Article

Application of Binary Searching for Item Exposure Control in Cognitive Diagnostic Computerized Adaptive Testing

Applied Psychological Measurement
2017, Vol. 41(7) 561–576

DOI: 10.1177/0146621617707509
journals.sagepub.com/home/apm
SAGE

Chanjin Zheng[1] and Chun Wang[2]

Abstract

Cognitive diagnosis has emerged as a new generation of testing theory for educational assessment after the item response theory (IRT). One distinct feature of cognitive diagnostic models (CDMs) is that they assume the latent trait to be discrete instead of continuous as in IRT. From this perspective, cognitive diagnosis bears a close resemblance to searching problems in computer science and, similarly, item selection problem in cognitive diagnostic computerized adaptive testing (CD-CAT) can be considered as a dynamic searching problem. Previously, item selection algorithms in CD-CAT were developed from information indices in information science and attempted to achieve a balance among several objectives by assigning different weights. As a result, they suffered from low efficiency from a tug-of-war competition among multiple goals in item selection and, at the same time, put an undue responsibility of assigning the weights for these goals by trial and error on users. Based on the searching problem perspective on CD-CAT, this article adapts the binary searching algorithm, one of the most well-known searching algorithms in searching problems, to item selection in CD-CAT. The two new methods, the stratified dynamic binary searching (SDBS) algorithm for fixed-length CD-CAT and the dynamic binary searching (DBS) algorithm for variable-length CD-CAT, can achieve multiple goals without any of the aforementioned issues. The simulation studies indicate their performances are comparable or superior to the previous methods.

Keywords

CD-CAT, searching algorithms, binary searching, restrictive progressive (RP) method, restrictive threshold (RT) method, SHTVOR

[1]Jiangxi Normal University, Nanchang, China
[2]University of Minnesota, Minneapolis, MN, USA

Corresponding Author:
Chanjin Zheng, School of Psychology, Jiangxi Normal University, 99 Ziyang Avenue, Nanchang, Jinagxi 330022, China.
Email: russelzheng@gmail.com

Introduction

Cognitive diagnosis has emerged as a new generation of testing theory for educational assessment after the item response theory (IRT). It represents a departure from previous testing theories because it has relevance not only to teachers and students but also to cognitive psychologists who investigate the cognitive process of problem solving (Greeno, 1980) and psychiatrists who need to identify specific psychological disorders (Templin & Henson, 2006). Although cognitive diagnosis has made much progress in developing the state-of-the-art technology (Rupp & Templin, 2008), its application still faces some practical challenges. One of the major applications of cognitive diagnosis is to implement cognitive diagnosis through technology-enhanced computerized adaptive testing (CAT). Over about five decades, the technologies for the IRT-based CAT have matured and the psychometrics behind it have also carefully explicated (Chang, 2015). With the popularity of CAT, the cognitive diagnostic CAT (CD-CAT) has been attracting more and more practitioners' attention (Wang, Chang, & Douglas, 2012), which has generated considerable interest in building item selection algorithms for CD-CAT.

Measurement accuracy has been the major theme for item selection algorithm in CD-CAT (Zheng, 2015; Zheng & Chang, 2016). Item selection algorithm development, however, has to take measurement accuracy and other practical considerations into account, among which item exposure control is the most intensively studied one. The restrictive progressive (RP) method and the restrictive threshold (RT) method from Wang, Chang, and Huebner (2011) are the two methods developed specifically to address the item exposure control issue in fixed-length CD-CAT; for the case of variable-length CD-CAT, Hsu, Wang, and Chen (2013) proposed a method based on the Sympson–Hetter (SH) method, which comprises test overlap control, variable length, online update, and restricted maximum information (SHTVOR). It is worth pointing out that the basic strategy shared by RP, RT, and SHTVOR is to put all the relevant elements (an information-based index to ensure measurement accuracy and others for item exposure control, test overlap, etc.) in one single index and then attempt to strike a balance among these competing objectives. These item exposure control methods are very similar to a tug-of-war and inevitably suffer from low efficiency due to the competition and compromises among the multiple objectives as the SH in CAT does (Chang & Ying, 1999).

The current study attempts to adapt the binary searching algorithm in searching problems in computer science to the item exposure control problem in CD-CAT. The new methods can be considered as an analogy of the *b*-matching method (Hulin, Drasgow, & Parsons, 1983; Urry, 1971; Weiss, 1974) and they share the same advantage as the α-stratification method (Chang & Ying, 1999), which is that they do not achieve the multiple purposes through competition and thus is free from the low efficiency issue.

This adaptation is motivated by one distinct feature of cognitive diagnostic models (CDMs), which is that they assume the latent trait to be discrete instead of continuous as in IRT, and thus the target space consists of distinct mutually exclusive elements labeled as *cognitive patterns*. From this perspective, cognitive diagnosis bears a close resemblance to searching problems in computer science, in which the goal is to identify a targeted element (cognitive pattern) among limited number of ordered ones. Searching problems are a well-studied topic in computer science; a myriad of well-established searching algorithms have been developed (Knuth, 1973) and can be exploited for the CD-CAT. The major goal of the current article is to investigate the possibility of applying one of the most commonly used searching algorithms, binary searching, to the CD-CAT.

The remaining sections of the article are laid out as follows: The following section will give a brief introduction to the item exposure control studies in CD-CAT. In the ''Dynamic Binary

Searching Algorithm'' section, the binary search algorithm is presented and how it can be adapted to CD-CAT is illustrated. In the ''Simulation Studies'' section, two simulation studies are conducted to evaluate the new algorithms against the previous methods in the fixed-length and variable-length CD-CAT. Finally, the ''Discussion'' section is concluded with a discussion of the findings of this work and directions for future research.

