Behav Res (2012) 44:95–109
DOI 10.3758/s13428-011-0143-3

Combining CAT with cognitive diagnosis: A weighted item selection approach

Chun Wang · Hua-Hua Chang · Jeffery Douglas

Published online: 19 August 2011

Abstract Computerized adaptive testing (CAT) was originally proposed to measure θ, usually a latent trait, with greater precision by sequentially selecting items according to the student's responses to previously administered items. Although the application of CAT is promising for many educational testing programs, most of the current CAT systems were not designed to provide diagnostic information. This article discusses item selection strategies specifically tailored for cognitive diagnostic tests. Our goal is to identify an effective item selection algorithm that not only estimates θ efficiently, but also classifies the student's knowledge status α accurately. A single-stage item selection method with a dual purpose will be introduced. The main idea is to treat diagnostic criteria as constraints: Using the maximum priority index method to meet these constraints, the CAT system is able to generate cognitive diagnostic feedback in a fairly straightforward fashion. Different priority functions are proposed. Some of them are based on certain information measures, such as Kullback–Leibler information, and others utilize only the information provided by the *Q*-matrix. An extensive simulation study is conducted, and the results indicate that the information-based method not only yields higher classification rates for cognitive diagnosis, but also achieves more accurate θ estimation. Other constraint controls, such as item exposure rates, are also considered for all the competing methods.

C. Wang (✉) · H.-H. Chang
Psychology, University of Illinois,
Champaign, IL, USA
e-mail: cwang49@illinois.edu

J. Douglas
Statistics, University of Illinois,
Champaign, IL, USA

Keywords CAT · Cognitive diagnosis · Constraint-weighted item selection · *a*-stratification

Combining a test with cognitive diagnosis is a matter of interest not only to students and teachers, but also to cognitive psychologists who investigate the cognitive processes of problem solving (Greeno, 1980) and psychiatrists who needs to pinpoint the specific psychological disorders of patients (Templin & Henson, 2006). In educational testing, for instance, formative feedback will help improve learning as a diagnostic or screening mechanism (Cheng, 2009). In the past few decades or even earlier, several research studies have been devoted to statistical methodologies for the purpose of diagnosing students' cognitive errors. Yamamoto (1989) used the latent-class model; Mislevy (1995) used the inference network; DiBello, Stout, and Roussos (1995) proposed the unified model; Haertel (1989) and Junker and Sijtsma (2001) invented the "deterministic input, noisy 'and' gate" (DINA) model; Tatsuoka (1983, 1990) and Tatsuoka and Tatsuoka (1987, 1997) published a series of articles looking into the statistical pattern classification approach and its application in cognitive diagnosis under the name of "rule-space methodology." The common ground for these cognitive diagnostic approaches is that they infer the unobservable cognitive processes from logical interrelationships among cognitive tasks and subtasks (also termed *attributes*) involved in the test problems. Depending upon the different attributes (say, K in all) required for a certain test, each person is given an attribute mastery pattern at the end of the test, and the pattern consists of K binary elements representing unobservable attributes that indicate the strength and weakness of each examinee.

Although cognitive diagnosis has seen much progress in developing the state-of-the-art technology (Rupp & Templin, 2008), its application still faces a practical challenge. That is, most of the current large scale tests are built upon item response theory (IRT). IRT explains the individual's behavior by defining latent traits (usually denoted by θ), estimating the individual's standing on each of these traits, and then using the numerical values obtained to predict performance in relevant situations (Lord & Novick, 1968). The traits are similar to the true score in classical test theory and can be used to rank examinees or make pass/fail decisions in mastery tests. However, the traits are different from mastery status given by cognitive diagnosis. Thus, it is natural to doubt whether and how to extract cognitive information from the current IRT-based tests. In fact, Tatsuoka's rule-space methodology provides a statistically sound and practically valuable means for tackling this challenge.

In rule space methodology, first a set of knowledge states (K. K. Tatsuoka, 1991; also called *attribute mastery patterns*, denoted by α) are identified from a test—say, there are 2^K different elements in the set. (Note that in practice, the total number of possible patterns might be less than 2^K.) Then an incidence matrix Q of size $K \times n$(n denotes the number of items) is constructed, which defines the item characteristics with respect to the underlying cognitive processes involved in the item. Next, Boolean descriptive functions (BDFs) are employed to systematically relate the knowledge states to observable item response patterns, which are called *ideal item-score patterns*. Therefore, BDF serves as a bridge to connect two spaces, a set of attribute patterns and a set of ideal item-score patterns. Once both spaces are determined via a Q-matrix and a BDF, a student's observed response pattern is compared with the ideal item-score pattern. Lastly, one ideal pattern that is closest to the student's actual pattern is determined, and the underlying knowledge state with respect to this ideal pattern is recognized as the student's estimated pattern. One of the key components of this methodology is how to define *closeness*; to this end, a rule space, also known as a classification space, is constructed. It is a two-dimensional space with coordinates (θ, ζ), θ is the latent trait defined by IRT, and ζ is a function of the item response function reflecting the distinctiveness of various response patterns. In this sense, ζ measures the information that cannot be extracted from the total score. Now, every observed response pattern and ideal item-score pattern can be mapped as points in this two-dimensional rule space, and the Mahalanobis distance between each of the two points is a measure of closeness. In this regard, this rule space methodology extracts diagnostic information from the IRT-based tests, and each examinee will obtain both θ and knowledge state α. To validate this approach, Tatsuoka and Tatsuoka (1997) conducted an experiment with three steps: pretest, remediation, and posttest. They found that "knowing students' knowledge state prior to remediation is very helpful and that the rule-space method can effectively diagnose students' knowledge states and point out ways for remediating their errors quickly with minimum effort."

In fact, it is common that tests that were originally designed to produce reliable scores for ranking examinees are expected to provide information that is useful for remediation or other purposes as well. The last few decades have seen an increasing trend in reporting subscores for the sake of diagnostic purposes—for example, ACT reports English, mathematics, reading, and science scores, and SAT reports critical reading, mathematics, and writing scores. The subscores are often derived from subsections of tests. There are also some tests reporting both an overall score and subscores, such that the purposes of ordering examinees and providing diagnostic feedback are met simultaneously—for example, Measures of Academic Progress (MAP), a state-aligned CAT targeting K–12 applications developed by the Northwest Evaluation Association (see www.nwea.org). Besides providing a general achievement score, MAP has the capability of calculating a "goal performance score" based on subscales, which is analogous to the attribute vector α in cognitive diagnosis. The subscale scores provided by MAP, however, are obtained simply by summing up the responses to the items measuring each skill. Indeed, this subscore approach is an approximate measure of a student's mastery level per each particular skill; however, using a model-based approach, such as the rule space model, provides a level of control in scaling, linking, and item banking that is unavailable with simpler methods. In this article, instead of adopting the rule space method, we will use another widely used model—namely, the DINA model (Haertel, 1989; Junker & Sijtsma, 2001)—for diagnostic purposes. There are two differences between the DINA model and rule space approach. First, the DINA model uses a power function (please refer to Eq. 7 for details) to relate the knowledge state to the ideal item score pattern, whereas rule space models use a BDF. Second, the DINA model incorporates a stochastic part in the item response function, which naturally explains the "distance" between the observed response pattern and ideal pattern in rule space.

