Graduate
Management
Admission
Council®

Obtaining Reliable Diagnostic Information through Constrained CAT

Chun Wang, Hua-Hua Chang,
and Jeff Douglas
University of Illinois at Urbana-Champaign

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

Abstract

Computerized adaptive testing (CAT) has the advantage of delivering tests in an interactive manner such that the ability or latent traits are more effectively estimated. Until now, most research concerning item selection rules in CAT has been built upon either item response theory (IRT) or cognitive diagnostic models (CDM) separately. The only study that combined these two approaches together was done by McGlohen and Chang (2008). They proposed a two-stage method, in which a "shadow" test functioned as a bridge to connect information gathered at θ for IRT, and information accumulated at α for CDM. In this paper, we develop a one-stage method to build a CAT featuring reliable cognitive diagnosis. The major idea is to treat diagnostic information as various constraints, and by using a maximum priority index (MPI) method to meet these constraints the cognitive diagnosis can be done reliably at the end of the test. Several priority functions are proposed, some based upon formal measures of information, like Kullback-Leibler information, and others only utilize the knowledge of which items measure what attributes, as provided by the Q matrix. Simulation studies and their results are reported. We show how utilization of information-based methods both yields higher classification rates for cognitive diagnosis and achieve accurate θ estimation. Item exposure rates are also considered for all competing methods.

Acknowledgment

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

Copyright © 2009 by the Authors

Citation

Wang, C., Chang, H., & Douglas, J. (2009). Obtaining reliable diagnostic information through constrained CAT. 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

Jeffery A. Douglas. 116E Illini Hall, 725 S. Wright St. Champaign, IL 61820 , U.S.A.
Email: jeffdouglas24@gmail.com

Obtaining Reliable Diagnostic Information Through Constrained CAT

Computerized adaptive testing (CAT) has become popular in many high-stakes educational testing programs. In this research, we study the utility and efficiency of modifications of the constraint weighted *a*-stratification approach to CAT that guarantees efficient estimation of a trait (θ) will provides sufficient diagnostic information to report scores on fine-grained skills.

In educational testing research, specialized latent class models for cognitive diagnosis have been developed to classify mastery or non-mastery of each attribute in a set of attributes the exam is designed to assess. As in item response theory (IRT), item parameters can be of interest. However, the ultimate goal of applying diagnostic models is to classify examinees into one of several different categories describing their attribute profiles. These attributes can take many forms, depending on the application, but often correspond one-to-one with specific skills needed to answer items on an exam or other psychological assessment. In many cases, dependency in the skills or attributes can be explained by a single continuous and broadly-defined θ, and such a fine breakdown of the skills required for an exam is often secondary to the primary goal of estimating the standard unidimensional θ. We propose techniques for use in CAT that primarily aim to efficiently estimate θ, but also satisfy test constraints that allow to examinees to be classified according to specific skills that have been identified for a particular exam.

A trend for greater diagnostic feedback has led to the recent development of numerous models of skills diagnosis, with the aim of assessing mastery of several very specific skills. We consider the novel problem of conducting CAT for estimation of θ, but with the recognition that a breakdown of performance on more specific skills might also be desired.

Weighted *a*-Stratification

CAT is often utilized for testing because it can tailor items to the ability of the examinee to obtain an efficient estimate of an examinee's ability. The preferred estimator of ability is the maximum likelihood estimator $\hat{\theta}^{mle}$, which is the value of θ that maximizes the conditional likelihood function of the responses. Under smoothness conditions on the item response functions (IRF), $\hat{\theta}^{mle}$ is asymptotically $\mathrm{N}(\theta_0, I(\theta_0)^{-1})$ where θ_0 is the true value of θ and $I(\theta_0)$ is the Fisher information at θ_0.

The function $I(\theta_0)$ is a useful measure of the precision with which the *J* items can serve to measure θ, as a function of θ. An efficient but impractical method for conducting CAT is to implement the maximum information criterion (MIC), which is to select the next item as the item that would maximize the item specific information function at the current value of $\hat{\theta}^{mle}$. Though it is efficient, the MIC results in extremely poor utilization of the item bank.

Due to the need to balance item exposure, among other concerns, the success of a CAT algorithm must be measured in several ways. Through simulation, the mean-squared error of θ estimates can be examined for tests of fixed length using the various item selection procedures. However, practical testing concerns require that non-statistical criteria should also be considered in the evaluation. Test construction often must satisfy content-related constraints. Conducting CAT under these constraints has been studied by many researchers. Methods based on the use of linear programming have been developed by van der Linden(2000). These techniques set out to optimize an objective function for accuracy while controlling for several constraints. We focus on less technical and highly practical extensions of the item bank stratification method of Chang and Ying (1999) that addresses the issue of test constraints.