Item Exposure Control in CD-CAT

In past decades, there have been a number of different exposure control approaches proposed in the literature. As identified by Georgiadou, Triantafillou, and Economides (2007), there are at least five different types of exposure control strategies: (a) randomization (Kingsbury & Zara, 1989; McBride & Martin, 1983), (b) conditional selection (Chang & Ansley, 2003; Chen & Lei, 2005; Stocking, 1993; Sympson & Hetter, 1985; van der Linden & Veldkamp, 2004), (c) stratification strategies (Chang, Qian, & Ying, 2001; Chang & Ying, 1999; Yi & Chang, 2003), (d) combined strategies (Eggen, 2001; Leung, Chang, & Hau, 2002; Revuelta & Ponsoda, 1998; van der Linden & Chang, 2003), and (e) multiple-stage adaptive test designs (Luecht & Nungester, 1998). Existing item control methods in CD-CAT fall into one or two of these categories.

Item Exposure Control in Fixed-Length CD-CAT

The two restrictive stochastic methods for item selection in CD-CAT, RP and RT, fall into the first and fourth categories introduced above. As their names indicate, the basic idea of the methods is to change the original deterministic approach based purely on item information to a stochastic approach. This is accomplished by imposing a random component in item selection or selecting an item from a candidate set rather than strictly selecting the item with the maximum information.

RP method consists of two controls, progressive control and restrictive control. The primary idea of progressive control is to add a stochastic component to the item selection criterion (Revuelta & Ponsoda, 1998), such that it will not always choose the items with the highest information. The restrictive control seeks to suppress overexposure by adding a restriction on the maximum exposure rate. Combining the two ideas leads to the RP item selection index for the *j*th item being denoted as

$$\mathrm{RP}_j = \left(1 - \frac{\exp_j}{r}\right)\left[(1 - x/L)R_j + \mathrm{PWKL}_j \times \beta x/L\right],$$

where x is the number of items administered, L is the test length, β is the importance parameter, PWKL is the posterior-weighted Kullback–Leibler method (Cheng, 2009), and $R_j \sim \mathrm{uniform}(0, \max(\mathrm{PWKL}(X_j))$. The restriction component is the term $(1 - \exp_j/r)$ which ensures that the maximum of the current item exposure rate $\exp_j$ for the *j*th item will be kept under a certain value, r. The progressive component is the changing weight $(1 - x/L)$ of the random component. In so doing, the stochastic component can achieve a decent item exposure rate at the early stage while the measurement precision can still be maintained or only slightly decreased due to the increasing importance of the information in the later stage.

RT is also comprised of two parts, a restrictive component and a threshold component. The threshold component is designed to construct sets of items within an information interval:

$$\left[\max\left(\mathrm{PWKL}_j\right) - \delta, \max\left(\mathrm{PWKL}_j\right)\right],$$

where δ is the threshold parameter and is defined as $[\max(\text{PWKL}_j) - \min(\text{PWKL}_j)] \times f(x)$, in which x is the number of items already administered and $f(x) = (1 - (x/L))^{\beta}$ is a monotone decreasing function. β is the importance parameter and controls the relative importance of the exposure balance versus the estimation precision. And the items with exposure rates which exceed the maximum exposure rate will be excluded from item selection. For some constant δ, and the items whose information lies in this interval form a candidate set, $S_{(c)}$, from which one item is selected randomly as the next one to be administered in a CD-CAT.

Previous simulation studies show that RP and RT perform well in terms of maintaining measurement accuracy and decent item exposure balance. It is not difficult, however, to notice that both of them resort to explicit control over these two competing objectives. This issue is specifically reflected in choosing a proper value for the importance parameter β, which is contingent on many factors such as item parameters, test length, and so on. To make it more difficult for users, one has to accomplish this by trial and error, although Wang et al. (2011) offered some recommendations. As a result of an improper importance parameter, RP and RT can suffer from low efficiency in item use indicated by the simulation study (see Simulation Study I for details). In addition, because test length has to be predetermined, both of them are only applicable to fixed-length CD-CAT, but not the variable-length case which is the topic for the following section.

Item Exposure Control in Variable-Length CD-CAT

Hsu et al. (2013) proposed SHTVOR for variable-length CD-CAT. This procedure is based on the SH method (Sympson & Hetter, 1985) and is capable of controlling test overlap for variable-length termination, online updating the exposure control parameters, and using restricted maximum information to freeze items with an exposure rate greater than the prespecified maximum until their exposure rate decreases. As the name suggests, SHTVOR falls into the fourth category (the combined strategy) and it suffers from the inherent problem of low efficiency in the SH method. Its implementation is more complicated than the SH, RP, and RT methods. SHTVOR consists of the following seven steps (Hsu et al., 2013):

1. Initialize/set the parameters, such as the number of items in the bank J, the target maximum item exposure rate $r_{\max}$, target test overlap rate $\bar{T}_{\max}$, the exposure control parameter of item p_k, and so on.
2. Administer CAT to an examinee by comparing p_k with a randomly generated number from $U(0, 1)$. If p_k is larger, then administer the item; otherwise, select another item from the item pool and compare again to determine whether the item can be administered. Repeat this procedure until an item is administered. Exclude the administered item for this examinee from the item pool.
3. Update the examinee's cognitive pattern estimate and select another item as described in Step 2 until the examinee has reached the prespecified fixed precision or until the maximum test length is reached.
4. Compute $P(A)$ and $P(S)$ for item $j(j = 1, 2, \ldots, J)$ as the percentage an item has been administered and selected, respectively. Update $\bar{T}$ and p_k as follows:

$$\bar{T} = \frac{N \times \sum_{j=1}^{J} P(A_j)^2}{\bar{L} \times (N-1)} - \frac{1}{(N-1)},$$

where $\bar{L}$ is the mean test length across all examinees, and N is the total number of examinees who have undergone CAT thus far. Update p_k as follows:

$$p_k = \begin{cases} 0, \text{if} P(A) \geq r_{\max} \\ \frac{r_{\max}}{P(S)}, \text{if} P(A) \leq r_{\max} \text{ and } P(S) > r_{\max} \\ 1, \text{if} P(A) \leq r_{\max} \text{ and } P(S) \leq r_{\max}. \end{cases}$$