Another innovation of this article is that instead of using a paper-and-pencil test, we advocate using computerized adaptive testing (CAT), which is more efficient and more secure (Chang, 2004). However, it is more challenging, because the current item selection algorithms for CAT aim at estimating either the continuous latent trait θ only (van der Linden & Chang, 2003; Yi, Zhang, & Chang, 2008) or the

cognitive knowledge state α only (Cheng, 2009; Xu, Chang, & Douglas, 2003). In this article, 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 α might also be desired. Kingsbury (2009) once dubbed adaptive tests geared toward cognitive diagnosis as "idiosyncratic computerized adaptive testing" (ICAT), and ICAT has found promising applications in providing teachers with information for targeting instruction for students with unique knowledge patterns. In this article, we propose several techniques for use in CAT whose aim is primarily to efficiently estimate θ but also to satisfy test constraints that allow one to classify examinees according to specific skills that have been identified for a particular exam. These techniques can be used as complements to McGlohen and Chang's (2008) approach. The remainder of the article is organized as follows. The next section will begin with the introduction of the maximum priority index (MPI) method, which is the primary method in this study, and cognitive diagnosis and various models will be described in more detail. Then the major priority indices that incorporate cognitive information will be proposed, followed by simulation studies. Finally, the results of various methods will be provided, and the last section will discuss the generalization of the techniques and give concluding remarks.

Maximum priority method for CAT

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. One commonly used estimator of ability is the maximum likelihood estimator $\widehat{\theta}^{mle}$. Under smoothness conditions on the item response functions, $\widehat{\theta}^{mle}$ is asymptotically normal $N\left(\theta_0, I(\theta_0)^{-1}\right)$ where θ_0 is the true value of θ and $I(\theta_0)$ is the Fisher information at θ_0.

An efficient method for conducting CAT on the basis of $I(\theta_0)$ is to implement the maximum information criterion (MIC), which is to select the next item according to the one that would maximize the item-specific information function at the current value of $\widehat{\theta}^{mle}$. Although it is efficient, the MIC always results in extremely poor utilization of the item bank.

Besides the need to balance item exposure, among other concerns, the success of a CAT algorithm must be measured in several ways. Practical testing concerns require that nonstatistical criteria should also be considered in the evaluation, such as content balancing and answer key balancing. 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, but it is computationally intensive and sometimes infeasible (Timminga, 1998). Another branch is the heuristic methods, such as the weighted deviation modeling (WDM) method (Stocking & Swanson, 1993), in which the constraints are viewed only as desirable properties that do not have to be met strictly. WDM shares the advantage of heuristic methods; it is fast, and the algorithm always offers a solution. However, the drawback of WDM is that one has to go through a time-consuming process to adjust the weight (Leung, Chang, & Hau, 2005). We focus on a less technical and highly practical constraint management method—the MPI method (Cheng & Chang, 2009). It also belongs to heuristic methods, but, as compared with the WDM, it leads to fewer constraint violations and better exposure control while maintaining similar measurement precision.

In this study, cognitive information is treated as the constraints that the test intends to meet, and these constraints always come in the form of lower and upper bounds of the number of items from each attribute category or the information accumulated at each attribute, known as the flexible constraints. Cheng, Chang, and Yi (2007) proposed a two-stage method for doing this, first addressing the lower bounds of the constraints before turning to the upper bounds. To further reduce the complexity and increase the efficiency of managing various constraints simultaneously, Cheng, Guo, Chang, and Douglas (2009) streamlined the process to condense MPI into only a single phase, termed *one-phase item selection*. This MPI can be multiplied by various types of information measures to form the item selection criterion.

MPI within the one-phase item selection framework

Let μ_s denote the number of items to be selected from constraint 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 ($s = 1,2 \ldots S$, and S is the total number of content areas). L is the test length, and J is the total number of items in the item bank. Denote a constraint relevancy matrix by $\boldsymbol{C}$. $\boldsymbol{C}$ is a $J \times S$ matrix defined over the item bank, with entries c taking the value 1 if item j is relevant to constraint s and 0 otherwise.

The one-phase item selection uses the priority index for item j, p_j, defined as

$$p_j = \prod_{s=1}^{S} (f_{1s} \cdot f_{2s})^{c_{js}}, \tag{3}$$

where

$$f_{1s}^{(t+1)} = \frac{u_s - x_s^{(t)} - 1}{u_s}, \quad (4)$$

and

$$f_{2s}^{(t+1)} = \frac{(L - l_s) - \left(t - x_s^{(t)}\right)}{L - l_s}, \quad (5)$$

and "·" indicates multiplication. Here, c_{js} is the $(j, s)^{th}$ element of $\boldsymbol{C}$; t is the number of items that have already been selected, and $x_s^{(t)}$ is the number of items that have been selected associated with constraint s at stage t (i.e., the current test length is t); thus, the function $f_{1s}^{(t+1)}$ measures the predicted distance from the upper bound with respect to constraint s at stage (t + 1). $L–l_s$ is the upper bound of the sum of number of items that can be selected from other content areas, so $f_{2s}^{(t+1)}$ measures the expected distance from the upper bound with respect to other content areas except s. When $f_{2s}^{(t+1)} = 0$, the sum of items from other content categories has reached its maximum. Because $f_{1s}^{(t+1)}$ decreases and $f_{2s}^{(t+1)}$ increases with x_s, the index p_j strikes a balance between the two to keep the number of items from constraint category s within the lower and upper bounds. To demonstrate this one-phase method, let us consider the following artificial example. For simplicity, suppose that there are two constraints in the test, that the test length $L = 10$, and that the lower and upper bounds for the two constraints are $l_1 = 3, u_1 = 6; l_2 = 4, u_2 = 7$, respectively. Assume that there are two candidate items in the item bank; item 1 measures constraint 1, and item 2 measures constraint 2. Assume that $t = 6$, $x_1 = 2$ and $x_2 = 4$. Also assume that both items provide equal amount of Fisher information. Intuitively, item 1 should be selected because it measures the first content area that has not been measured with enough items. The priority index in Eq. 3 will also select item 1 according to the following calculation (see Table 1).

Constraint weighted a-stratification design

In the large-scale application of CAT, many issues should be considered, such as reliability and test security. One of the most commonly recognized drawbacks to the maximum-information-based item selection method is that, no matter how large the item pool size is, only a small fraction of the items tend to be used and this poses a security risk (Wainer et al., 1990). In order to deal with these issues, an *a*-stratified item selection method (AST) is proposed. With proper stratification and blocking techniques, AST can equalize item exposure distribution and, hence, can yield higher test security and maintain estimation efficiency (e.g., Chang & Ying, 1999; Hau & Chang, 2001).

Table 1 A demo of the one-phase method

	f_{11}	f_{12}	f_{21}	f_{22}	p_j
Item 1	1/2	3/7	N/A[1]	N/A	3/14
Item 2	N/A	N/A	2/7	2/3	4/21

[1] N/A here means we do not need to calculate the value in the corresponding cell because $c_{12} = c_{21} = 0$

The rationale behind *a*-stratification is that high discrimination parameters are most useful late in an exam and are not needed as badly in the early stages when there is considerable uncertainty about the ability 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), 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 $\widehat{\theta}$ with a similar difficulty parameter within the current stratum. Through stratification item exposure, balance is achieved.

Test constraints can easily be incorporated into the *a*-stratification. This involves dividing the item bank into strata in the same way as in the unweighted version. However, rather than matching the ability estimate with the nearest difficulty parameter within the current stratum, the item selection index becomes: $T_j = \frac{1}{\left|b_j - \widehat{\theta}\right|} p_j$ where p_j is the priority index of the j^{th} item obtained from Eq. 3. We need to select the next item such that T_j is maximized.

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 (CDMs) have been proposed, and they model the probability of correctly answering an item as a function of an attribute mastery pattern. In fact, these CDMs are special cases of finite mixture models developed specifically for the purpose of identifying the presence or absence of fine-

grained skills. They are built upon the following three assumptions: (1) Each individual can be assigned to one of the latent classes defined by attribute profiles; (2) the behavior of an individual responding to items is represented by the latent class responding to the items; and (3) the attributes that make up the latent classes are correlated with one another.

