The Application of the Monte Carlo Approach to Cognitive Diagnostic Computerized Adaptive TestingWith Content Constraints

Xiuzhen Mao and Tao Xin

Applied Psychological Measurement published online 1 May 2013
DOI: 10.1177/0146621613486015

Article

Applied Psychological Measurement
XX(X) 1–15

DOI: 10.1177/0146621613486015
apm.sagepub.com
SAGE

The Application of the Monte Carlo Approach to Cognitive Diagnostic Computerized Adaptive Testing With Content Constraints

Xiuzhen Mao[1,2] and Tao Xin[2]

Abstract

The Monte Carlo approach which has previously been implemented in traditional computerized adaptive testing (CAT) is applied here to cognitive diagnostic CAT to test the ability of this approach to address multiple content constraints. The performance of the Monte Carlo approach is compared with the performance of the modified maximum global discrimination index (MMGDI) method on simulations in which the only content constraint is on the number of items that measure each attribute. The results of the two simulation experiments show that (a) the Monte Carlo method fulfills all the test requirements and produces satisfactory measurement precision and item exposure results and (b) the Monte Carlo method outperforms the MMGDI method when the Monte Carlo method applies either the posterior-weighted Kullback–Leibler algorithm or the hybrid Kullback–Leibler information as the item selection index. Overall, the recovery rate of the knowledge states, the distribution of the item exposure, and the utilization rate of the item bank are improved when the Monte Carlo method is used.

Keywords

computerized adaptive testing, cognitive diagnostic, content constraints, the Monte Carlo approach

Computerized adaptive testing (CAT) is a test model that combines test theory and computer technology. CAT has received extensive attention because it has several outstanding features. It tailors a test to the latent trait level of each examinee, which not only shortens the test length but also provides a higher expected measurement precision compared with a traditional test. CAT also has the ability to supply different forms of assessment to the examinee based on the implementation of different test theories. For example, when the Item Response Theory (IRT)

[1]Sichuan Normal University, Chengdu, China
[2]Beijing Normal University, Beijing, China

Corresponding Author:
Tao Xin, Institute of Developmental Psychology, Beijing Normal University, Beijing 100875, China.
Email: xintao@bnu.edu.cn

is used in CAT, examinees can receive either a summative score that presents their total ability level or several summative subscale scores that denote their performance on each individual proficiency profile. As another example, when Cognitive Diagnostic Theory (CDT) is used in CAT, examinees can receive more detailed diagnostic information regarding their mastery of every attribute. Therefore, it is possible to conduct a remedy instruction that is based on both the hierarchical structure of attributes and the estimates of examinees' knowledge states.

The measurement precision of CAT should be as accurate as possible to ensure that reliable inferences or decisions are made. In addition to the response to the item, the validity of the test also has an influence on the measurement precision. It is known that many factors contribute to the test validity, such as (a) the test length; (b) the word count; (c) the proportions of the content domains in the test; (d) the inclusion of ''enemy items'', in which one enemy item provides a hint about the answer to the other items; and (e) the balance of item keys in the test. For a paper-and-pencil test, it is easy to meet content constraints and guarantee test validity by assembling items according to the test blueprint. However, items in CAT are usually selected individually in order and cannot be completely assigned beforehand. Therefore, the performance of different item selection methods in CAT when nonstatistical constraints are present is a worthwhile investigation because the item selection method that is used may improve the validity of test. There have been many item selection strategies used in traditional CAT when content constraints are present. However, reports that have analyzed the item selection rules in cognitive diagnostic CAT (CD-CAT) have focused mainly on improvements to test precision or to the control of item exposure (Cheng, 2009; McGlohen & Chang, 2008; Wang, Chang, & Huebner, 2011; Xu, Chang, & Douglas, 2003). Except for a study by Cheng (2010), there are few reports on the impact of the item selection method in CD-CAT when content requirements are present.