5. If $\bar{T} > \bar{T}_{\max}$, then
   a. Calculate the target variance of the item exposure rate across item S_0^2 while $\bar{T} = \bar{T}_{\max}$.
   b. Set $P(A)' = S_0 \left[\frac{P(A)-\bar{r}}{S}\right] + \bar{r}$ and $P'_k = \frac{P(A)'}{P(S)'}$ where $P(A)'$ is the adjusted percentage of times that an item has been administered based on S_0, and P'_k is the adjusted exposure control parameter.
   c. If $P'_k > 1$, then $P'_k = 1$; if $P_k > P'_k$, then $P_k = P'_k$.
6. Set the $L_{\max}$ largest P_ks as 1 to guarantee that all examinees will complete the CAT before exhausting the entire bank.
7. With the updated P_ks, repeat Steps 2 to 6 to administer the CAT again until all of the examinees have finished the CAT.

It is easy to notice that the SHTVOR involves multiple components which serve different objectives. These objectives might have to compete among themselves and then result in low efficiency as in RT and RP. Specifically, a very conservative criterion for the target test overlap rate cannot necessarily guarantee balanced item bank use evidenced by the number of over- or underused items in the bank. Furthermore, similar to RT and RP, the SHTVOR also put the burden of assigning a proper value to the target test overlap rate, $\bar{T}_{\max}$, on users.

To resolve aforementioned issues, the current study proposes some new methods based on the binary searching to address the item exposure control issue in CD-CAT. They relieve users of the difficult task of assigning proper values for the importance parameter or the target overall test overlap, but can produce comparable or better results than the previous methods.

Dynamic Binary Searching Algorithm (DBSA)

The Linear and Binary Searching Algorithms

A brief introduction to the linear and binary searching in computer science is presented as the background information and the starting point for the development of the new methods. This brief introduction is heavily borrowed from Rosen (2011) and Knuth (1973), but in a more accessible manner. The problem of locating an element in an ordered list occurs in many contexts. For instance, a program that checks the spelling of words searches for them in a dictionary, which is just an ordered list of words. Problems of this kind are called searching problems. The general setup for searching problems can be described as follows: Locate an element x in a list of distinct elements $a_1, a_2, \ldots, a_n$, or determine that it is not in the list. The solution to this search problem is the location of the term in the list that equals x (i.e., i is the solution, if $x = a_i$) and is 0 if x is not in the list. A searching algorithm is one for finding an item with specified properties among a collection of items. The linear and binary searching are two fundamental searching algorithms.

Linear searching or sequential searching is the simplest search algorithm and it is a special case of brute-force search. It is a method for finding a particular value in a list, which consists

Table 1. The Efficiency Analysis of the Linear and Binary Searching for n Objects.

Searching algorithms	Average case	Best case	Worst case
Linear searching	$n+2$	1	$2n$
Binary searching	$2\log n$	1	$2\log n$

of checking every one of its elements, one at a time and in sequence, until the desired one is found. The elements in the list are ordered in a certain way. More specifically, the linear searching algorithm begins by comparing x and a_1. When $x = a_1$, the solution is the location of a_1, namely, 1. When $x \neq a_1$, compare x with a_2. If $x = a_2$, the solution is the location of a_2, namely, 2. When $x \neq a_2$, compare x with a_3. Continue this process, comparing x successively with each term of the list until a match is found where the solution is the location of that term, unless no match occurs. If the entire list has been searched without locating x, the solution is 0. There are two common cases for linear searching. The first one in which all list elements are equally likely to be searched for (uniformly distributed) is denoted as linear searching with equal probabilities. The second case in which some values are much more likely to be searched than others is denoted as linear searching with unequal probabilities.

Binary searching, also known as half-interval searching and logarithmic searching, is a dichotomic divide-and-conquer searching algorithm for an ordered list. The binary searching algorithm proceeds by comparing the element located in the middle of the list, namely, $a_{n/2}$ if n is even or $a_{(n+1)/2}$ if n is odd. The list is then split into two smaller sublists of the same size, or where one of these smaller lists has one fewer term than the other. The search continues by restricting the search to the appropriate sublist based on the previous comparison until the solution is obtained.

The efficiency of the two searching algorithms is evaluated by three types of complexity analysis: a worst-case, an average-case, and a best-case analysis. A worst-case analysis refers to the largest number of operations needed to solve the given problem using this algorithm on input of specified size. A worst-case analysis tells how many operations an algorithm requires to guarantee that it will produce a solution. Similar definition can be given to the best-case analysis and the average-case analysis. Assuming that the key must be in the list of n objects which is the case for CD-CAT, the worst-case, the average-case, and the best-case complexity for linear and binary searching are presented in Table 1. The average-case and worst-case analyses for the linear searching with equal probabilities show that the largest/expected number of operations required to complete the linear searching is proportional to n while those for the binary searching is $\log n$. Therefore, if the list has a large number of elements, the binary searching algorithm is much more efficient than the linear searching algorithm.