In this article, we focus on CDMs that are latent class models structured in part by a Q-matrix (K. K. Tatsuoka, 1985), a matrix that relates items to attributes. Different CDMs 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 the lack of others. Let α be a K-dimensional vector for which α_k indicates whether or not a subject possesses the k^{th} attribute for $k = 1, 2, ..., K$.Let Q be a $J \times K$ matrix, with $(j, k)^{th}$ entry q_{jk} denoting whether the j^{th} item requires the k^{th} attribute. An example of a conjunctive model is the DINA model (Junker & Sijtsma, 2001). The item response function of the DINA model is

$$P\left(Y_{ij} = 1|\alpha_i\right) = \left(1 - s_j\right)^{\eta_{ij}} g_j^{1-\eta_{ij}}, \tag{6}$$

where α_i is a K-dimensional vector denoting the mastery profile of examinee i and Y_{ij} is the response of examinee i to the item j; $s_j = P\left(Y_{ij} = 0\middle|\eta_{ij} = 1\right)$ and $g_j = P\left(Y_{ij} = 1\middle|\eta_{ij} = 0\right)$are the slipping and guessing parameters, respectively, for the j^{th} item; and η_{ij} is the ideal response that relates the attribute pattern of a subject and the j^{th} row of the Q-matrix as follows:

$$\eta_{ij} = \prod_{k=1}^{K} \alpha_{ik}^{q_{jk}}. \tag{7}$$

The variable η_{ij} indicates whether the subject possesses all the attributes needed for answering the j^{th} item. Computing item parameter estimates for the DINA model can be done with an EM algorithm (de la Torre, 2009; Haertel, 1989) or by use of Markov chain Monte Carlo (de la Torre & Douglas, 2004; C. Tatsuoka, 2002). Templin, Henson, and Douglas (2008) and Templin, Henson, Templin, and Roussos (2008) discuss how to fit CDMs, including the DINA model, as well as the remaining models in this section, and provide software.

The DINA model is widely used in practice (de la Torre & Douglas, 2004); however, when the data set is not large enough to support estimation of such a large number of parameters as would go with the DINA model, a simpler alternative, introduced by Maris (1999) and named in Junker and Sijtsma (2001) the *noisy input, deterministic output "and" gate* (NIDA) model, can be used. The NIDA model considers slips and guesses at the attribute level, and they are the same for every item for which $q_{jk} = 1$. Specifically, the slipping and guessing parameters are defined by $s_k = P\left(\eta_{ijk} = 0|\alpha_{ik} = 1, q_{jk} = 1\right)$ and $g_k = P\left(\eta_{ijk} = 1|\alpha_{ik} = 0, q_{jk} = 1\right)$. $P\left(\eta_{ijk} = 1\middle|q_{jk} = 0\right)$ 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 that the η_{ijk}s are independent conditional on the vector α_i, the IRF is

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

Since the parameters vary only across attributes, this model has severe restrictions—in particular, when the aim is to evaluate the effectiveness of the items. The NIDA model is extended by a reduced version of the reparameterized unified model, called the *reduced RUM* (DiBello et al., 1995).

Compensatory models differ from conjunctive models by allowing subjects to partly compensate for a lack of some attributes by possession of others. Such models usually include an additive term in the item response function. The general diagnostic model (GDM) of von Davier (2005) is taken as the representative of a latent class compensatory model. The GDM has item response function

$$P\left(Y_{ij} = 1|\alpha_i\right) = \frac{\exp\left[\beta_j + \sum_{k=1}^{K} \gamma_{jk} q_{jk} \alpha_{ik}\right]}{1 + \exp\left[\beta_j + \sum_{k=1}^{K} \gamma_{jk} q_{jk} \alpha_{ik}\right]}. \tag{8}$$

In this model, q_{jk} and α_{ik} have the same meanings as in conjunctive models. The parameters β_j and γ_{jk} may be interpreted as threshold and slope parameters, respectively (von Davier, 2005). The GDM has an item response function quite similar to that in the logistic multidimensional IRT model but has discrete latent variables rather than continuous latent variables.

Priority indices with cognitive constraints

For items calibrated with the latent class CDM, the cognitive information contained in them can be partly quantified by a Q-matrix, CDM item parameters, or a statistical combination of them. On the basis of different levels of cognitive information, priority indices will take different forms, as defined in detail in the following section.

Q-control

Intuitively, if more items measure an attribute, more information will be accumulated with respect to this attribute, and the more reliably this attribute will be measured. Thus, how accurately an attribute is measured depends in part on the number of items measuring that attribute. Therefore, one method is to set upper and lower bounds on how many items should measure each attribute, denoted as l_s and u_s, and use them in the priority index along with the IRT information.

The priority index defined in Eq. 3 or 5 is intended to control the nonstatistical constraints such as content balance and answer key balance. Since every item can be assigned to one and only one category of each constraint, the number of constraints is the same across the items. Therefore, the P_j will have an equal number of multipliers for every item, which lays the foundation for a fair comparison. However, in our case, the number of attributes measured by items varies. If we simply adopt the original priority index, the problem of noncomparability will arise; that is, the items measuring more attributes will always have a smaller P_j as compared with the items measuring fewer attributes. To solve this problem, we reconstruct the priority index in the following way to make them comparable across the items.

Let q_{jk} denote the Q-matrix entry for the j^{th} item. The priority index is defined as

$$P_j = \prod_{k=1}^{K} \left[\frac{u_k - x_k - q_{jk}}{u_k}\right] \left[\frac{(L - l_k) - \left(t - x_k - q_{jk}\right)}{L - l_k}\right], \quad (9)$$

with all symbols denoting the same meanings as in Eq. 5. We refer to this method as *Q-control*. As compared with the priority index in Eqs. 3–5, the P_j here in Eq. 9 takes a slightly different form: Instead of putting q_{jk} in the exponent, we put it in each multiplier of P_j. We make such a modification here in order to make sure that all items, no matter how many attributes they measure, will have the same number of multipliers in P_j; this will force all the P_js to have the same metric to facilitate comparison. Combining this priority index with the *a*-stratification method, we name it *StraQ*, and with the maximum information method, we have *MIQ*. In MIQ, the item selection criterion becomes $T_j = FI_j\left(\widehat{\theta}\right)P_j$, and $FI_j\left(\widehat{\theta}\right)$ is the item Fisher information.

This *Q*-control priority index can be expected to balance items over the different skills determined by the *K* attributes but does not explicitly discriminate between items with good and bad diagnostic qualities.

Q discrimination-control

As an extension of the *Q-control* method, simple modifications can be created that utilize the quality of the items. Depending on which CDM is used, the priority index will take slightly different forms. For example, if the DINA model is used, we can define the *Q discrimination-control* method by using exactly the same elements as those given in Eq. 9, but altering P_j to

$$P_j = \left(1 - s_j\right)\left(1 - g_j\right) \times \prod_{k=1}^{K} \left[\frac{u_k - x_k - q_{jk}}{u_k}\right] \left[\frac{(L - l_k) - \left(t - x_k - q_{jk}\right)}{L - l_k}\right], \quad (10)$$

where s_j and g_j denote probabilities of deviation from ideal response patterns. Intuitively speaking, in the DINA model, ($1–s_j$) is the power of differentiating the persons who get the j^{th} item correctly ("winners") from those who do not ("losers") among the "masters" of item j, while ($1–g_j$) gives the power of discriminating the winner from losers among the nonmasters. So $\left(1 - s_j\right)\left(1 - g_j\right)$ serves as a measure of discrimination or reliability of the j^{th} item, and Q discrimination-control incorporates this. In fact, the correlation between $\left(1 - s_j\right)\left(1 - g_j\right)$ and the item cognitive diagnosis information index (CDI; Henson & Douglas, 2005) is as high as .93, according to our calculation. We will use "StraQD" to represent this index with the *a*-stratification method, and "MIQD" to denote the maximum information with the Q discrimination index.