Item Selection Methods in Traditional CAT With Content Constraints

There are three major item selection method types that have been implemented in traditional CAT when content constraints are presented. The first type uses a heuristic algorithm that selects items in a restrictive way from different item subsets (Moyer, Galindo, & Dodd, 2012). Methodologies of this kind include the constrained CAT (Kingsbury & Zara, 1991), the modified multinomial model (S. Y. Chen & Ankenmann, 1999), and the modified constrained CAT (Leung, Chang, & Hau, 2003). This set of rules is easy to implement and suitable for tests that are composed of a few properly proportioned content categories. Unfortunately, once the number of content categories is increased, the item selection process will become increasingly difficult. Moreover, some of the constraints may be violated due to sampling error during the real testing process. The second kind of item selection methods that have been used in CAT applies a heuristic algorithm that assigns items according to item weights. These item weights are constructed by combining the item information with the goal of the constraints. The second type of item selection methods include the weighted deviation model (Stocking & Swanson, 1993), the maximum priority index method (Cheng & Chang, 2009), and the constraint-weighted *a*-stratification method (Cheng, Chang, Douglas, & Guo, 2009). This method type avoids the computational complexity and infeasibility issues that are encountered during the process of item selection. However, these item selection methods cannot ensure that all the requirements of the test are fulfilled and may even decrease the measurement precision (Cheng & Chang, 2009). The third kind of item selection methods is the mathematical programming method. The shadow-test method (van der Linden, 2000; van der Linden & Reese, 1998), the multiple shadow-test method (van der Linden, 2005), and the Monte Carlo approach (Belov, Armstrong, & Weissman, 2008) are instances of this kind. Mathematical programming methods single out test items from a number of shadow tests that have been assembled in advance of the test. As a

result, these mathematical programming methods select the approximately optimal item with which all the constraints can be satisfied when the test is over. However, the computation will be very complex and will possibly have no solution if there are too many constraints.

Particularly, it is necessary for the shadow-test method to assemble a shadow test prior to the selection of each item. The assembled shadow test has the following characteristics: (a) It includes all of the administered items, (b) it fulfills all of the content constraints, and (c) it maximizes the test information at the current proficiency level of the examinee. Then, the most informative one at the estimated proficiency level from all the free items of this shadow test is assigned for administration. Although all the content constraints could be satisfied using the shadow-test method, Belov et al. (2008) noted that this method produced an uneven distribution of item exposures and had a low utilization rate of the total number of items in the item bank. Furthermore, Belov et al. defined the fundamental problem of CAT (FPCAT) when content requirements are present as finding the next item φ such that the item information at $\hat{\theta}$ is maximized and all the constraints are fulfilled at the end of the test. Therefore, it usually declines the precision a little when the shadow-test method is used as the solution of the FPCAT may be not the item that is selected by the shadow-test method.

To address the FPCAT in traditional CAT when content constraints are present, Belov et al. (2008) proposed the Monte Carlo method. Furthermore, they proved that the Monte Carlo method would converge to the solution of the FPCAT if the number of available items was large enough. They also presented many other advantages of the Monte Carlo approach over the shadow-test method. For example, the Monte Carlo method produced a more robust estimate of proficiency level compared with the shadow-test method when an examinee with an extremely high (or low) ability performed abnormally at the beginning of the test. In addition, the Monte Carlo method not only provided a lower maximum item exposure rate and a higher rate of utilization of the item bank but also shortened operation time (Belov et al., 2008).

Item Selection Rules for CD-CAT With Content Constraints

The modified maximum global discrimination index (MMGDI) method which was proposed by Cheng (2010) is a heuristic algorithm that chooses items based on item weights. The MMGDI method aims to balance attribute coverage by selecting the item with the largest modified global discrimination index (MGDI) sequentially from the remaining items in the pool. In other words, if the test is presumed to contain at least B_k items that test the kth ($k = 1, 2, \ldots, K$) attribute and b_k items that measure this attribute have already been administered prior to the selection of the tth item, then the item selection rule can be described as

$$i_t \equiv \arg\max \left\{ \prod_{k=1}^{K} \left(\frac{B_k - b_k}{B_k} \right)^{q_{kj}} \cdot \mathrm{GDI}_j(\hat{\alpha}^{t-1}),\ j \in R_t \right\}. \tag{1}$$