DBSA for CD-CAT

The new methods are built upon the linear or binary searching algorithm, consisting of two sequential steps: The first is the dynamic searching which determines the optimal vector from a Q-matrix (denoted as a Q-vector hereafter) for an item bank based on the current cognitive pattern estimate, and the second is to randomly select one item for administration from the group of items with that optimal Q-vector. The first step ensures measurement accuracy to some extent, and the second can equalize the item exposure rates. The item exposure control mechanism is randomization (Kingsbury & Zara, 1989; McBride & Martin, 1983), the first category,

but the major difference from the original randomization is that it takes place within a preselected group of items that might potentially enhance measurement accuracy. The distinct feature of the new methods is that there is no competition among the two objectives, and users do not have the responsibility of striking a balance between them. Below, this paper will focus the discussion on the first step, the development of the dynamic searching algorithms, because the second step is simple and well known to researchers and practitioners.

The idea of linear and binary searching can be naturally extended to CD-CAT because the latent trait is discrete. In the language of searching problems, the ordered objects in CD-CAT are the elements in the cognitive pattern distribution, namely, cognitive patterns. The ultimate goal of CD-CAT is to identify the appropriate cognitive pattern of an examinee. The bridge between cognitive patterns and items is the Q-matrix, so item selection in CD-CAT amounts to finding the most appropriate Q-vector by matching with the current interim estimate of an examinee's cognitive pattern. This makes an obvious analogy to the *b*-matching method in the IRT-based CAT (Hulin et al., 1983; Urry, 1971; Weiss, 1974) in which only the difficulty parameter and the latent trait estimate are involved. Therefore, one may name this step as pattern-matching.

In the context of CD-CAT, however, there are some important differences between searching problem in CD-CAT and general searching problems in computer science. First, it is worth noting that the elements in CD-CAT (cognitive patterns) are not strictly ordered as in general searching problems, but partially ordered (Tatsuoka, 2002; Tatsuoka & Ferguson, 2003). For example, four distinct cognitive patterns can be identified for two attributes: $(00), (10), (01), (11)$. The first and fourth patterns can be properly ordered, representing the smallest and biggest elements in the list, but there is no strict order in the second and third ones. This may pose a challenge in general searching problems, but it can be easily handled in CD-CAT (explanations follow in the sequel). Second, the searching in CD-CAT is dynamic. It is a common and necessary practice to update the cognitive pattern posterior (the target list) after an item is administered and the response is received in CD-CAT while the probabilities for the target list remain constant in general searching problems. Usually, the probability mass tends to concentrate in a few cognitive patterns or even one after a few items have been administered.

Dynamic linear searching. The linear version of the dynamic searching or the pattern-matching is extremely simple and straightforward. Assuming the interim estimate for examinee's cognitive pattern based on the current posterior is accurate, the optimal Q-vector of ideal candidate items is exactly the same to the cognitive pattern with the largest posterior probability. It is worth pointing out that the partial order in cognitive patterns whose probabilities are not largest can be conveniently ignored because the item selection does not involve them at all. In case of multiple cognitive patterns with the largest probability, one may consider all of them as the optimal Q-vectors. All items with such a Q-vector (or Q-vectors) in the item bank are selected to form a subset from which an item will be chosen randomly to administer in the second step. So, the item selection algorithm based on the dynamic linear searching can be described as an iterative process as follows:

Step 1: Identify the optimal Q-vector which is the same to the cognitive pattern with the largest probability;
Step 2: Randomly choose an item from all the items with optimal Q-vector.
Step 3: Update the cognitive pattern posterior.

Repeat Steps 1 to 3 until the CAT satisfies the termination rule.

Although the linear pattern-matching is simple and straightforward, there are two inherent defects associated with it. One prominent problem with the linear pattern-matching is its measurement efficiency particularly at the early stage of CD-CAT. As not much information on the posterior is gained and the posterior is close to the uniform distribution, the linear searching strategy is essentially a linear searching with equal probabilities. A preliminary simulation study confirmed these conjectures. The other problem with the linear pattern-matching is that the items with the corresponding cognitive pattern might not be available in the item bank. This is not uncommon in CD-CAT. In practice, Q-matrix might not contain all the possible cognitive patterns; in fact, more commonly, it only involves one, two, or three attributes, so it creates some practical difficulties to carry out the pattern-matching step. For these two problems, the dynamic linear searching algorithm will not be included in the ''Simulation Studies'' section.

Dynamic binary searching (DBS). The DBS is to select the items such that their Q-vector can split all the possible cognitive patterns into two equal groups, with respect to the posterior probabilities, given the current estimated posterior. Such splitting is called separation in the partially ordered set theory for CDMs (Tatsuoka, 2002; Tatsuoka & Ferguson, 2003) and can handle the issue of partial order in cognitive patterns. The splitting rule is very similar to the calculation of the latent response η in the deterministic input, noisy, and gate (DINA) model (Haertel, 1989; Junker & Sijtsma, 2001). The separation S_{jm} for item j and pattern m is defined as

$$S_{jm} = \prod_{k=1}^{K} I\{q_{jk} \leq \alpha_{mk}\} = \begin{cases} 1 & \text{if pattern } m \text{ possesses all the required skills required for item } j \\ 0 & \text{if pattern } m \text{ lacks at least one of the required skills for item } j \end{cases}.$$

Usually, it is difficult to obtain two groups with exactly equal posterior probabilities, and the Q-vector closest to the ''middle point'' is preferred. So, the binary searching index B_j for the tth administration can be formulated as

$$B_j^t = \left| \sum_{S_{jm}=1} g(\alpha_m | y_{t-1}) - 0.5 \right|,$$

where $g(\alpha_m|y_{t-1})$ is the posterior probability for pattern m after $t-1$ items have been administered, and the smaller the index for an item is, the better the item is. Particularly so in an ideal case (i.e., the list is just half split by the Q-vector), it is 0. So, the item selection algorithm based on the DBS can be summarized as an iterative process as follows:

Step 1: Identify the optimal Q-vector by calculating the index B_j which requires the separation S_{jm};
Step 2: Randomly choose an item from all the items with optimal Q-vector;
Step 3: Update the cognitive pattern posterior.