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_u and α_v denote two distinct attribute patterns. The Kullback–Liebler distance between the distribution of the j^{th} item's response, assuming that α_u is the correct pattern, is given by

$$D_{juv} = E_{\alpha_u}\left[\log\left[\frac{P_{\alpha_u}\left(x_j\right)}{P_{\alpha_v}\left(x_j\right)}\right]\right]. \quad (11)$$

Here, x_j is the response to the j^{th} item; $P_{\alpha_u}\left(x_j\right)$ is the probability of getting the response x_j given the ability pattern α_u. By the definition of KL distance, D_{juv} is the discrimination power of item j to differentiate the attribute pattern α_u from α_v. The CDI, which is a summary of the item's overall discriminating power over all possible pairs of attribute pattern (Henson & Douglas, 2005), is constructed as follows:

$$CDI_j = \frac{1}{2^K 2^{(K-1)}} \sum_{u \neq v} D_{juv}. \quad (12)$$

However, since this index is only an overall measure of item information, it ignores which attributes are required by

which items that are provided in the Q-matrix. To break down CDI into an 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, Roussos, Douglas, & He, 2008). Denoting the set of these pairs as Ω_k, we have

$$d_{jk} = \frac{1}{2^{(K-1)}} \sum_{\Omega_k} D_{juv}, \tag{13}$$

where d_{jk} summarizes the ability of the j^{th} item to distinguish between examinees who have possessed the k^{th} attribute and others who have not. The attribute-level information index indicates the contribution of an item to the correct classification for each attribute. The corresponding priority index is

$$P_j = \sum_{k=1}^{K} (u_k - x_k) d_{jk}. \tag{14}$$

Note that this index is entirely different from the previous two. The primary reason is that information is a continuous variable and our intention is to maximize the information accumulated at each attribute while also balancing the amount of information across the attributes. In this case, u_k and x_k are the upper bound and the accumulated information for the k^{th} attribute, respectively. The lower bound is not needed here, and $(u_k–x_k)$ serves as a weight for attribute-level information, indicating how important the information for the k^{th} attribute is. When the weight is large, this means that the information accumulated for the k^{th} attribute is far from enough, and those items that carry large information for attribute k are preferred. This priority index has a built-in "minimax" mechanism—that is, to minimize the maximum distance from the upper bound—and as a result, the attribute information lagging behind will be given more weight. Equipping the *a*-stratification method with this information-based index forms stratified information control, denoted as "StraInfor"; and the maximum information based method is named "MIinfor."

Also note that this Kullback–Leibler index, as opposed to 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 the priority index for item selection.

Simulation study

The simulation study involves selecting an appropriate model for investigating the distinct aims of cognitive diagnosis and unidimensional IRT. This will be covered next, followed by the simulation details. The results will be given at the end, which concern accuracy of θ estimation and classification accuracy for binary attributes, as well as a breakdown of item exposure rate pattern under one test condition as a representative.

Higher-order DINA model

Due to the twofold aim and correlation between the unidimensional θ and attribute vector α, it is better to find a single underlying model that can incorporate both θ and α. This can be accomplished by viewing the attributes as the specific knowledge required for examinee performance and modeling these attributes as arising from a broadly defined latent trait resembling the θ of the 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 further developed for hierarchical models with various structures between θ and α (Templin, et al., 2008a, b). The higher-order latent trait models combine the IRT model and diagnostic model by assuming conditional independence of response Y given α and also by assuming that the components of α are independent conditional on θ. The particular relationship between θ and α in de la Torre and Douglas's article is logistic regression, given as

$$P(\alpha_k = 1|\theta) = \frac{\exp(\lambda_{0k} + \lambda_k \theta)}{1 + \exp(\lambda_{0k} + \lambda_k \theta)}. \tag{15}$$

When conditioning on α, the examinee's response follows the CDM. Therefore, the higher-order model contains a hierarchy, where the CDM forms level one and the logistic regression model forms level two. If the DINA model is used as a level one model, the whole model is called the *higher-order DINA model* (HO-DINA), but the first-level model can be any other cognitive diagnostic model, such as the NIDA model or the RUM model.

Equation 14 shows that the probability of mastering the k^{th} attribute is based on the general ability θ. This model enables one to classify each α_k and obtain an estimated $\widehat{\theta}$ that has a meaning close to that of the unidimensional latent trait in IRT models. In particular, de la Torre and Douglas (2004) demonstrated how an estimate of the θ from the HO-DINA model correlates highly with ability estimate $\widehat{\theta}$ obtained from a 2PL item response model fitted with the same data. Therefore, by generating data from the HO-DINA model, we can have two sets of parameters, one from the 2PL model, including a, b, and $\widehat{\theta}$, which are ready for the unidimensional IRT, and the other set from the DINA model, including slipping, guessing, and α, which are ready for cognitive diagnosis.

Simulation setup

Three thousand higher-order θ_0s were drawn from $N(0, 1)$. It is reasonable to believe that the attributes are correlated in the population, and the correlation of attributes is controlled by the slope λ in the higher-order part of the model. To make the results more general, we considered two cases, one in which attributes were highly correlated and one in which they were not. Results showed that the correlation coefficient between two attributes in high-correlation cases ranged from .62 to .730, while all were below .20 in low-correlation cases. These correlations are quite similar to the real test. For example, in the operational CAT–ASVAB (Armed Services Vocational Aptitude Battery) test, nine dimensions were extracted, with each representing one content area. The correlations between these nine dimensions were around .150 to .908 (Segall, 1996). Assuming that there are five attributes and each attribute is moderately difficult to master, the i^{th} examinee's mastery for attribute k is

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

In this way, the data are simulated to have $2^5 = 32$ discrete diagnostic groups (i.e., the latent classes). The numbers of examinees falling into each diagnostic group are almost balanced.

The item bank size is predetermined to be 800. Q-matrix is carefully defined such that the number of items measuring each attribute is balanced, ranging from 386 to 422. DINA model slipping and guessing parameters were then simulated from a 4-Beta (0, 0.9, 1.5, 2) distribution, following de la Torre and Douglas (2004). Two 3,000 × 800 complete response matrices were generated separately according to the DINA model, and each one was retrofitted with a two-parameter logistic (2PL) model using BILOG. As a result, we obtained a- (discrimination) and b- (difficulty) parameters for the same 800-item pool under the two conditions. The descriptive statistics of the item pool and examinee sample are presented in Tables 2 and 3.

Table 2 Number of items measuring (or examinees mastering) each attribute

	Attribute				
	1	2	3	4	5
Number of items	386	415	422	412	396
Number of examinees (high correlation)	1,961	1,897	1,839	1,452	1,535
Number of examinees (low correlation)	1,918	1,944	1,868	1,192	1,530

Table 3 Descriptive statistics of item parameters

Item Parameter	a	b	Slipping	Guessing
High correlation				
Mean	1.1825	0.0098	0.0907	0.1428
SD	0.5915	0.7994	0.0848	0.1413
Low correlation				
Mean	0.9477	0.1274	0.0907	0.1428
SD	0.5306	0.8055	0.0848	0.1413

Meanwhile, we have $\widehat{\theta}_0$ from the output, which is the limiting value of $\widehat{\theta}$ estimated from the entire bank of items, using the 2PL model. Data showed that the correlation between the higher-order true θ_0 and this limiting $\widehat{\theta}_0$ is .79 and .81 in the two cases, respectively, indicating that the $\widehat{\theta}_0$ from the somehow misspecified 2PL model converges approximately to the true θ_0 in the limiting sense (de la Torre & Douglas, 2004). Therefore, the higher-order model yields 2PL-like data, and thus the "wrong" 2PL model gives a meaningful quantification of ability, which will be used as the "truth" in our method evaluation below.

Six target item selection methods are considered in this simulation study, with three of them based on a-stratification and the other four based on the maximum Fisher information criterion. All of them utilize cognitive diagnosis information in varying levels. Two baseline methods are also considered: pure a-stratification (denoted as "Stratified" hereafter) and the pure maximum information method (denoted as "MI" hereafter). One would expect both MI and Stratified to perform well in estimating θ, but they do not recognize any cognitive diagnosis information that should assist in classification of α. The random item selection method is also included as a general baseline.