Here, q_{kj} denotes whether item j measures the kth attribute, and R_t refers to the remaining items in the item bank. The global discrimination index (GDI) of item j at $\hat{\alpha}^{t-1}$ is calculated by

$$\mathrm{GDI}_j\left(\hat{\alpha}^{t-1}\right) = \sum_{c=1}^{2^K} \left(\sum_{x=0}^{1} \log \left(\frac{P\left(X_j = x | \hat{\alpha}^{t-1}\right)}{P\left(X_j = x | \alpha_c\right)} \right) P\left(X_j = x | \hat{\alpha}^{t-1}\right) \right). \tag{2}$$

It usually confines the available item selection area and leads to some losses in measurement precision that impose some content constraints on traditional CAT. However, it may be another case for CD-CAT when constraints that require the balance of the attribute coverage are

present. As a matter of fact, the attribute mastery pattern can be categorized exactly only when all the attributes are correctly classified. Hence, if there are enough items that measure each attribute and attribute coverage keeps balanced, the recovery rate of the entire pattern is likely to increase. So, it is not surprising that the MMGDI method performs better than the original maximum GDI method both in satisfying the requirements of the attribute coverage and in improving measurement precision (Cheng, 2010). However, it is worthy of noticing that when the number of attributes measured by item j increased, the index $\prod_{k=1}^{K}[(B_k - b_k)/B_k]^{q_{kj}}$ decreases if some of the b_ks are lower than the corresponding B_ks ($k = 1, 2, \ldots, K$). Thus, for items whose GDI information are equal, the MMGDI method is likely to select the item that measures fewer attributes, which gives rise to an uneven distribution of item exposures. Furthermore, the MMGDI method will become more complicated and even impractical if the administrator includes additional test requirements, such as well-proportioned content domains and an appropriate balance of answer keys.

It is known that content balancing and attribute balancing have different meanings in psychometrics. Furthermore, the shadow-test and the Monte Carlo methods were put forward mainly for handling nonstatistical requirements, especially for content constrains (Belov et al., 2008; van der Linden, 2005). Therefore, the present study investigates the performance of the Monte Carlo method in CD-CAT when content constraints other than only attribute balancing are present based on the drawbacks of the shadow-test and MMGDI methods that have been described above and the advantages of the Monte Carlo method that were introduced in Belov et al. (2008). The remainder of this article is organized as follows. The section ''Cognitive Diagnostic Model (CDM) and Item Selection Rules'' introduces the CDM and the item selection methods that were used in this study. Then, ''The performance of the Monte Carlo Method in CD-CAT With Content Constraints'' and ''A Comparison of the Monte Carlo and MMGDI Methods'' sections present two simulation studies. The final section presents the conclusions and discussion.

CDM and Item Selection Rules

Deterministic Inputs, Noisy "And" Gate (DINA) Model

Generally, the relationship among the item characteristic, the item responses, and the knowledge states is constructed by the CDM. It follows that the CDM is a basis and a kernel of the CDT that is used to analyze the profiles of the examinees. Until now, researchers have proposed a great many CDMs. The DINA model (Haertel, 1989; Junker & Sijtsma, 2001; Macready & Dayton, 1997) is an example of the stochastic conjunctive latent class model. Although all the attributes that are specified by an item are required, the mastery of all the attributes does not always result in a correct response. Conversely, lacking at least one of them does not necessarily result in an incorrect response. Accordingly, the DINA model simulates random circumstances during the test by slipping and guessing parameters. The parameter s_j represents the probability of a careless mistake and incorrectly answered item j for a capable examinee and g_j is the probability of responding to item j correctly because of guessing when $\prod_{k=1}^{K}\alpha_{ik}^{q_{jk}} = 0$. That is, $s_j = P(X_{ij} = 0 | \prod_{k=1}^{K}\alpha_{ik}^{q_{jk}} = 1)$ and $g_j = P(X_{ij} = 1 | \prod_{k=1}^{K}\alpha_{ik}^{q_{jk}} = 0)$. The entry q_{jk} indicates whether item j measures attribute k. If the correct application of attribute k influences the probability of answering item j, then the entry q_{jk} will equal 1; otherwise, the entry q_{jk} will equal 0. The variable α_{ik} equals 1 when examinee i has mastered attribute k, and equals 0 otherwise. As a result, the item response function (IRF) of the DINA model is expressed as