The notion behind *a*-stratification is that high discrimination parameters are most useful late in a test and are not needed as badly in the early stages when there is considerable uncertainty about the θ parameter. The item bank is divided into several strata, usually three or more, so that the distribution of difficulty parameters in these strata remains roughly constant. This is accomplished by ordering the items by their difficulty parameters and taking adjacent groups of *M* (number of strata), and separating them into *M* different bins according to the size of their corresponding discrimination parameters. This results in *M* strata that have nearly balanced difficulty parameters, and nearly ordered distributions of discrimination parameters. Item selection in *a*-stratification involves ascending through these *M* strata, matching $\hat{\theta}$ with a similar difficulty parameter within the current stratum. Through stratification, item exposure balance is achieved.

Test constraints can easily be incorporated by an extension, constraint weighted *a*-stratification. This involves dividing the item bank into strata in the same way as the unweighted version. However, rather than matching the θ estimate with the nearest difficulty parameter within the current stratum, it computes priority indices for the items corresponding to constraints, and among the set of items with difficulty parameters within some distance, $\delta,$ of the θ estimate, it selects the item with the highest priority. Cheng and Chang (2006), proposed a two-stage method for doing this, first addressing lower bounds on item content constraints before turning to upper bounds.

Two-Phase Item Selection

Each flexible content balancing constraint involves a lower bound and an upper bound. Let μ_s denote the number of items to be selected from content area *s*. It must satisfy the following two (in)equalities:

$$l_s \le \mu_s \le u_s \tag{1}$$

and

$$\sum_{s=1}^{S} \mu_s = L \tag{2}$$

where l_s and u_s are the lower bound and upper bound respectively (*s = 1,2 .., S,* and *S* is the total number of content categories) and *L* is the test length.

Two-phase item selection handles the lower bounds in the first phase and the upper bounds in the second phase. The first phase of item selection involves L_1 items, where $L_1 = \sum_{s=1}^{S} l_s$; the second phase involves L_2 items, where $L_2 = L - L_1$. In the first phase, the priority index becomes

$$p_j = \prod_{s=1}^{S} (f_s)^{c_{js}} \tag{3}$$

where C is a $J \times S$ constraint relevancy matrix with entries *c* taking the value 1 if item *j* is relevant to constraint *m*, and 0 otherwise. When constraint *s* reaches its lower bound l_s, f_s becomes 0, then this constraint category will become unused and no more items can be selected from it until the other constraint categories are fulfilled. In this way all the lower bounds will be met at the end of the first phase.

In the second phase, the f_s s are computed by:

$$f_s = \frac{u_s - x_s}{u_s}. \tag{4}$$

Similar to the first phase, when constraint *s* reaches its upper bound u_s, f_s and p_j will be 0, and no more items from this constraint category will be selected as long as there are other categories left unfulfilled.

One-Phase Item selection

To further reduce the complexity and increase the efficiency of *a*-stratification, Cheng et al. (2008) streamlined the process to include only a single phase. In this case the priority score becomes

$$p_j = \prod_{s=1}^{S} (f_{1s} f_{2s})^{c_{js}}, \tag{5}$$

where

$$f_{1s} = \frac{u_s - x_s - 1}{u_s}, \tag{6}$$

and

$$f_{2s} = \frac{(L - l_s) - (t - x_s)}{L - l_s}, \tag{7}$$

where *t* is the number of items currently administered. The function f_{1s} measures the distance from the upper bound. $L - l_s$ is the upper bound of the sum of the number of items that can be selected from other content categories. When $f_{2s} = 0$, the sum of items from other content categories has reached its maximum. Because f_{1s} is decreasing and f_{2s} is increasing with x_s, the index p_j strikes a balance between the two to keep the number of items from constraint category *s* between the lower and upper bounds.

Cognitive Diagnosis

The demand for more formative assessments to be used for in-class diagnostic purposes implies a need for a more fine-grained analysis at the subscale level. Cognitive diagnosis regards the subscales as attributes; by partitioning the latent space into smaller cognitive “attributes”, it can evaluate the student with respect to each attribute. Therefore, students receiving the same total score may have entirely different attribute profiles. Various cognitive diagnosis models

(CDM) have been proposed, and they model the probability of correctly answering an item as a function of an attribute mastery pattern. In this paper, we focus on cognitive diagnosis models that are latent class models structured in part by a *Q* matrix, a matrix that relates items to attributes. Different cognitive diagnosis models are often distinguished by whether attributes enter the response process by a conjunctive rule that requires simultaneous possession of them or by a compensatory function in which some attributes can partly compensate for lack of others.