For a-stratification-based methods, the item bank was partitioned into four equally large strata such that the a-parameters are in an ascending order, whereas b-parameter distributions are roughly the same across the strata. For a detailed introduction of the procedure, please see Chang and Ying (1999). To make things more general, we considered two test lengths, 21 and 41. The two values were intentionally chosen to simplify the simulation; for a-stratification-based methods, besides the first item, 5 items were chosen from each stratum when the test length was 21, and 10 items were selected from each stratum when the test length was 41. The maximum exposure rate was set to be 0.2 for long tests and 0.1 for short tests.

Evaluation criterion

To check the attribute estimation, the proportion of attributes that were correctly identified was recorded—namely, the recover rate. The recover rate of attribute

mastery was computed marginally for each attribute and for the entire attribute pattern as well. The pattern recovery (PR) rate and attribute recovery (AR) rate are computed as

$$PR = \frac{\sum_{i=1}^{N} R_i}{N} = \frac{\sum_{i=1}^{N} \left(I_{\left(\widehat{\alpha}_i, \alpha_i\right)} \right)}{N}$$

where $\widehat{\alpha}_i$ and α_i are the estimated and true knowledge states of examinee i, respectively; I is the indicator function. If $\widehat{\alpha}_i$ is equal to α_i —say, we classify the whole pattern correctly—then $R_i = 1$; otherwise, R_i is set to 0.

$$AR_k = \frac{\sum_{i=1}^{N} A_{ik}}{N} = \frac{\sum_{i=1}^{N} \left(I_{\left(\widehat{\alpha}_{ik}, \alpha_{ik}\right)} \right)}{N} (k = 1, 2, ..., K)$$

If $\widehat{\alpha}_{ik}$ equals to α_{ik} —say, we classify the attribute k correctly one time—let $A_{ik} = 1$; otherwise, A_{ik} is set to 0.

Concerning the ability estimation, note that the θ of the higher-order DINA model was used to generate data but the 2PL model was assumed when conducting CAT. Consequently, when evaluating the performance of θ estimation, we refer to both the higher-order trait of the DINA model used to generate data and the limiting value of $\widehat{\theta}$ obtained by estimating θ with responses to the entire bank of items—that is, the $\widehat{\theta}_0$ from BILOG output. One would expect that the mean squared error (*MSE*) from the former would be much larger than the *MSE* from the latter. *MSE* is calculated as

$$MSE = \frac{1}{m} \sum_{i=1}^{m} \left(\widehat{\theta}_i - \theta_{0i} \right)^2,$$

where m is the total number of examinees, and $\widehat{\theta}_i$ is the final MLE estimate for examinee i; θ_{0i} is the corresponding true value from either 2PL or HO-DINA. In order to check the exposure rate balance, the exposure rate distribution was plotted against the item discrimination parameters. We also report the chi square index at the top of each figure; it is defined as

$$\chi^2 = \sum_{j=1}^{N} \left(er_j - \overline{er}_j \right)^2 / \overline{er}_j$$

where er_j is the exposure rate of item j and $\overline{er}_j = L/N$ 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, it quantifies the efficiency of the item bank usage, and the smaller the value, the more efficiently the item pool is used.

Results

Six different target methods (those with different priority indices) are considered in this simulation study. To show that people are more accurately classified into diagnostic groups when the priority indices are used, *non-adaptive* diagnostic testing would be considered as a baseline; they are the Stratified and MI methods. The random method is the overall baseline, which is nonadaptive with respect to both θ and α.

For the test length of 41 and the high-correlation case, estimation results are given in Table 4. We can see that as compared with the baseline MI (or Stratified) method, those methods with various priority controls, such as MIQ, MIQD, and MIinfor (or StraInfor), produced higher recovery rates of α and an even lower *MSE* of θ. The reason is that θ and α are highly correlated, and the estimation accuracy of α will affect the estimation accuracy of θ. By adding the constraint control concerning α, recovery rates of α increase, and as a result, the estimation of θ is more accurate. The methods that have the highest recovery rates are with KL information control, particularly when looking at the whole attribute patterns; this is because

Table 4 Item selection results (test length of 41 and high correlation)

	Recovery Rate						MSE	
	AT1	AT2	AT3	AT4	AT5	Pattern	From 2PL	From HO-DINA
Stratified	0.9943	0.9997	0.9840	0.9597	0.9583	0.9060	0.140	0.565
straQ	0.9893	0.9940	0.9833	0.9857	0.9587	0.9207	0.128	0.594
straQD	0.9897	0.9967	0.9860	0.9827	0.9663	0.9277	0.113	0.561
StraInfor	1.0000	0.9963	0.9960	0.9957	0.9907	0.9800	0.067	0.495
MI	0.9987	0.9973	0.9963	0.9250	0.9370	0.8653	0.096	0.568
MIQ	0.9980	0.9980	0.9993	0.9803	0.9763	0.9540	0.093	0.526
MIQD	0.9967	0.9977	0.9983	0.9870	0.9773	0.9597	0.089	0.500
MIinfor	1.0000	0.9983	0.9987	0.9920	0.9957	0.9853	0.061	0.484
Random	0.9353	0.9473	0.9720	0.9577	0.9423	0.8273	0.195	0.693

attribute-level information takes both *Q*-matrix and CDM parameters into consideration. The results for the shorter test length of 21 are given in Table 5. Here, the same trends hold, and we see a more pronounced advantage for our new approaches. The KL information control method is still the best in terms of correctly identifying the attribute pattern and estimating ability.

One interesting result in the tables above is that the MI method yielded slightly smaller attribute recover rate for some conditions and some attributes than did the Stratified method. In other words, with exposure control in the Stratified method, the estimation accuracy of α surprisingly increased. This might be because when the exposure rate is balanced by stratifying the item pool, it actually takes some steps to balance the attribute coverage as in the *Q*-control; thus, the recovery rate of α increases. As was pointed out by one reviewer, attributes 4 and 5 tend to have slightly lower recovery rates—in particular, when cognitive diagnostic information is not considered. This is because of the item pool, items that measure the fourth and fifth attributes have slightly higher slipping and guessing parameters; thus, those items have slightly lower information.

For the long test and low -correlation case, with results tabulated in Table 6, we see that the KL information control method is even more valuable. This is a more challenging case because of the near independence of the attributes. Classifying the whole vector is more difficult because less information can be compiled about the joint distribution. The *Q* discrimination method comes in a distant second. Similar results obtain in the short test case, with the results given in Table 7. Test length has a more significant result for the estimation of θ than for the estimation of α, since it is shown that the recovery rates of α for the two test lengths did not differ much, especially when the KL information control was employed. The *MSE* of θ, on the other hand, is almost doubled with the short test.

Because under each condition, the exposure rate distribution stays almost the same, we present only the results of the *long test and high-correlation* case as a representative, which is shown in Figs. 1, 2 and 3. The chi-square value is presented above each figure.

It is consistent with our expectation that the MI method led to an increasing trend of item exposure with the increase of item discrimination parameters (Chang & Ying, 1999, 2008). Adding different types of priority indices will balance the exposure rate a little bit, especially for the *Q*-control and *Q* discrimination-control. Fig. 2 shows *a*-stratification-based methods, and the same pattern display is shown here. One thing that needs notice is that in *a*-stratification-based methods, the majority of the administered items seem to have smaller discrimination parameters, which indicates that we might combine the *a*-stratification methods and maximum information methods in the future. The Fig. 4 is a summary of the exposure rate distribution under various methods; each curve represents a cumulative distribution function of the exposure rate. The difference between three sets of methods is easier to detect from this figure. Not surprisingly, MI series methods led to the most skewed exposure distribution, and *a*-stratification-based methods reasonably controlled the item exposure balance. The random selection method produced an exposure rate for all the items in the item bank of roughly around 0.05. To summarize, all three types of priority indices worked well, as compared with the nonadaptive baseline, but classification of α and estimation of θ were best when *KL information* control was used, and *Q* discrimination methods came in second in all cases.