$$P\left(X_{ij}=1|\alpha_i,s_j,g_j\right)=\left(1-s_j\right)^{\prod_{k=1}^{K}\alpha_{ik}^{q_{jk}}}\cdot g_j^{\left(1-\prod_{k=1}^{K}\alpha_{ik}^{q_{jk}}\right)}. \quad (3)$$

Here, $\prod_{k=1}^{K}\alpha_{ik}^{q_{jk}}$ reflects whether examinee i has mastered all the attributes that are required by item j.

From Equation 3, it can be inferred that the probability of a correct response for item j is either $1-s_j$ or g_j. It involves two parameters s_j and g_j for item j and these parameters are easy to interpret. Furthermore, de la Torre (2009b) introduced a technique to estimate the DINA model parameters using either the expectation-maximization or Markov chain Monte Carlo algorithm. Therefore, the DINA model has been extensively researched and applied (de la Torre, 2008; de la Torre, 2009a; de la Torre & Douglas, 2004, 2008; de la Torre & Karelitz, 2009). In addition, the DINA model has also been applied in CD-CAT (P. Chen, Xin, Wang, & Chang, 2012; Cheng, 2009, 2010). Hence, the DINA model was chosen as the CDM for the following simulations due to the preceding virtues. Except the CDM, the item selection method is another important component of CD-CAT, and the Monte Carlo method put forward by Belov et al. (2008) was used in the present study.

The Monte Carlo Method for Item Selection

The main idea of the Monte Carlo method is to construct many shadow tests uniformly prior to the selection of each item. Each shadow test must include all the items that have already been presented and meet all of the constraints. To determine which item to administer next, the most informative item from all of the free items within these shadow tests is found. The processes of item selection are divided into input and output in the same way described by Belov et al. (2008). The former includes the set of t administered items, the current cognitive profile $\hat{\alpha}^t$, the set S that is composed of shadow tests, and the parameters m and r. Parameter m equals the number of shadow tests that are required to be assembled beforehand, and parameter r equals the number of shadow tests already in set S. The latter contains the selected item φ and the renewed set S.

Generally speaking, the steps that are taken to choose the $(t+1)$th item can be described as follows. Step 1: Assemble $m-r$ shadow tests uniformly to guarantee that all the shadow tests have an equal chance to be chosen. For example, suppose the test length is L, this can be done by constructing a test first which composed of all the t utilized items and $(L-t)$ items that are chosen randomly. Then, keep the test if all the content constraints are fulfilled. Otherwise, the test is discarded. The procedure is repeated until $m-r$ eligible tests have been assembled. Step 2: Add these $m-r$ shadow tests to set S and get m shadow tests in set S altogether. Step 3: Draw all the items that have not yet been administered from every test in set S. So, there are $m(L-t)$ items and some items may appear many times. Step 4: Assign the most informative one at knowledge state $\hat{\alpha}^t$ from all the $m(L-t)$ items to φ. Step 5: Renew the set S by keeping all the shadow tests that contain item φ.

The Monte Carlo method used here and the method used in Belov et al. (2008) differ mainly in Steps 3 and 4. For one thing, although the study by Belov et al. implements a method that choose the first $2t+1$ of the $m(L-t)$ available items, the present study decided to choose items from all $m(L-t)$ items which is always more than $2t+1$ items because increasing the available items will generally improve the measurement precision. For another, the differences in Step 4 are resulted from the different measurement theory used in the two studies. However, there are many common properties between the present study and the study by Belov et al. First, Belov et al. proved the Monte Carlo approach would converge to the solution of the

FPCAT if $m(L - t)$ tended to infinity. The demonstration has only depended on the item bank size and the parameters m, L, and t. Hence, it is justifiable to draw the same conclusion in CD-CAT. Although $m(L - t)$ is expected to be infinite, m is not required to be infinite in practice because multiple shadow tests may contain the optimal item. Second, it certainly ensures an equal probability of choosing each item from the remaining items in the item bank by assembling test uniformly. Hence, the Monte Carlo method can decrease the maximum item exposure rate. Third, it assures to meet all the constraints by way of assembling shadow tests. Last, it will definitely lengthen the operation time of the Monte Carlo method, even make it unsolvable if there are not enough items in the item bank.