Repeat Steps 1 to 3 until the CAT satisfies the termination rule.

It is necessary to point out that the DBS is a rediscovery from the searching problem perspective. Tatsuoka and Ferguson (2003) proposed the halving algorithm, which is identical to the DBS algorithm from the current study. They also gave the mathematical proof on the convergence of the several algorithms to the true value of the cognitive pattern from the partially ordered set theory, which constitutes the mathematical foundation for the measurement accuracy of the DBS. Unfortunately, it has gone unnoticed at the early time of CD-CAT in which measurement efficiency was the key due to its relatively low measurement efficiency compared with

the Shannon entropy (SHE) method (Tatsuoka, 2002; Tatsuoka & Ferguson, 2003). Even in their own study, Tatsuoka and Ferguson used the SHE method for its high measurement efficiency and only mentioned the halving algorithm in passing, presenting no empirical simulation study or envisioning the possibility of this novel application to item exposure issue. The rediscovery and its application opens up the possibility of studying CD-CAT in the perspective of searching problems and is exactly the major contribution of the current study to the CD-CAT item selection research in general and item exposure control issue in particular.

Compared with the dynamic linear searching, the DBS enjoys several advantages. First, it is free from the practical constraint that it might not find the pattern in the item bank. Second, it takes advantage of all the information of the posterior distribution, unlike the dynamic linear searching which is only concerned with the single element with the largest probability. It can be expected that it can achieve higher measurement efficiency in selecting items especially at the early stage of CD-CAT.

A further modification for fixed-length CD-CAT. The DBS algorithm can be applied to both fix-length and variable-length CD-CAT, but it can be further enhanced by stratifying item bank by an item discrimination index based on item quality, as in the α-stratification method, in fix-length CD-CAT. A stratified dynamic binary searching (SDBS) method, thus, can be proposed for fix-length CD-CAT. The key is to develop counterparts of item discrimination index and *b*-matching method in CD-CAT. As stated above, the DBS can fulfill the role of *b*-matching.

The natural candidate for CD-CAT item bank stratification is item discrimination indices for CDMs. Rupp, Templin, and Henson (2010) provided a summary of item discrimination indices for CDMs. There are two types of indices: the classical testing theory (CTT)–based global indices and the Kullback–Leibler (KL) information–based indices. The CTT-based global indices can be regarded as the counterpart of the α parameter in the IRT and thus as the bank stratification criterion. The CTT-based global indices for DINA and the noncompensatory reparameterized unified model (NC_RUM; Hartz, 2002) are $d_{j,DINA} = (1 - s_j) - g_j$ in which s_j and g_j are the slipping and guessing parameters, and $d_{j,NC-RUM} = \pi_j^* - \pi_j^* \prod_{a=1}^{A} r_{ja}^{*q_{ja}}$ in which π_j^*, r_{ja}^*, and q_{ja} are the baseline parameter, penalty parameter, and the elements in the cognitive pattern vector. This index can be derived for the majority of the cognitive diagnosis models. Thus, the method proposed here can be readily extended to other models. However, in this study, only the DINA model and NC_RUM will be used.

With item discrimination index and DBS ready, a general framework for the stratification method for CD-CAT can be set up as follows:

1. Partition the item bank into M levels according to the item discrimination index;
2. Partition the test into M stages;
3. In the kth stage, select n_k items from the kth level based on the dynamic searching method (note that $n_1 + n_2 + \cdots + n_K$ equal the test length);
4. Repeat Step 3 for $k = 1, 2, \ldots, M$.

Simulation Studies

Two simulation studies were conducted to demonstrate the feasibility of the proposed algorithms compared with the existing methods in fixed-length and variable-length CD-CAT, respectively, because the methods for fixed-length CD-CAT are not applicable for variable-length CD-CAT. To validate the results in different CDMs, two important models were used in

two studies: the DINA model (Haertel, 1989; Junker & Sijtsma, 2001) and NC_RUM (Hartz, 2002). The results for the two models in each study were similar, so the results were reported for only one model in each study and other detailed results were published as online appendix.

Study I

Study I is a simulation for a fixed-length CD-CAT of 40 items that aims to investigate the SDBS strategy's performance.

Item bank and examinees generation. The item bank consisting of 480 items for four-attribute NC_RUM (Hartz, 2002) is generated similar to Cheng (2009). Implementation of cognitive diagnosis requires the construction of the design matrix named as a Q-matrix (Tatsuoka, 1983). For an item bank consisting of J items, the Q-matrix is a $J \times K$ matrix of 1s and 0s that specifies the association between items and K attributes. The entry corresponding to the kth attributes for the jth item, q_{jk}, indicates whether the kth attributes are required to answer item j correctly or not. The Q-matrix used in this study is generated item by item and attribute by attribute. Each item has 20% chance of measuring each attribute. This mechanism is employed to make sure that every attribute is adequately and equally represented in the item pool.

The NC_RUM needs two types of item-level parameters: (a) the baseline parameter π_j^* represents the probability of correct response to item j if all measured attributes have been mastered, and (b) the penalty parameter r_{jk}^* represents the probability of correct response to item j for not having mastered attribute k. The probability of a correct response conditional on the cognitive pattern and item parameters is defined as

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

where α_i denotes the examinee's latent cognitive pattern whose kth element is α_{ik}. The item parameters π_j^* and r_{jk}^* were generated from $U(0.75, 0.95)$ and $U(0.2, 0.95)$, respectively. As the test length is 40, the item bank was partitioned into five strata with equal number of items by the item discrimination index described above.