Discussion

Nowadays, to meet the standards of accountability under the provisions of "No Child Left Behind," many testing programs are interested in estimating a broadly defined latent trait to summarize scores, while also gleaning some diagnostic information from an exam. In this article, we proposed an innovative idea to deliver CAT such that it will

Table 5 Item selection results (test length of 21 and high correlation)

	Recovery Rate						MSE	
	AT1	AT2	AT3	AT4	AT5	Pattern	From 2PL	From HO-DINA
Stratified	0.9735	0.9607	0.9683	0.9691	0.9387	0.8287	0.144	0.638
straQ	0.9774	0.9753	0.9677	0.9871	0.9177	0.8543	0.132	0.598
straQD	0.9827	0.9813	0.9771	0.9743	0.9487	0.8647	0.250	0.597
StraInfor	0.9917	0.9799	0.9839	0.9903	0.9696	0.9197	0.105	0.568
MI	0.9837	0.9923	0.9783	0.9263	0.9210	0.8253	0.164	0.571
MIQ	0.9650	0.9647	0.9860	0.9820	0.9830	0.9060	0.124	0.571
MIQD	0.9680	0.9617	0.9843	0.9857	0.9883	0.9110	0.121	0.562
MInfor	0.9987	0.9873	0.9957	0.9847	0.9877	0.9593	0.069	0.556
Random	0.8593	0.8987	0.9000	0.8880	0.8850	0.6630	0.349	0.743

Table 6 Item selection results (test length of 41 and low correlation)

	Recovery Rate						MSE	
	AT1	AT2	AT3	AT4	AT5	Pattern	From 2PL	From HO-DINA
Stratified	0.9913	0.9977	0.9847	0.9440	0.9637	0.8947	0.159	0.753
straQ	0.9783	0.9897	0.9820	0.9677	0.9703	0.9050	0.111	0.732
straQD	0.9803	0.9903	0.9857	0.9700	0.9747	0.9137	0.156	0.684
StraInfor	0.9977	0.9933	0.9917	0.9837	0.9890	0.9587	0.109	0.641
MI	0.9953	0.9967	0.9977	0.9197	0.9500	0.8747	0.118	0.719
MIQ	0.9950	0.9983	0.9987	0.9807	0.9863	0.9623	0.064	0.711
MIQD	0.9983	0.9977	1.0000	0.9860	0.9887	0.9730	0.062	0.701
MInfor	0.9983	0.9983	0.9983	0.9977	1.0000	0.9930	0.049	0.623
Random	0.9403	0.9503	0.9673	0.958	0.9433	0.8273	0.312	0.913

give both formative and summative information. The idea is that we consider cognitive diagnostic needs as test constraints added to the traditional unidimensional CAT. By satisfying the constraints, we can gather enough information to classify mastery or nonmastery of specific skills, thereby facilitating formative assessment.

To meet this objective, one important issue needs to be addressed. When a single examination is used for a standard unidimensional score and for skills diagnosis, one has to consider how these distinct aims can be addressed at once. The higher-order DINA model used in our simulation illustrates how data generated under DINA model item parameters can appear to follow an IRT model, indicating that we can obtain two sets of parameters for both item pool and population. This model also illustrates how θ and α relate in a systematic way. The standard θ still has a useful interpretation as a general and broad level of knowledge or intelligence, and fitting logistic IRT models can be quite a good approximation when CAT is conducted and $\widehat{\theta}$ is estimated. In this way, CAT may be conducted in a manner that assigns this general score but can also diagnose the presence or absence of each component of an attribute vector. Therefore, it can afford practitioners with the opportunity to rank order the examinees while diagnosing them in a way that leads to tailored remediation.

The DINA model is used in simulation because it is parsimonious, tractable, and interpretable, but the methods proposed in this study can be generalized to other CDMs. It will make no difference in Q-control, and the only thing that needs to be changed in Q discrimination-control is the "discrimination" component. For the DINA model, it would be $(1\text{-}s)(1\text{-}g)$, and for the RUM model, it would be r_k^*, which represents the penalty for not possessing the k^{th} attribute (Henson & Douglas, 2005). For Information-control, the only modification is the item response function defined in Eq. 8, which varies with the model.

One may argue that the common factor multidimensional item response theory model (MIRT) or bifactor model might also provide cognitive information for a test. However, as compared with MIRT, CDM works better when there are many attributes—say, more than three—and when these attributes are very fine and specifically defined. Because of the numerical integration required by MIRT, it has not found applications in higher dimensions where CDMs can still be reliably estimated. For smaller numbers of attributes that are defined rather broadly, MIRT might

Table 7 Item selection results (test length of 21 and low correlation)

	Recovery Rate						MSE	
	AT1	AT2	AT3	AT4	AT5	Pattern	From 2PL	From HO-DINA
Stratified	0.9710	0.9853	0.9677	0.9077	0.9140	0.7910	0.195	0.761
straQ	0.9663	0.9797	0.9613	0.9420	0.9387	0.8267	0.176	0.746
straQD	0.9687	0.9867	0.9700	0.9540	0.9487	0.8550	0.178	0.685
StraInfor	0.9897	0.9833	0.9790	0.9893	0.96507	0.9067	0.157	0.652
MI	1.0000	1.0000	0.9893	0.8397	0.8467	0.7190	0.168	0.731
MIQ	0.9640	0.9677	0.9790	0.9750	0.9677	0.9440	0.121	0.715
MIQD	0.9797	0.9710	0.9807	0.9860	0.9787	0.9550	0.125	0.709
MInfor	0.9967	0.9947	0.9917	0.9903	0.9880	0.9797	0.104	0.627
Random	0.8175	0.9019	0.9116	0.8907	0.8187	0.5953	0.379	0.941