Study 1: The Performance of the Monte Carlo Method in CD-CAT With Content Constraints

For each examinee, CAT includes the following steps: (a) choosing an initial value of the latent trait, (b) selecting the most appropriate item from the remaining items in the item bank according to the estimated latent trait value, (c) estimating the potential trait level on the basis of his (her) responses to those presented items, and (d) repeating steps (b) and (c) until the termination rule is satisfied. It follows that the examinees, the generation of the item bank, the item selection strategy, the method of parameter estimation, and the termination rule are essential components for simulating CAT. The following section details the data generating procedures and specifications for each simulation. The following simulation experiments used CAT codes that were written in the MATLAB (version 7.10.0.499).

Method

Generation of the Item Bank. Study 1 simulated an item bank that included 300 items. This number of items was chosen to satisfy the rule that the pool needs to contain at least 12 times as many items as the test contains (Stocking, 1994); however, Chang and Zhang (2002) advised that even larger ratios should be used. These items in the item bank measured six independent attributes.

The attributes that were measured by each item were determined using the same method as Cheng (2009, 2010). That is to say, each item tested at least one attribute and the probability of measuring every attribute was 0.2. For example, item j (j = 1, 2, . . . , 300) would measure attribute k (k = 1, 2, . . . , 6) if 0.2 was larger than a random w_{jk} generated from a uniform distribution on the interval (0, 1), else item j would not measure attribute k. As a result, the **Q**-matrix was generated item by item and attribute by attribute.

The results from de la Torre and Douglas (2004) showed that most of the item slipping and guessing parameters were smaller than 0.25. In addition, the assumption that item parameters between 0.05 and 0.25 were equally likely has been used in Cheng (2010) and as a special case in P. Chen et al. (2012). Therefore, it is a reasonable assumption to draw both the item slipping and guessing parameters from a uniform distribution on the interval (0.05, 0.25). There were 300 pairs of item parameters in all.

Examinees and Item Responses. The study assumed that every examinee had a 0.5 chance to master each attribute. Hence, there were 64 knowledge states that appeared equally in the population. Study 1 generated 2,000 examinees in total. Then, the item responses of these 2,000 examinees were obtained through the following steps: (a) applying the DINA model to compute the correct response probability of examinee i (i = 1, 2, . . . , 2,000) to item j (j = 1, 2, . . . , 300); (b) extracting p_{ij} from a uniform distribution on the interval (0, 1); and (c) assigning a value of 1

to the corresponding response if the correct response probability was larger than p_{ij} and assigning a value of 0 otherwise.

Item Selection Methods. The Monte Carlo method that was presented in Section 2 was utilized in Study 1 with the following constraints: (a) The test length was 20; (b) there were no less than 3 items that tested each attribute; (c) there were constraints on the answer key that, for example, an item had four choices with only one correct answer key and the number of each answer key of the test would be between 3 and 7; and (d) the 37th and 189th items were chosen randomly as ''enemy items''. When items are labeled as enemy items, the two items are restricted from appearing in the same test, as one enemy item shows a clue to the solution of the other enemy item (Finkelman, Kim, Roussos, & Verschoor, 2010). In addition, the parameter m to select the tth item was determined according to

$$m = \begin{cases} \max([(2t+1)/(20-t)]), 1 & \text{if } t < 20; \\ 41 & \text{otherwise.} \end{cases} \tag{4}$$

In Equation 4, $[(2t+1)/(20-t)]$ denotes the nearest whole number that is larger than $(2t+1)/(20-t)$, which is computed in the same way as in Belov et al. (2008). Specially, the parameter m equals $2t+1$ when the last item is selected because in that case $20 - t$ equals zero.