These models for cognitive diagnosis all require knowledge of which attributes are needed for which items, and differ by how the attributes are utilized. Let α be a *K*-dimensional vector for which α_k indicates whether or not an examinee possesses the k^{th} attribute for $k = 1, 2, ..., K.$ Let $\mathbf{Q}$ be a $J \times K$ matrix referred to as a *Q* matrix (Tatsuoka, 1985), with (j, k) entry q_{jk} denoting if j^{th} item requires the k^{th} attribute. An example of a conjunctive model is the DINA (Deterministic Input, Noisy output And gate) model (Junker & Sijtsma, 2001). The IRF of the DINA model is,

$$P(Y_{ij} = 1 | \alpha_i) = (1 - s_j)^{\eta_{ij}} \, g_j^{1-\eta_{ij}}, \tag{8}$$

where for all *i*, $s_j = P(Y_{ij} = 0 | \eta_{ij} = 1)$ and $g_j = P(Y_{ij} = 0 | \eta_{ij} = 0)$ are the probabilities of slipping and guessing, respectively, for the j^{th} item, and η_{ij} is the ideal response which relates the attribute pattern of an examinee and the j^{th} row of $\mathbf{Q}$,

$$\eta_{ij} = \prod_{k=1}^{K} \alpha_{ik}^{q_{jk}} \tag{9}$$

The variable η_{ij} indicates whether the examinee possesses all the attributes needed for answering the j^{th} item. Computing item parameter estimates for the DINA model can be done with the EM algorithm (Haertel, 1989), or by use of Markov chain monte carlo methods (de la Torre & Douglas, 2004; Tatsuoka, 2002). Templin et al. (2008) discuss how to fit cognitive diagnosis models, including the DINA model as well as the remaining models in this section ,and provide software.

The NIDA (Noisy Input, Deterministic output And gate) model, introduced by Maris (1999), and named in Junker and Sijtsma (2001), considers slips and guesses at the attribute level, where skills are applied in sequence to construct an overall response.

For the NIDA model, η_{ij} indicates whether the i^{th} examinee correctly applied the k^{th} attribute in completing the j^{th} item. Slipping and guessing parameters are indexed by attribute rather than by item, in the case of the DINA model, and are defined by $s_k = P(\eta_{ijk} = 0 | \alpha_{ik} = 1, q_{jk} = 1)$ and $g_k = P(\eta_{ijk} = 1 | \alpha_{ik} = 0, q_{jk} = 1)$. $P(\eta_{ijk} = 1 | q_{jk} = 0)$ is set equal to 1, regardless of the value of α_{ik}.

In the NIDA model an item response Y_{ij} is 1 if all η_{ijk}s are equal to 1, $Y_{ij} = \prod_{k=1}^{K} \eta_{ijk}$. By assuming the η_{ijk}s are independent conditional on the vector α_i, the IRF is

$$P(Y_{ij}=1|\alpha_i,s,g)=\prod_{k=1}^{K}P(\eta_{ijk}=1|\alpha_{ik},s_k,g_k)=\prod_{k=1}^{K}[(1-s_k)^{\alpha_{ik}}g_k^{1-\alpha_{ik}}]^{q_{jk}}. \qquad (10)$$

The NIDA model is extended by a reduced version of the reparameterized unified model, called the reduced RUM. The IRF of the reduced RUM is

$$P(Y_{ij}=1|\alpha_i)=\pi_j^*\prod_{k=1}^{K}r_{jk}^{*(1-\alpha_{ik})q_{jk}}. \qquad (11)$$

The parameter π_j^* is the probability of answering correctly for someone who possesses all of the required attributes, and r_{jk}^* is a parameter between 0 and 1 that represents the penalty for not possessing the k^{th} attribute. Though the parameters of the reduced RUM appear different than the slipping and guessing parameters of the NIDA, they correspond to a reparameterization that keeps the model identifiable when slipping and guessing probabilities are allowed to vary across items.

Compensatory models differ from conjunctive models by allowing examinees to partly compensate for lack of some attributes by possession of others. Such models usually include an additive term in the IRF. Compensatory models can often allow for more equivalent classes in response probabilities, but might not be derived under a specific theory for the response process. The general diagnostic model (GDM) of von Davier (2005) is taken as the representative of a latent class compensatory model. The GDM has IRF

$$P(Y_{ij}=1|\alpha_i)=\frac{\exp[\beta_j+\sum_{k=1}^{K}\gamma_{jk}q_{jk}\alpha_{ik}]}{1+\exp[\beta_j+\sum_{k=1}^{K}\gamma_{jk}q_{jk}\alpha_{ik}]}, \qquad (12)$$

which looks much like a logistic multidimensional IRT model, but has binary latent variables rather than continuous latent variables. In this paper, we will focus on the DINA model as our underlying cognitive diagnosis model.