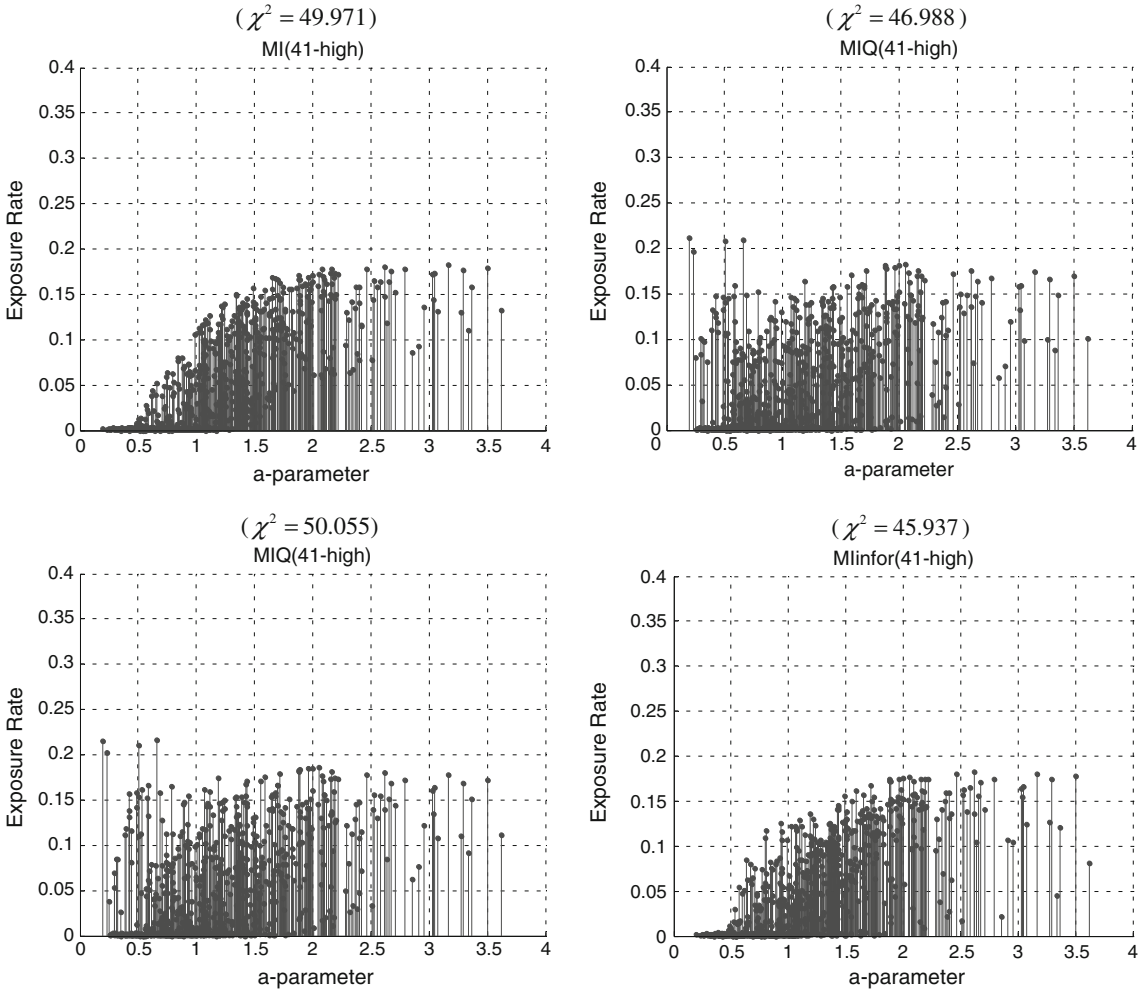

Fig. 1 Exposure rate distribution for MI-based methods

make much more sense. Also, the DINA model is a conjunctive model that corresponds to the theory concerning how items measuring multiple skills are answered (Pardos, Beck, & Heffernan, 2008), while the usual compensatory MIRT models are not consistent with this. In the bifactor model, all items load on both the primary dimension and, at most, one of the rest of the dimensions (Gibbons & Hedeker, 1992; Holzinger & Swineford, 1937). Also, the bifactor model requires that the general factor and specific factors are orthogonal to each other, which is more restrictive for real data. Moreover, unlike the HO-DINA model, all factors in the bifactor model relate to the examinees' performance directly and in a compensatory fashion, while the item response in HO-DINA relies solely on the specific skill status in a conjunctive way. In conclusion, the higher-order DINA models the attributes at the first level, and the higher level is used to model the association of the attributes, so the two models have different interpretations and aims.

To implement the cognitive diagnostic constraints, we focus on less technical and highly practical extensions of the MPI in the traditional CAT. Several modifications are made for our purpose here. First, note that in order to have a fair comparison, we need to enforce the priority index for the different items to be on the equivalent metric. When dealing with nonstatistical constraints, it is often the case that each row in the constraint relevancy matrix **C** will sum up to the same constant, indicating that each item has the same number of constraints. However, in *Q*-control or *Q* discrimination-control, considering that the number of attributes measured by each item often varies, instead of putting the *Q*-matrix elements as exponents, we incorporate them into multipliers themselves. In information control, due to the continuous measure of the information, we can no longer regard it as a nonstatistical constraint; instead, we construct an entirely new index that is a weighted sum of attribute level KL information. The several extensions we made to the original priority index are the by-products of