The cognitive patterns for 2,000 examinees were generated assuming that every examinee has 50% chance of mastering each attribute. For example, in a four-attribute test, there were 16 distinct types of latent classes which were assumed to be equally likely in the population.

Item selection algorithms. Five item selection methods were used in this study, including random, PWKL, RP, RT, and SDBS algorithm. For RP and RT, the importance parameter was set to be 2, recommended by Wang et al. (2011).

Evaluation criteria. These item selection algorithms were evaluated in three aspects: estimation accuracy, item exposure balance, and item bank usage. The evaluation criteria for estimation accuracy include recovery rates of attributes and cognitive patterns. The evaluation criterion for item exposure balance is the scaled χ^2 (Chang & Ying, 1999) that quantifies the equalization of exposure rates. Let m denote the number of examinees, N the size of the item bank, and

$$er_j=\frac{\text{number of times the } j\text{th item is used}}{m},$$

which represents the observed exposure rates of the jth item. Therefore, the desirable uniform exposure rate for all items is $\overline{er}_j=L/N$, in which L is the test length and the scaled χ^2 is designed to measure the similarity between the observed and desired exposure rates:

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

Those for item bank usage are the number of items with less than 2% exposure rate and the number of items with more than 20% exposure rate.

Results. The estimation accuracy, and the measures of exposure balance and the item bank usage of each method are reported in Table 2. As expected, we observe a trade-off between measurement accuracy and item exposure rates. In terms of the estimation accuracy, except the random method, the cognitive pattern recovery rates for the SDBS and RT are 0.910 and 0.92 while that for RP is 0.964. In terms of exposure balance, the SDBS outperforms RT and RP with a scaled χ^2 index of 1.193. In terms of item bank usage, there were no overused items or underused items for the SDBS and RT while RP underused 65 items which is the price for the high recovery rate for the cognitive patterns. It is worth pointing out that the same specification for the importance parameter β in RT and RP produced different results: RT struck a decent balance between measurement accuracy and item use while RP put too much weight on the measurement accuracy which needs more fine-tuning.

Study II

Item bank and examinees generation. The item bank consisting of 480 items for the six-attribute DINA model (Haertel, 1989; Junker & Sijtsma, 2001) was generated in the same manner as Cheng (2009). The DINA model–predicted probability that examinee i will respond correctly to item j is defined by

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

where s_j is the slipping parameter and g_j is the guessing parameter. η_{ij} is the latent response for examinee i, α_i, to item j. The item parameters s_j and g_j were both generated from $U(0.05, 0.25)$. The cognitive patterns for 2,000 examinees were generated in the same manner in which 64 distinct cognitive patterns were assumed to be equally likely in the population.

Item selection algorithms. Three item selection methods were used in this study, including PWKL (the baseline condition), SHTVOR, and DBS algorithm. For SHTVOR, the target maximum item exposure rate $r_{\max}$ was set to be 0.2, and the target test overlap rate $\bar{T}_{\max}$ was set to be 0.03, which is very conservative to ensure a low average test overlap.

Termination rule. Tatsuoka (2002) recommended that a variable-length CD-CAT stops if the posterior probability value associated with any one cognitive pattern exceeds 0.8. A similar rule was adopted in this study, but the stopping criterion was set on three different levels: 0.7, 0.8, and 0.9.

Evaluation criteria. These item selection algorithms were evaluated in terms of two aspects: estimation accuracy and item bank usage. The evaluation criteria for estimation accuracy include attribute recovery rates and pattern correct classification rates (PCCRs), and those for item bank usage are the number of items with less than 2% exposure rate (underused items) and the number of items with more than 20% exposure rate (overused items).

Results. The estimation accuracy and test length statistics for all of the algorithms under different stopping criteria are presented in Table 3 and item bank use in Table 4. The PCCRs for all

Table 2. Recovery Rates and Exposure Balance Indices for the NC_RUM (Fixed Length).

Item selection	Attribute				Pattern	Exposure balance χ^2	Number of overused (>0.2)	Number of underused (<0.02)
	1	2	3	4				
PWKL	0.999	0.999	1.000	0.999	0.998	314.060	57	369
RT	0.994	0.986	0.992	0.991	0.964	43.988	0	65
RP	0.979	0.970	0.981	0.981	0.921	7.801	0	0
SDBS	0.969	0.974	0.971	0.973	0.910	1.193	0	0
Random	0.913	0.905	0.936	0.926	0.735	0.221	0	0

Note. PWKL = posterior-weighted Kullback–Leibler method; RT = restrictive threshold method; RP = restrictive progressive method; SDBS = stratified dynamic binary searching.

Table 3. The Measurement Accuracy and Test Length for the DINA Model (Variable Length).

Stopping criteria	Item selection	Attribute						Pattern	Test length	
		1	2	3	4	5	6		*M*	*SD*
0.7	PWKL	0.942	0.941	0.961	0.965	0.963	0.948	0.774	8.04	1.804
	SHTVOR	0.951	0.951	0.948	0.954	0.953	0.952	0.765	11.76	3.548
	DBS	0.957	0.961	0.951	0.953	0.945	0.954	0.776	12.39	3.549
0.8	PWKL	0.963	0.971	0.973	0.979	0.967	0.971	0.843	9.52	2.297
	SHTVOR	0.971	0.975	0.961	0.971	0.975	0.971	0.852	13.96	4.150
	DBS	0.976	0.973	0.968	0.975	0.970	0.971	0.857	14.45	3.998
0.9	PWKL	0.984	0.988	0.993	0.993	0.984	0.983	0.929	11.72	2.863
	SHTVOR	0.990	0.986	0.989	0.986	0.982	0.984	0.921	17.88	5.434
	DBS	0.987	0.986	0.990	0.987	0.985	0.982	0.920	17.50	4.816

Note. DINA = deterministic input, noisy, and gate; PWKL = posterior-weighted Kullback–Leibler method; SHTVOR = Sympson–Hetter method, which comprises test overlap control, variable length, online update, and restricted maximum information; DBS = dynamic binary searching.

of the algorithms under each stopping criterion are close to each other, so the results for item bank use are comparable.