Higher-Order Latent Trait Models

Due to the two-fold aim in our research, we needed a model which can incorporate both a unidimensional θ and an attribute vector α. This can be accomplished by viewing the attributes as the specific knowledge required for examination performance, and modeling these attributes as arising from a broadly defined latent trait resembling the θ of item response models, so that we can construct the relationship between general aptitude and specific knowledge. This approach was proposed by de la Torre and Douglas (2004) and termed as higher-order latent trait models. They combined the IRT model and diagnostic model by assuming conditional independence of response Y given α , and also assuming that the components of α are independent conditional on θ. The particular relationship between θ and α they consider is logistic regression, given as

$$P(\alpha_k=1|\theta)=\frac{\exp(\lambda_{0k}+\lambda_k\theta)}{1+\exp(\lambda_{0k}+\lambda_k\theta)}. \qquad (13)$$

When conditioning on a latent vector of binary skills α , the conditional distribution of the data follows the cognitive diagnosis model. Therefore, the higher-order model contains a hierarchy, where the cognitive diagnosis model forms Level 1 and the logistic regression model forms Level 2. If the DINA model is used as a Level 1 model, the whole model is called a "higher-order DINA model", and we will adopt it as our underlying model.

Priority Indices with Cognitive Constraints

Five different methods for item selection were studied in simulation. One method was simply random item selection. This can obviously be expected to perform well for balancing item exposure, but it cannot be expected to provide efficient estimation of θ or classification of α . A second method was one-stage *a*-stratification, as described above, using four strata of IRT discrimination parameters. *a*-stratification would be expected to work well in balancing item exposure and estimating θ, but it does not recognize any cognitive diagnosis constraints that should assist in classification of α .

The next three methods utilize cognitive diagnosis information in varying levels. All adapt one-stage weighted *a*-stratification, but involve priority indices that incorporate some knowledge of the cognitive diagnosis model. The items are stratified according to IRT discrimination parameters, and within each stratum item selection involves optimizing a priority function.

Q Control

It is intuitive that how accurate an attribute is measured depends in part on the number of items measuring that attribute, since when more items measure an attribute, more information is accumulated with respect to that attribute. Therefore, one method is to set upper and lower bounds on how many items should measure each attribute, l_s and u_s , and use this information in the priority index along with the IRT information. Let b_j denote the difficulty parameter of the j^{th} item, and q_{jk} denote the *Q* matrix entry for the j^{th} item. The priority index is

$$P_j = \frac{1}{\left|b_j - \hat{\theta}\right|} \prod_{k=1}^{K} f_{jk} I(q_{jk}), \tag{14}$$

where

$$f_{jk} = [\frac{u_k - x_k - 1}{u_k}][\frac{(L - l_k) - (t - x_k - q_{jk})}{L - l_k}] , \tag{15}$$

with *L* denoting total test length, x_k denoting number of items used so far that are relevant to the k^{th} attribute, and *t* denoting the number of items already administered. We refer to this method as *Q control*. Here, $I(q_{jk})$ is an indication function, so f_{jk} will take different forms depending upon the *Q* matrix element. We made a modification here to make sure that every item, no matter how many attributes they measure, will have the same number of multipliers of f_{jk} ; this will enforce all the P_j to be on the same scale for comparison. This method can be expected to balance items over the attributes and over the different skills determined by the *K* attributes, but it does not explicitly discriminate between items with good and bad diagnostic qualities.

Q(1 – *s*)(1– *g*) Control

As an extension to the *Q-control* method, and depending on which cognitive diagnosis model is used, simple modifications can be created that utilize the quality of the items. For example if the DINA model is used, we can define the method *Q(1-s)(1-g) control* by using exactly the same f_{jk}s given above, but altering P_j to

$$P_j = \frac{1}{\left|b_j - \hat{\theta}\right|}(1-s_j)(1-g_j)\prod_{k=1}^{K} f_{jk} I(q_{jk}). \tag{16}$$

Because s_j and g_j denote probabilities of deviation from ideal response patterns, $(1-s_j)(1-g_j)$ serves as a measure of the discrimination or reliability of the j^{th} item, and *Q(1-s)(1-g) control* incorporates this.