The item information function is another essential component of the Monte Carlo method. According to a study by Cheng (2009), the posterior-weighted Kullback–Leibler (KL) algorithm (PWKL) and the hybrid KL (HKL) index considerably improved the measurement precision of the KL information; both the PWKL algorithm and HKL index also outperformed the Shannon entropy in both measurement precision and computation time. Hence, the KL information, the PWKL algorithm and the HKL index were each used with the Monte Carlo method to compute the item selection function in three separate simulation experiments. The expression for item j at knowledge state $\hat{\alpha}^{t-1}$ for each of these methods is given by

$$\mathrm{KL}_j(\hat{\alpha}^{t-1}) = \sum_{c=1}^{2^K} \left[\sum_{x=0}^{1} \log\left(\frac{P(X_j = x|\hat{\alpha}^{t-1})}{P(X_j = x|\alpha_c)} \right) \cdot P(X_j = x|\hat{\alpha}^{t-1}) \right]; \tag{5}$$

$$\mathrm{PWKL}_j(\hat{\alpha}^{t-1}) = \sum_{c=1}^{2^K} \left\{ \left[\sum_{x=0}^{1} \log\left(\frac{P(X_j = x|\hat{\alpha}^{t-1})}{P(X_j = x|\alpha_c)} \right) \cdot P(X_j = x|\hat{\alpha}^{t-1}) \right] \times \pi_{i,t}(\alpha_c) \right\}; \tag{6}$$

$$\mathrm{HKL}_j(\hat{\alpha}^{t-1}) = \sum_{c=1}^{2^K} \left\{ \left[\sum_{x=0}^{1} \log\left(\frac{P(X_j = x|\hat{\alpha}^{t-1})}{P(X_j = x|\alpha_c)} \right) \cdot P(X_j = x|\hat{\alpha}^{t-1}) \right] \times \pi_{i,t}(\alpha_c) \times \frac{1}{d(\hat{\alpha}^{t-1}, \alpha_c)} \right\}. \tag{7}$$

If the prior distribution of cognitive profile $\alpha_c (c = 1, 2, \ldots, 2^K)$ is signified by $\pi_{i,0}(\alpha_c)$, the prior distribution is constrained to $\pi_{i,0}(\alpha_c) \geq 0$ and $\sum_{c=1}^{2^K} \pi_{i,0}(\alpha_c) = 1$. Then, the posterior distribution after the first t responses is proportional to $\pi_{i,0}(\alpha_c) L(u^t|\alpha_c)$ and can be denoted as $\pi_{i,t}(\alpha_c)$. If X_j represents the response to item j, the likelihood function $L(u^t|\alpha_c)$ is equal to $\prod_{j=1}^{t} [P_j(\alpha_c)]^{X_j} [1 - P_j(\alpha_c)]^{1-X_j}$. The expression $d(\hat{\alpha}^{t-1}, \alpha_c)$ in Equation 7 represents the Euclidean distance between knowledge states $\hat{\alpha}^{t-1}$ and α_c; that is $d(\hat{\alpha}^{t-1}, \alpha_c) = \sqrt{\sum_{k=1}^{K} (\hat{\alpha}_k^{t-1} - \alpha_{ck})^2}$. A larger value of $d(\hat{\alpha}^{t-1}, \alpha_c)$ indicates a greater distance between knowledge states.

Estimation of the Knowledge States. The initial value of the knowledge states was chosen randomly from the 64 knowledge states. The interim knowledge states were estimated using the maximum likelihood estimation (MLE) method.

After the test, the Monte Carlo method using the different item information indices, Equations 5, 6, and 7, was evaluated according to the following criteria.

Measures of Performance. The dependent variables are the recovery rates of each attribute and the entire knowledge state. Let n_k denote the number of examinees whose attribute k was classified correctly as mastery or nonmastery and N represent the total number of examinees. In addition, let n equal the number of examinees who were classified into the correct attribute mastery pattern. Then, the recovery rates of attribute k and the entire cognitive state were calculated by n_k/N and n/N, respectively.