As regards the item exposure control, it is expected that the PWKL without the item exposure control mechanism generates the largest test overlap rate. When item exposure control is adopted, the test overlap rate for the two algorithms is reduced substantially. It is worth noting that DBS produces a similar test overlap rate to SHTVOR, even though there is no explicit mechanism for controlling it as does SHTVOR.

The number of overused and underused items can provide more information about item bank usage. PWKL has the greatest number of overused and underused items. Item exposure control mechanism can improve these indices significantly. There are no overused items for any of the algorithms with item exposure control methods. In terms of underused items, they exhibit differential performances. The numbers of underused items for SHTVOR are 73, 54, and 25, respectively, for three stopping criteria. Please note that the specifications for the target maximum item exposure rate $r_{\max}$ and the target test overlap rate $\overline{T}_{\max}$ in this study were very conservative to enhance item use, and the final realized test overlap rate is almost identical to DBS, but there is still a few underused items. This is a result of the inability of the SH method to improve the utilization of underused items, even though there is an explicit control over the test overlap rate. By contrast, there are no underused or overused items in DBS.

Table 4. Item Exposure and Item Bank Use for the DINA Model (Variable Length).

Stopping criteria	Item selection	Test overlap	Overused (>0.2)	Underused (<0.02)
0.7	PWKL	0.671	10	412
	SHTVOR	0.033	0	73
	DBS	0.029	0	0
0.8	PWKL	0.622	13	395
	SHTVOR	0.037	0	54
	DBS	0.035	0	0
0.9	PWKL	0.593	15	390
	SHTVOR	0.045	0	25
	DBS	0.042	0	0

Note. DINA = deterministic input, noisy, and gate; PWKL = posterior-weighted Kullback–Leibler method; SHTVOR = Sympson–Hetter method, which comprises test overlap control, variable length, online update, and restricted maximum information; DBS = dynamic binary searching.

In summary, the new method can strike a better balance between measurement accuracy and item bank use than SHTVOR. It is also much simpler to implement, in comparison with SHTVOR.

Discussion

Inspired by the binary searching algorithm in computer science, the authors proposed two new methods based on the DBS algorithm for CD-CAT. The halving algorithm, an algorithm proposed from the partially ordered set theory and identical to the DBS, did not win recognition due to its relatively low measurement efficiency compared with other information-based methods. But this study rediscovered this algorithm from a searching problem perspective and turned this disadvantage into a ''less-is-more'' case. Specifically, this study exploited this fact and proposed two 2-step item selection methods with item exposure control based on the dynamic searching algorithm to address the low efficiency issue of the previous item exposure methods in CD-CAT. The results indicate that the new algorithms can achieve multiple goals of measurement accuracy and item exposure control as effectively as the RT and RP in fixed-length CD-CAT, and more effectively than SHTVOR in variable-length CD-CAT.

On one hand, the DBS can make use of all the information from the posterior distribution. In a sense, it is similar to the PWKL, the Bayesian variant of the KL algorithm, which improves its measurement efficiency to a great extent, although it cannot produce as high measurement accuracy as other information-based methods. On the other hand, the DBS avoids ''overdoing'' in terms of measurement accuracy as most information-based methods do and only determine the ideal Q-vector for next candidate item instead of the best item itself. Some randomness arises naturally to eliminate extremely high or low item exposure rates. Thus, the two-step sequential algorithms based on the dynamic searching conveniently avoid competition of and compromise among the several goals in RT, RP, and SHTVOR, elegantly striking a decent balance between the measurement accuracy and item bank use. The other related practical advantage is that the new methods relieve users of the burden of specifying the importance parameter which can only be done in a trial-and-error manner for the previous methods.

In addition, the new method is very flexible to handle item exposure control issue in both fixed-length and variable-length CD-CAT. RT and RP were developed solely for fixed-length

CD-CAT. Some further work is needed to make them feasible in variable-length CD-CAT. The DBS can easily be used in both scenarios as demonstrated in two simulation studies.

Interesting future directions include identifying more efficient searching algorithms other than the binary searching algorithm. Searching problems have been thoroughly studied in computer science, and a myriad of algorithms have been developed (Knuth, 1973). It is possible to find another searching algorithm more efficient than binary searching in CD-CAT.

Another possible future direction is to tap the great potential of the DBS in various applications. For example, it can be combined with other algorithms to deal with multiple constraints such as item exposure rates, content constraint, key balancing, and so on, in CD-CAT. There is also a possibility of applying it in other scenarios such as automated test assembly in a highly constrained real-world diagnostic testing project as in Wang, Zheng, Zheng, Su, and Li (2016). One another interesting application is in the dual-purpose CAT, which aims to obtaining general overall score and diagnostic information in one single administration (McGlohen & Chang, 2008; Wang et al., 2012; Wang, Zheng, & Chang, 2014). It is a much simpler alternative to the shadow test which selects a group of candidate items in the first stage of the θ- and α-based method (McGlohen & Chang, 2008). It is also a better option than the IRT-based bank stratification method in the weighted approach (Wang et al., 2012).