KL Information Control

A final method, *KL Information-control*, is more formal and uses indices of reliability for cognitive diagnosis models developed by Henson and Douglas (2005). Let α_w and α_v denote two distinct attribute patterns. The Kullback-Liebler distance between the distribution of the j^{th} item's response, assuming α_w is the correct pattern, is given by,

$$D_{juv} = E_{\alpha_u}\left[\log\left[\frac{P_{\alpha_u}(x_j)}{P_{\alpha_v}(x_j)}\right]\right]. \tag{17}$$

Here, x_j is the response to the j^{th} item; it can take the value of either 1 or 0; $P_{\alpha_u}(x_j)$ is the probability of getting the response x_j given the ability pattern α_u. This item-level information can be summed up to form the test information, and this additive property sets the foundation for the method. The cognitive diagnosis information index (CDI), which is a summary of the item's overall discriminating power, is constructed as follows:

$$\bar{D}_j = \frac{1}{2^K 2^{(K-1)}}\sum_{u \neq v} D_{juv}. \tag{18}$$

It is obtained by simply averaging over all possible combinations of attribute patterns (Henson & Douglas, 2005). However, since this index is only an overall measure of item information, it does not specify the amount of information the item provides for each attribute. Therefore, the overall information needs to be broken down to retrieve the attribute-level information, which indicates the contribution of an item to the correct classification for each attribute. To form the attribute level information index, D_{juv} can be summarized by averaging over all pairs of attribute patterns α_w and α_v that differ only on the k^{th} attribute (Henson., et.al, 2008):

$$d_{jk} = \frac{1}{2^{(K-1)}}\sum_{\Omega_k} D_{juv}, \tag{19}$$

where Ω_k indexes all pairs of attribute patterns that differ only on the k^{th} attribute. The quantity

d_{jk} then summarizes the ability of the j^{th} item to distinguish between examinees who are very similar to one other, but with one possessing the k^{th} attribute and the other not. The corresponding priority index is

$$P_j = \frac{1}{\left|b_j - \hat{\theta}\right|} \sum_{k=1}^{K} (u_k - x_k) d_{jk}. \tag{20}$$

Note that the u_k and x_k are no longer integers, as those in previous methods. In order to make this method feasible, u_k is carefully chosen according to the attribute-level test information. Also note that this Kullback-Leibler index, as opposed to specific indices based on model-specific item parameters, is a general index that can apply to all CDMs. For any given model, one could define this attribute-level information index and incorporate it into priority index for item selection.

Simulation Study

The simulation study, which compared the five methods, involved generating an item bank under a higher-order DINA model for investigating the distinct aims of cognitive diagnosis and unidimensional IRT.

Examinee Generation

In order to simulate data for which a cognitive diagnosis model and a unidimensional IRT model both have applications, we chose the higher-order DINA model of de la Torre and Douglas (2004). In this case, we needed to generate two sets of parameters for the examinees, one was the broad general ability, θ, the other was the fine-grained dichotomous vector, α, and they followed the second level of the model. It is reasonable to believe that the attributes are correlated in the population. Therefore, in order to make the results more general, we considered two cases, one in which attributes were highly correlated and one in which they were not. The correlation of attributes is controlled by the slope λ in the higher-order part of the model. 3,000 higher-order θs were generated from N(0, 1). Then to generate α,

λ=[1.2712, 1.4176, 1.2656, 1.8755, 0.8083] and λ=[0.6306, 0.7083, 0.6328, 0.9377, 0.4042]

were chosen respectively for the high- and low- correlation case. The correlation coefficient between two attributes in the high-correlation case ranged from 0.22 to 0.370, while all below 0.10 in the low-correlation case. We assumed that each attribute was moderately difficult to master. Therefore, the i^{th} examinee's mastery for attribute k was

$$\alpha_k = \begin{cases} 1 & \text{if } P(\alpha_k = 1|\theta) > 0.5 \\ 0 & \text{otherwise} \end{cases}. \tag{21}$$

We checked the entire breakdown of the proportion of examinees with each attribute pattern in both conditions, and the distributions were reasonable; when the correlation was low, the distribution over these 32 patterns was more flat.

Item Bank Construction

The item bank size was predetermined to be 800. The first step was to carefully define a Q

matrix for the entire item bank such that the number of items measuring each attribute was balanced, ranging from 386 to 422, and the number of items measuring different numbers of attributes followed a certain pattern which mimicked the real item bank well, i.e., more than half of the items measured two or three attributes, approximately 20% measured only one attribute, and the rest measured either four or five attributes. We then simulated a bank of DINA model slipping and guessing parameters for 900 items, which followed lognormal distributions, and discarded those items with extremely high or low slipping/guessing parameters, resulting in 800 items.