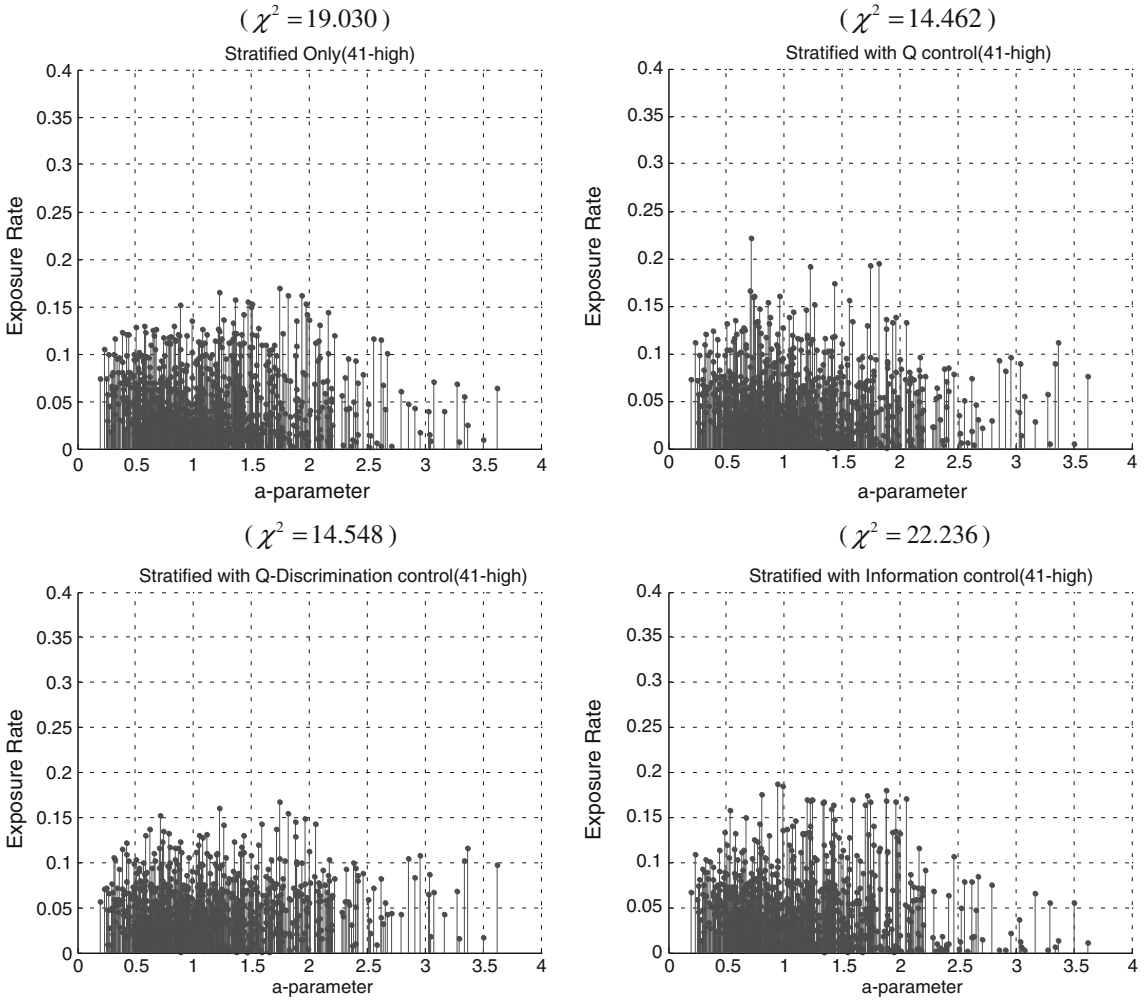

Fig. 2 Exposure rate distribution for a-stratification-based methods

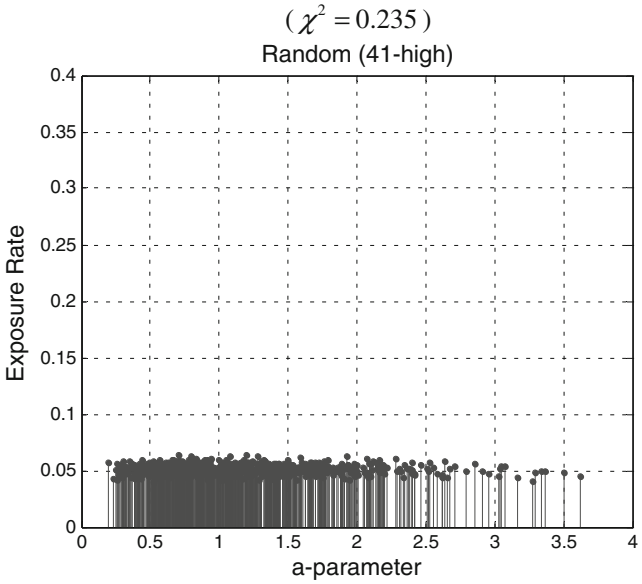

Fig. 3 Exposure rate distribution for random method

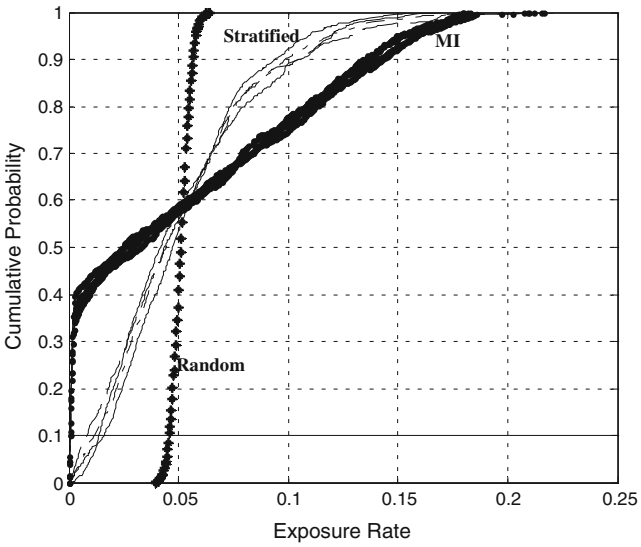

Fig. 4 Cumulative distribution function of exposure rate for each method

our research, and they can be readily used in various conditions when the original one is not appropriate.

The primary objective of cognitive diagnosis is to classify examinees into latent classes determined by vectors of binary skill indicators, and in terms of general latent class modeling, the CDMs belong to multiple classification latent class models. The reliability of the diagnostic information is reflected by the recovery rate of α; the higher the recovery rate, the more reliable is the cognitive information that was gained. Our main results showed that the most efficient method is the one that directly utilizes Kullback–Liebler information, although a simple priority score based on slipping and guessing parameters performs nearly as well. Using only Q-matrix information was not as effective, but it takes one step toward achieving balance across attributes. These conclusions are most apparent when results for the shorter exam are inspected and for classification of the entire attribute vector. In this study, we consider only the case where the Q-matrix is correctly specified in advance. However, in practice, if the Q-matrix is misspecified, the α estimation will be negatively affected (Rupp & Templin, 2008). In this situation, the Q-control or Q discrimination-control may not be as beneficial as in our simulation results, and further research is needed in this regard.

CAT can be expanded to perform cognitive diagnosis if an underlying model that contains information about a continuous θ and binary latent attribute vector α is carefully chosen. The unidimensional IRT provides a simplification of test data by assuming a single latent trait. CDMs 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. (2004). *Understanding Computerized Adaptive Testing – From Robbins-Monro to Lord and beyond*. In David Kaplan (Ed.) The SAGE handbook of quantitative methodology for the social sciences (pp. 117–133), Sage Publications.www.sagepub.com/book.aspx?pid=10048

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

Chang, H.-H., & Ying, Z. (2008). To weight or not to weight? Balancing influence of initial items in adaptive testing. *Psychometrika, 73*, 441–450.

Cheng, Y. (2009). When cognitive diagnosis meets computerized adaptive testing: CD-CAT. *Psychometrika, 74*, 619–632.

Cheng, Y., & Chang, H.-H. (2009). The maximum priority index method for severely constrained item selection in computerized adaptive testing. *British Journal of Mathematical and Statistical Psychology, 62*, 369–383.

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

Cheng, Y., Guo, F., Chang, H., & Douglas, J. (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.

de la Torre, J. (2009). DINA model and parameter estimation: A didactic. *Journal of Educational and Behavioral Statistics, 34*, 115–130.

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

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

Gibbons, R. D., & Hedeker, D. (1992). Full-information item bi-factor analysis. *Psychometrika, 57*, 423–436.

Greeno, J. G. (1980). Trends in the theory of knowledge for problem solving. In D. T. Tuma & F. Reif (Eds.), *Problem solving and education: Issues in teaching and research* (pp. 9–23). Hillsdale, NJ: Erlbaum.

Haertel, E. H. (1989). Using restricted latent class models to map the skill structure of achievement items. *Journal of Educational Measurement, 26*, 333–352.

Hau, K. T., & Chang, H. (2001). Item selection in computerized adaptive testing: Should more discriminating items be used first? *Journal of Educational Measurement, 38*, 249–266.

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.

Holzinger, K. J., & Swineford, F. (1937). The bi-factor method. *Psychometrika, 2*, 41–54.

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.

Kingsbury, G. G. (2009). Adaptive item calibration: A process for estimating item parameters within a computerized adaptive test. In D. J. Weiss (Ed.), *Proceedings of the 2009 GMAC Conference on Computerized Adaptive Testing*.

Leung, C., Chang, H., & Hau, K. (2005). Computerized adaptive testing: A mixture item selection approach for constrained situations. *British Journal of Mathematical and Statistical Psychology, 58*, 239–257.

Lord, F. M., & Novick, M. R. (1968). Statistical theories of mental test scores. Reading, MA: Addison-Wesley.

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

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

Pardos, Z. A., Beck, J. E., & Heffernan, N. T. (2008). The composition effect: Conjunctive or compensatory? An analysis of multi-skill math questions in ITS. Retrieved from http://www.educationaldatamining.org/EDM2008/uploads/proc/15_Pardos_44.pdf

Rupp, A. 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.

Segall, D. O. (1996). Multidimensional adaptive testing. *Psychometrika, 2*, 331–354.

Stocking, M. L., & Swanson, L. (1993). A method for severely constrained item selection in adaptive testing. *Applied Psychological Measurement, 17*, 277–292.

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. (1985). A probabilistic model for diagnosing misconceptions in the pattern classification approach. *Journal of Educational Statistics, 10*, 55–73.

Tatsuoka, K. K., & Tatsuoka, M. M. (1987). Bug distribution and pattern classification. *Psychometrika, 52*(2), 193–206.

Tatsuoka, K. (1990). *Toward an integration of item-response theory and cognitive error diagnosis*. In N. Frederiksen, R. Glaser, A. Lesgold, & M. Safto (Eds.), Monitoring skills and knowledge acquisition (pp. 453–488). Hillsdale, NJ: Erlbaum.

Tatsuoka, K. K. (1991). Boolean algebra applied to determination of universal set of knowledge states (Tech. Rep. RR-91-44-ONR). Princeton, NJ: Educational Testing Service.

Tatsuoka, K. K., & Tatsuoka, M. M. (1997). Computerized cognitive diagnostic adaptive testing: Effect on remedial instruction as empirical validation. *Journal of Educational Measurement, 34* (1), 3–20.

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

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

Templin, J., Henson, R., & Douglas, J. (2008). *General theory and estimation of cognitive diagnosis models: Using Mplus to retrieve model estimates*. Manuscript under review

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

Timminga, E. (1998). Solving infeasibility problems in computerized test assembly. *Applied Psychological Measurement, 22*, 280–291.

van der Linden, W. J. (2000). Constrained adaptive testing with shadow tests. In W. J. van der Linden & C. A. W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 27–52). Dordrecht: Kluwer.

van der Linden, W. J., & Chang, H. (2003). Implementing content constraints in Alpha-Stratified adaptive testing using a shadow test approach. *Applied Psychological Measurement, 27*(2), 107–120.

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

Wainer, H. D., Dorans, N. J., Flaugher, R., Green, B. F., Mislevy, F. J., Steinberg, L., et al. (1990). *Computerized adaptive testing: A primer*. Hillsdale, NJ: Erlbaum.

Xu, X., Chang, H.-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.

Yamamoto, K. (1989). *HYBRID model of IRT and latent class model*. ETS research report series (RR 89-41). Princeton, NJ: Educational Testing Service.

Yi, Q., Zhang, J., & Chang, H. (2008). Severity of organized item theft in computerized adaptive testing: A Simulation study. *Applied Psychological Measurement, 32*(7), 543–558.