Declaration of Conflicting Interests

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

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Supplemental Material

The online appendix is available at http://journals.sagepub.com/doi/suppl/10.1177/0146621617707509.

References

Chang, H. H. (2015). Psychometrics behind computerized adaptive testing. *Psychometrika*, *80*, 1-20.

Chang, H. H., Qian, J., & Ying, Z. (2001). a-stratified multistage computerized adaptive testing with b blocking. *Applied Psychological Measurement*, *25*, 333-341.

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

Chang, S. W., & Ansley, T. N. (2003). A comparative study of item exposure control methods in computerized adaptive testing. *Journal of Educational Measurement*, *40*, 71-103.

Chen, S. Y., & Lei, P. H. (2005). Controlling item exposure and test overlap in computerized adaptive testing. *Applied Psychological Measurement*, *29*, 204-217.

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

Eggen, T. (2001). *Overexposure and underexposure of items in computerized adaptive testing* (Measurement and Research Department Reports 2001-1). Arnhem, The Netherlands: CITO Group.

Georgiadou, E. G., Triantafillou, E., & Economides, A. A. (2007). A review of item exposure control strategies for computerized adaptive testing developed from 1983 to 2005. *The Journal of Technology, Learning and Assessment*, *5*(8), 4-28.

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*, 301-321.

Hartz, S. M. C. (2002). *A Bayesian framework for the unified model for assessing cognitive abilities: Blending theory with practicality* (Doctoral dissertation). University of Illinois at Urbana–Champaign.

Hsu, C. L., Wang, W. H., & Chen, S. Y. (2013). Variable-length computerized adaptive testing based on cognitive diagnosis models. *Applied Psychological Measurement*, *37*, 563-582.

Hulin, C. L., Drasgow, F., & Parsons, C. K. (1983). *Item response theory: Application to psychological measurement*. Belmont, CA: Dorsey Press.

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., & Zara, A. R. (1989). Procedures for selecting items for computerized adaptive tests. *Applied Measurement in Education*, *2*, 359-375.

Knuth, D. (1973). *The art of computer programming: Vol. 3. Searching and sorting*. Reading, MA: Addison-Wesley.

Leung, C. K., Chang, H. H., & Hau, K. T. (2002). Item selection in computerized adaptive testing: Improving the a-stratified design with the Sympson-Hetter algorithm. *Applied Psychological Measurement*, *26*, 376-392.

Luecht, R. M., & Nungester, R. J. (1998). Some practical examples of computer-adaptive sequential testing. *Journal of Educational Measurement*, *35*, 229-249.

McBride, J. R., & Martin, J. T. (1983). Reliability and validity of adaptive ability tests in a military setting. In D. J. Weiss (Ed.), *New horizons in testing: Latent trait test theory and computerized adaptive testing* (pp. 223-226). New York, NY: Academic Press.

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

Revuelta, J., & Ponsoda, V. (1998). A comparison of item exposure control methods in computerized adaptive testing. *Journal of Educational Measurement*, *35*, 311-327.

Rosen, K. (2011). *Discrete mathematics and its applications* (7th ed.). New York, NY: McGraw-Hill Science.

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.

Rupp, A. A., Templin, J., & Henson, R. A. (2010). *Diagnostic measurement: Theory, methods, and applications*. New York, NY: The Guilford Press.

Stocking, M. L. (1993). *Controlling item exposure rates in a realistic adaptive testing paradigm*. Princeton, NJ: Educational Testing Service.

Sympson, J. B., & Hetter, R. D. (1985, October). *Controlling item-exposure rates in computerized adaptive testing*. Paper presented at the 27th Annual meeting of the Military Testing Association, San Diego, CA.

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

Tatsuoka, C., & Ferguson, T. (2003). Sequential classification on partially ordered sets. *Journal of the Royal Statistical Society, Series B: Statistical Methodology*, *65*, 143-157.

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

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

Urry, V. W. (1971). *A Monte Carlo investigation of logistic mental test models*. Available from ProQuest Information & Learning.

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

van der Linden, W. J., & Veldkamp, B. P. (2004). Constraining item exposure in computerized adaptive testing with shadow tests. *Journal of Educational and Behavioral Statistics*, *29*, 273-291.

Wang, C., Chang, H. H., & Douglas, J. (2012). Combining CAT with cognitive diagnosis: A weighted item selection approach. *Behavior Research Methods*, *44*, 95-109.

Wang, C., Chang, H. H., & Huebner, A. (2011). Restrictive stochastic item selection methods in cognitive diagnostic computerized adaptive testing. *Journal of Educational Measurement*, *48*, 255-273.

Wang, C., Zheng, C., & Chang, H. H. (2014). An enhanced approach to combine item response theory with cognitive diagnosis in adaptive testing. *Journal of Educational Measurement*, *51*, 358-380.

Wang, S., Zheng, Y., Zheng, C., Su, Y.-H., & Li, P. (2016). An automated test assembly design for a large-scale Chinese Proficiency Test. *Applied Psychological Measurement*, *40*, 233-237.

Weiss, D. J. (1974). *Strategies of adaptive ability measurement*. Minneapolis: University of Minnesota.

Yi, Q., & Chang, H. H. (2003). a-Stratified CAT design with content blocking. *British Journal of Mathematical and Statistical Psychology*, *56*, 359-378.

Zheng, C. (2015). *Some practical item selection algorithms in cognitive diagnostic computerized adaptive testing—Smart diagnosis for smart learning* (Unpublished doctoral dissertation). University of Illinois at Urbana–Champaign.

Zheng, C., & Chang, H.-H. (2016). High-efficiency response distribution–based item selection algorithms for short-length cognitive diagnostic computerized adaptive testing. *Applied Psychological Measurement*, *40*, 608-624.