The item exposure rate represents the frequency of utilization and equals the ratio of the number of times the item is used to the total number of examinees (Chang & Ying, 1999). To quantify the equalization of exposure rates, the following indices were used: (a) the quartiles of the item exposure rates; (b) the number of unexposed items; (c) the number of overexposed items, which had exposure rates larger than 0.2; (d) the χ^2 statistic; and (e) the test overlap rate. The term er_j and $\overline{er}$ represent the exposure rate of item j and the expected exposure rate, respectively. Here, $\overline{er}$ equals the ratio of the test length L to the size of the item bank M. Then, χ^2 statistic is calculated by

$$\chi^2 = \sum_{j=1}^{M} \frac{\left(er_j - \overline{er}\right)^2}{\overline{er}}. \tag{8}$$

In Equation 8, a smaller value of χ^2 represents less discrepancy between the observed and expected item exposure rates. The test overlap rate is defined as the proportion of the expected number of overlapping items between two randomly selected examinees to the test length (Chang & Ying, 1999). To simplify the computation and to avoid time-consuming pairwise comparisons of common items that are administered to both examinees, the following equation

$$\hat{T} = \frac{L}{M} S_{er}^2 + \frac{M}{L}, \tag{9}$$

which was originally proposed by S. Y. Chen, Ankenmann, and Spray (2003), was applied to calculate the test overlap rate. Here, L denotes the test length, M denotes the item bank size, and S_{er}^2 is the variance of the item exposure rates. A smaller test overlap rate represents better control of item exposure by the item selection method.

Results

Tables 1 and 2 report the frequency statistics for the **Q**-matrix and the 2,000 true cognitive states. Not surprisingly, the number of items per attribute was slightly larger than the expected value of 60, because each item was required to test at least one attribute.

Table 3 illustrates the measurement precision of the Monte Carlo method using different item selection indices. Both the PWKL algorithm and HKL index outperformed the original KL information, as demonstrated by the level of improvement that was observed in the correct classification rates of each attribute and in the entire cognitive pattern when these indices were applied. However, the difference between the whole pattern recovery rate of the PWKL algorithm and that of the HKL index was negligible.

Table 1. The Number of Items That Test (or the Number of Examinees That Master) Each Attribute.

	Attributes					
	1	2	3	4	5	6
Number of items	70	82	84	78	86	80
Number of examinees	1,020	1,001	1,028	996	1,012	1,042

Table 2. The Number of Items That Test (or the Number of Examinees That Master) Each Possible Number of Attributes.

	Number of attributes						
	0	1	2	3	4	5	6
Number of items	0	163	101	30	5	1	0
Number of examinees	26	192	454	587	508	192	41

Table 3. Recovery Rates Using the Monte Carlo Method.

	Attributes						
Index	1	2	3	4	5	6	Whole pattern
KL	0.97	0.98	0.96	0.97	0.95	0.96	0.83
PWKL	0.99	0.99	0.99	0.99	0.98	0.99	0.94
HKL	0.99	0.99	0.98	0.99	0.98	0.99	0.93

Note: KL = Kullback–Leibler; PWKL = posterior-weighted Kullback–Leibler; HKL = hybrid Kullback–Leibler.

The results of item exposure are summarized in Table 4, which showed that (a) when the KL information was used, the maximum item exposure rate was 0.72 and approximately one fourth of the items were not used and (b) the largest item exposure rates that were observed when the PWKL and HKL criteria were used were 0.5 and 0.49, respectively, and the utilization rate of the item bank was more than 99% for both indices. Therefore, it was not surprising that both the test overlap rate and the χ^2 statistic for the PWKL and the HKL indices were smaller than the test overlap rate and the χ^2 statistic for the KL information. The distribution of the item exposure that was generated by the KL information was more scattered when compared with the distributions of the exposures generated by the PWKL or the HKL index. In addition, there was almost no difference between the distributions of item exposures produced by the PWKL and HKL indices.

