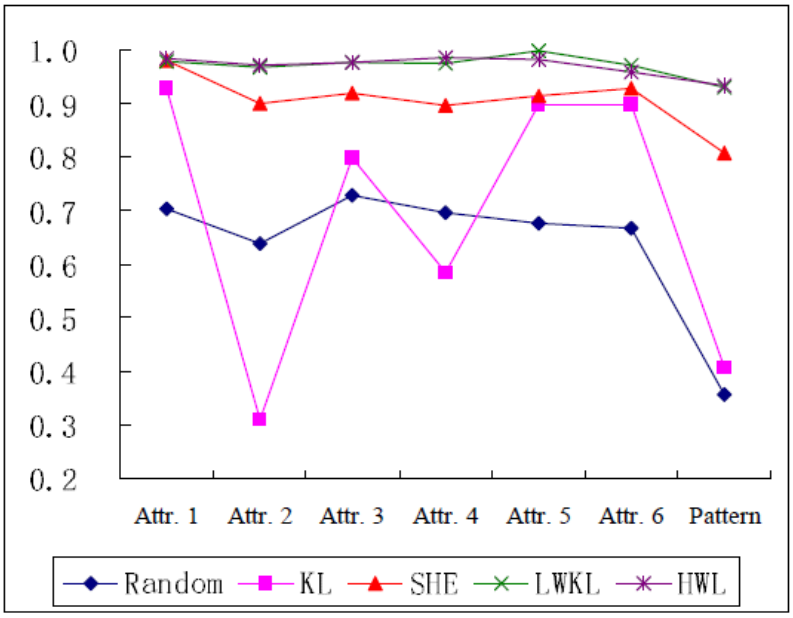
Figure 1 displays the data in Tables 1 and 2, before and after chance correction. Both figures show that the PWKL and the HWL algorithms worked well in terms of recovery of both each attribute and the entire pattern. They also shows that the original KL algorithm was not very consistent across attributes. Comparing the two figures, it is clear that chance correction magnified the difference among the algorithms.

Figure 1. Recovery Rates of the Examinees' Cognitive Profiles

a. Before Chance Correction



b. After Chance Correction



Conclusions

The current study showed that the PWKL and the HWL algorithms are very promising as item selection methods for cognitive diagnosis. They outperformed both the original KL and SHE algorithms. However, this study is limited in several aspects. First of all, all the item parameters and the Q matrix were generated. The reason why we choose to generate data is that the identification of the Q matrix is very complicated and time-consuming, and usually requires substantive knowledge. It is exceptionally difficult to specify a Q matrix for an entire item bank, which typically contains hundreds of items.

Secondly, the attributes were assumed to be independent during both the Q-matrix and examinee generation. In real tests, however, this assumption is often violated. Attributes might well be correlated or even ordered (de la Torre & Douglas, 2004; Karelitz & de la Torre, 2008). So the Q-matrix and examinee cognitive profiles generated this way might be unrealistic. Understanding the relation among the attributes might help improve the algorithms and make them more efficient.

In the future, we would like to replicate the study with a real item bank and real cognitive profiles, with a more complex Q matrix, and item parameters, and cognitive patterns calibrated from real test data. In addition, the algorithms are general enough to be applied to other situations, which we would like to consider in our future work as well:

1. *Variable-length CAT.* The key to variable-length CAT is the termination rule. In typical CAT, the termination rule is often $\mathrm{Var}\left(\hat{\theta}\right) < \varepsilon$ or $I\left(\hat{\theta}\right) > 1/\varepsilon$ where $\varepsilon \in \mathrm{R}^{+}$ is a very small number. We propose the following termination rules for use in cognitive diagnosis, i.e., the test stops when:

   (a) The Shannon entropy of the posterior or the change in the adjacent Shannon entropies becomes reasonably small:

   $$H(g_t) < \varepsilon \tag{41}$$

   or

   $$|H(g_{t-1}) - H(g_t)| < \varepsilon \tag{42}$$

   (b) The KL distance between two adjacent posteriors becomes small enough:

   $$KL(g_t || g_{t-1}) < \varepsilon \tag{43}$$

2. *Exposure control.* The current algorithms, including the original KL and SHE algorithms, as well as the new algorithms proposed here, do not consider any exposure control technique. Usually cognitive diagnostic tests are relatively low-stakes and test security is not a major concern. But if the cognitive diagnostic 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.

3. *Typical CAT.* Algorithms discussed here have important implications for CATs. For example, the SHE algorithm might well by applied to typical CATs. The only thing that needs to be changed is the computation of the posterior, and the corresponding Shannon

entropy. Note that in typical CATs, the Shannon entropy is defined on a continuous variable θ instead of the discrete latent states.

Similarly, the PWKL or LWKL index can be applied to usual CAT with continuous θ:

$$PWKL_j(\hat{\theta}) = \int_{\theta_l}^{\theta_u} KL_j(\theta||\hat{\theta}) g_t(\theta)\, d\theta \tag{44}$$

where $g_t(\theta)$ is the posterior distribution of θ after t items are answered. θ_l and θ_u are lower- and upper-bounds of θ, which defines the integration region.

References

Cheng, Y. (2008). Computerized adaptive testing: New developments and applications. Unpublished doctoral thesis, University of Texas at Austin.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20,* 37-46.

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

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, NJ: Lawrence Erlbaum Associates.

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, NJ: 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. National Council on Measurement in Education Annual Meeting in New York, NY.

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

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.

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 Francisco: Morgan Kaufmann.

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

U.S. House of Representatives (2001), "Text of No Child Left Behind Act".

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

van der Linden, W. J., & Chang, H.-H. (2003). Implementing content constraints in alpha-stratified adaptive testing using a shadow test approach. *Applied Psychological Measurement, 27*, 107-120.

Xu, X., Chang, H., & Douglas, J. (2005). Computerized adaptive testing strategies for cognitive diagnosis. Paper presented at the annual meeting of National Council on Measurement in Education, Montreal, Canada.