Then, based on the slipping/guessing parameters and the two sets of simulated true αs, we generated two 1000-by-800 complete response matrices separately according to the DINA model, and retrofitted the matrices with the two-parameter logistic (2PL) model by BILOG calibration. As a result, we obtained *a* (discrimination) and *b* (difficulty) parameters for the same 800-item bank for the two conditions. Table 1 provides the distribution of item parameters for the DINA model (*s* and *g*) and the 2PL model (*a* and *b*) under the two conditions. Meanwhile, we have $\hat{\theta}_0$ from the output, which is the limiting value of $\hat{\theta}$ obtained by estimating θ with responses to the entire bank of items using the 2PL model. Data showed that the correlation between the higher-order true θ_0 and this limiting $\hat{\theta}_0$ was 0.79 and 0.81 in the two cases, respectively, indicating that the $\hat{\theta}_0$ from the somehow misspecified 2PL model converged approximately to the true θ_0 in the limiting sense. Therefore, the higher-order model yielded 2PL-like data and thus the “wrong” 2PL model gave a meaningful quantification of θ, which was used as the “truth” in our method evaluation.

Item Selection Procedures

The item bank was partitioned into four equally large strata such that the *a* parameters were in ascending order while *b* parameter distributions were roughly the same across the strata. Without knowing any information about the examinee, the first item was randomly selected from the item bank, and the rest of the items were selected based on the constrained weighted *a*-stratification method, with different priority indices discussed above. To make things more general, we considered two test lengths, 21 and 41 items. The two values were intentionally chosen to simplify the simulation: in addition to the first item, for the shorter test with 21 items, 5 items were chosen from each stratum, while for longer test, 10 items were selected from each stratum.

Table 1. Summary Statistics for Item Parameters in the High Correlation and Low Correlation Conditions

High Correlation	Minimum	Maximum	Mean	Standard Deviation
Slipping	.001	.299	.0907	.085
Guessing	.001	.498	.143	.141
a	.199	3.626	1.282	.591
b	–2.672	1.978	.009	.899
Low Correlation	Min	Max	Mean	S.D.
Slipping	.001	.299	.0907	.085
Guessing	.001	.498	.143	.141
a	.20	3.63	1.1825	.591
b	–2.672	1.778	.009	.859

Generation of Item Scores in CAT

During the CAT procedure, given each examinee's true attribute pattern, scores for each administered item were generated based on the DINA model. Then, after obtaining the probability of a correct response, $P(Y_{ij}=1|\alpha_i)$, a random $\mathrm{U}(0,1)$ variable u was generated, and the score Y_{ij} was:

$$Y_{ij} = \begin{cases} 1 & \text{If } \mathrm{u} \le \mathrm{P}(Y_{ij}=1|\alpha) \\ 0 & \text{otherwise} \end{cases} \tag{22}$$

Note that once an item is selected, both the $\hat{\theta}$ and $\hat{\alpha}$ were updated for choosing the next item.

Estimation of General Ability and Attribute Patterns

For the attribute pattern estimation, since the slipping/guessing parameters of the DINA model were known, the posterior probability of each attribute pattern was calculated by computing the likelihood for all possible attribute patterns based on the examinee's performance and multiplying by the prior. The posterior mode [i.e., the maximum a posterior (MAP)] method was used to obtain $\hat{\alpha}$.

For θ estimation, at the beginning of the test, when the number of administered item was less than five or the response pattern was all 0 or 1, we used the expected a posteriori (EAP) method (see Bock & Mislevy, 1982) with a standard normal prior and 81 quadrature points; otherwise we used the maximum likelihood estimation (MLE) method.

Evaluation Criteria

Once an item selection method is proposed for CAT, it is important to determine how well the method performed in terms of estimation accuracy and test security. One natural way is to compare its performance to random item selection methods. To check the attribute estimation, the proportion of attributes that were correctly identified was recorded, namely, the *recovery rate*. The recovery rate of attribute mastery was computed marginally for each attribute and for the entire attribute pattern, as well. Concerning θ estimation, note that the θ of the higher-order DINA model was used to generate data, but the somewhat misspecified 2PL model was assumed when fitting item parameters and evaluating $\hat{\theta}$ when conducting CAT. Consequently, when evaluating the performance of θ estimation, we did not refer to the higher-order trait of the DINA model used to generate data, but rather the limiting value of $\hat{\theta}$ obtained by estimating θ with responses to the entire bank of items, i.e., the $\hat{\theta}_0$ from BILOG output. θ estimates were compared to their true values by computing the mean squared error,

$$\text{MSE}\left(\theta\right) = \frac{1}{m}\sum_{i=1}^{m}(\hat{\theta}_i - \hat{\theta}_{0i})^2. \tag{23}$$

In order to check the exposure rate balance, the minimum, maximum, mean, standard deviation and several percentiles of the exposure rate distribution were calculated in addition to the chi-square index,

$$\chi^2 = \sum_{j=1}^{N}(er_j - e\bar{r}_j)^2 / e\bar{r}_j, \tag{24}$$

where er_j is the exposure rate of item *j*, and $e\bar{r}_j = L / N$ is the desirable uniform rate for all items. This chi-square index captures the discrepancy between the observed and the ideal item exposure rates, therefore quantifying the efficiency of item bank usage; the smaller the value, the more efficiently the item bank was used.

Results

Estimation results are given in Table 2. For a 41-item test with high correlation, recovery rates for the attributes were high for all of the methods, but were highest for KL information and *Q*(1−*s*)(1−*g*) methods, particularly when looking at the whole attribute patterns. These methods also performed better in θ estimation. That is partly because both of these methods favor items with low slipping and guessing parameters, and when fitting the 2PL model, such good diagnostic items also result in high discrimination parameters, so that more information for θ is obtained, even when stratifying the discrimination parameters.

Table 3 displays item exposure results. Due to the fact that all experimental methods employed *a*-stratification, item exposure balance was under quite reasonable control, though random item selection optimized balance. The results for the shorter test length of 21 items show a more pronounced advantage for KL information, and item exposure results were still quite comparable for the methods, except for random item selection which performed very poorly at classification and estimation, but balanced exposure almost perfectly.

Table 2. Recovery Rates by Attribute, and MSE(θ) for Two Correlation Conditions and Two Test Lengths

	Recovery Rate						
Method	Attribute 1	Attribute 2	Attribute 3	Attribute 4	Attribute 5	Pattern	MSE(θ)
High Correlation, 41-Item Test							
Stratified only	0.995	0.9997	0.9833	0.9663	0.9667	0.919	0.14
Q control	0.9973	0.9957	0.988	0.9897	0.9757	0.9517	0.0975
$Q\ (1-s)(1-g)$	0.9987	0.9973	0.9903	0.9913	0.978	0.9583	0.0935
Information	0.9997	0.9973	0.9967	0.996	0.9933	0.984	0.067
Random	0.9353	0.9473	0.972	0.9577	0.9423	0.8273	0.195
High Correlation, 21-Item Test							
Stratified only	0.975	0.9707	0.9633	0.969	0.9387	0.8487	0.144
Q control	0.974	0.9733	0.9677	0.971	0.9377	0.8543	0.132
$Q(1-s)(1-g)$	0.9787	0.9783	0.971	0.9743	0.9387	0.8647	0.25
Information	0.9907	0.979	0.979	0.9883	0.966	0.9177	0.105
Random	0.875	0.919	0.916	0.9007	0.8887	0.6953	0.649
Low Correlation, 41-Item Test							
Stratified only	0.9933	0.9997	0.9793	0.9197	0.9427	0.8530	0.293
Q control	0.9867	0.9897	0.9777	0.955	0.939	0.8693	0.168
$Q(1-s)(1-g)$	0.9923	0.994	0.9757	0.953	0.942	0.878	0.151
Information	0.9990	0.9967	0.9960	0.9953	0.9883	0.9763	0.125
Random	0.9403	0.9503	0.9673	0.958	0.9433	0.8273	0.312
Low Correlation, 21-Item Test							
Stratified only	0.9690	0.9863	0.9560	0.8777	0.9087	0.7540	0.274
Q control	0.9480	0.9527	0.9490	0.9243	0.9113	0.7723	0.228
$Q(1-s)(1-g)$	0.9580	0.9640	0.9560	0.9247	0.9140	0.7830	0.235
Information	0.9877	0.9733	0.9700	0.9803	0.9507	0.8867	0.157
Random	0.8593	0.8987	0.9	0.888	0.885	0.663	0.379

Table 3. Exposure Rate Distribution for Two Correlation Conditions and Two Test Lengths

Method	Min	25%	50%	75%	90%	Max	Mean	S.D.	χ^2
High Correlation, 41-Item Test									
Stratified only	.0003	.0153	.0340	.0695	.1176	.2317	.05125	.0501	39.18
Q control	.0000	.0080	.0316	.0642	.1159	.5030	.05125	.0669	69.81
$Q(1 - s)(1 - g)$	.0000	.0097	.0326	.0620	.1152	.5040	.05125	.0660	67.96
KL Information	.0007	.0134	.0328	.0662	.1146	.4853	.05125	.0575	51.72
High Correlation, 21-Item Test									
Stratified only	.0013	.0100	.0183	.0310	.0479	.2197	.0262	.0302	27.841
Q control	.0007	.0090	.018	.0303	.0536	.2933	.0262	.0307	28.72
$Q(1 - s)(1 - g)$	.0003	.0097	.0186	.0306	.0543	.2107	.0262	.0287	25.22
KL Information	.0010	.0093	.0181	.0310	.0532	.2143	.0262	.0287	24.95
Random	.0183	.0243	.0263	.0280	.0300	.0367	.0262	.0027	0.253
Low Correlation, 41-Item Test									
Stratified only	.0007	.0176	.0363	.0743	.1253	.1853	.05125	.0431	29.075
Q control	.0007	.0150	.0343	.0685	.1259	.5297	.05125	.0543	46.134
$Q(1 - s)(1 - g)$	.0000	.0166	.0360	.0680	.1183	.4510	.05125	.0524	42.863
KL Information	.0003	.0140	.0345	.0673	.1143	.5867	.05125	.0593	54.969
Random	.0397	.0483	.0513	.0540	.0566	.0643	.05125	.0041	0.235
Low Correlation, 21-Item Test									
Stratified only	.0023	.0106	.0190	.0320	.0490	.1513	.0262	.0247	18.604
Q control	.0007	.0093	.0183	.0319	.0536	.4113	.0262	.0298	27.12
$Q(1 - s)(1 - g)$	.0007	.0106	.0181	.0325	.0543	.3073	.0262	.0271	22.38
KL Information	.0010	.0096	.0176	.0315	.0512	.4287	.0262	.0330	33.270
Random	.0183	.0243	.0263	.0280	.0300	.0367	.0262	.0027	0.235

For the 41-item test and low correlation case, Table 2 shows that KL information was even more valuable. This was a more challenging case because of the near independence of the attributes. As a result, classifying the whole vector was more difficult because less information can be compiled about the joint distribution. The $Q(1-s)(1-g)$ method came in a distant second. The trend remained in the 21-item test. To summarize, all four methods based on *a*-stratification yielded similar item exposure properties, but classification of α and estimation of θ was best when using KL information. In all cases, the $Q(1-s)(1-g)$ method performed second best, but did much worse when the attributes were not tightly associated.

Discussion

When a single examination is used for a standard unidimensional score and for skills diganosis, one has to consider how these distinct aims can be addressed at once. The higher-order DINA model illustrates how data can appear to follow an IRT model, when in fact there is local dependence that can only be eliminated by conditioning on a vector of binary skills. The standard θ still has a useful interpretation as a general and broad level of knowledge or ability and fitting logistic IRT models can be quite a good approximation when conducting CAT and assigning a $\hat{\theta}$. CAT can be conducted in a manner that assigns this critical score, but can also diagnosis presence or absence of each component of an attribute vector. Therefore, it can afford practitioners with the opportunity to rank order the examinee while diagnose them in a way that leads to tailored remediation.

Several priority scores were considered for modification of weighted *a*-stratification. The most efficient method directly utilized Kullback-Liebler information, though a simple priority score based on slipping and guessing parameters performed nearly as well. Using only *Q*-matrix information was not as effective, but can be used to achieve balance across the attributes. These conclusions are most apparent when inspecting results for the shorter exam, and for classification of the entire attribute vector.

CAT can be expanded to perform cognitive diagnosis, even when fitting ordinary IRT models, by calibrating the item bank with a cognitive diagnosis model such as the DINA, and considering both models in the item selection procedure. Unidimensional IRT provides a simplification of test data by assuming a single latent trait. Cognitive diagnosis models rectify this to some degree by recognizing several dimensions, but also make the assumption that latent variables are binary. By utilizing both models in a testing program, we can reach a useful compromise that achieves an ordering of the broad knowledge of examinees, but offers a finer breakdown that can be used for reporting, and provides valuable information for remediation.

References

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

Chang, H. H. and Ying, Z. (1999). *a*-stratified multistage computerized adaptive testing. *Applied Psychological Measurement*, *23*, 211-222.

Cheng, Y., & Chang, H. H. (2007). Two-phase item selection procedure for flexible content balancing in CAT. *Applied Psychological Measurement, 31*, 467-482.

Cheng, Y., Chang, H., Douglas, J., & Guo, F. (2009). Constraint-weighted *a*-stratification for computerized adaptive testing with nonstatistical constraints—balancing measurement efficiency and exposure control. *Educational and Psychological Measurement*, 69, 35-49.

van der Linden, W. J. (2000). Constrained adaptive testing with shadow tests. In van der Linden & Glas (Eds.), Computerized adaptive testing: Theory and practice (pp. 27-52). Boston:Kluwer-Nijhoff.

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

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

Henson, R. Roussos, L. Douglas, J.& He, X. (2008) Cognitive diagnostic attribute-level discrimination indices. *Applied Psychological Measurement, 32,* 275-288

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.

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

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

Tatsuoka, C. (2002). Data-analytic methods for latent partially ordered classification models. *Applied Statistics* (JRSS-C), *51,* 337-350.

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

Templin, J. L., Henson, R. A., Templin, S. E., & Roussos, L. (2008). Robustness of hierarchical modeling of skill association in cognitive diagnosis models. *Applied Psychological Measurement, 32,* 559-574.

von Davier, M. (2005). A general diagnostic model applied to language testing data. *Educational Testing Service, Research Report*, RR-05-16.