Study 2: A Comparison of the Monte Carlo and MMGDI Methods

Method

In Study 2, the item bank and examinees, as well as the methods for generating the item responses and estimating the cognitive states were the same as in Study 1. However, the two simulation studies differed in the item selection methods and the test requirements. In Study 2,

Table 4. Results of Item Exposure Using the Monte Carlo Method.

		Exposure rate							
Index	NU	Minimum	Q_1	Q_2	Q_3	Maximum	NO	TOR	χ^2
KL	78	0.000	0.000	0.010	0.100	0.720	33	0.264	59.010
PWKL	3	0.000	0.010	0.030	0.100	0.500	30	0.185	35.330
HKL	0	0.001	0.010	0.030	0.100	0.490	26	0.180	33.870

Note: NU = number of unused items; NO = number of overexposed items; TOR = test overlap rate; KL = Kullback–Leibler; PWKL = posterior-weighted Kullback–Leibler; HKL = hybrid Kullback–Leibler.

Table 5. Recovery Rates for Each Attribute and the Whole Pattern Using Different Methods.

		Attributes						
Method	Index	1	2	3	4	5	6	Whole pattern
MMGDI	MGDI	0.99	0.99	0.99	0.99	0.98	0.99	0.92
MC	KL	0.98	0.99	0.98	0.97	0.97	0.98	0.88
MC	PWKL	0.99	0.99	0.99	0.99	0.98	1.00	0.97
MC	HKL	0.99	0.99	0.98	1.00	0.98	0.99	0.96

Note: MMGDI = modified maximum global discrimination index; MGDI = modified global discrimination index; MC = the Monte Carlo method; KL = Kullback–Leibler; PWKL = posterior-weighted Kullback–Leibler; HKL = hybrid Kullback–Leibler.

the test length was 24 and each attribute was measured by no less than 4 items. The Monte Carlo and MMGDI methods were applied to select items. Parameter m in the Monte Carlo method was determined using a similar equation used in Study 1, defined as

$$m = \begin{cases} \max([(2t+1)/(24-t)], 1) & \text{if } t < 22; \\ 25 & \text{otherwise.} \end{cases} \quad (10)$$

Here, m equals 25 for the last three items for two reasons: (a) the measurement precision is much higher in the later stages of the test and (b) the resulting decrease the number of shadow tests in set S not only has negligible effects on the measurement precision but also shortens the operation time. For Study 2, the KL information, the PWKL algorithm, and HKL index were used to compute the item information index in the Monte Carlo method. Therefore, four experiments were conducted in Study 2 to compare the performances of the Monte Carlo and MMGDI methods.

Results

Table 5 presents the recovery rates of the Monte Carlo and the MMGDI methods. The correct classification rates of both single attribute and the whole pattern were uniformly lower when the Monte Carlo method using the KL information was used compared with the other methods that were tested. However, the recovery rates of individual skills were almost equal for the other three methods. Furthermore, the recovery rate of the attribute mastery pattern was 0.97 and 0.96 when the Monte Carlo approach using either the PWKL algorithm or the HKL index, respectively, was applied, while the recovery rate of the attribute mastery pattern was 0.92 when the

Table 6. Results of Item Exposure Using Different Methods.

Method	Index	NU	Exposure rate					NO	TOR	χ^2
			Minimum	Q_1	Q_2	Q_3	Maximum			
MC	KL	80	0.000	0.000	0.010	0.010	0.720	53	0.341	82.900
MC	PWKL	2	0.000	0.010	0.030	0.110	0.570	35	0.239	52.190
MC	HKL	1	0.000	0.010	0.030	0.120	0.570	36	0.233	50.420
MMGDI	MGDI	203	0.000	0.000	0.000	0.020	1.000	33	0.765	209.530

Note: NU = number of unused items; NO = number of overexposed items; TOR = test overlap rate; MC = the Monte Carlo method; KL = Kullback–Leibler; PWKL = posterior-weighted Kullback–Leibler; HKL = hybrid Kullback–Leibler; MMGDI = modified maximum global discrimination index; MGDI = modified global discrimination index.

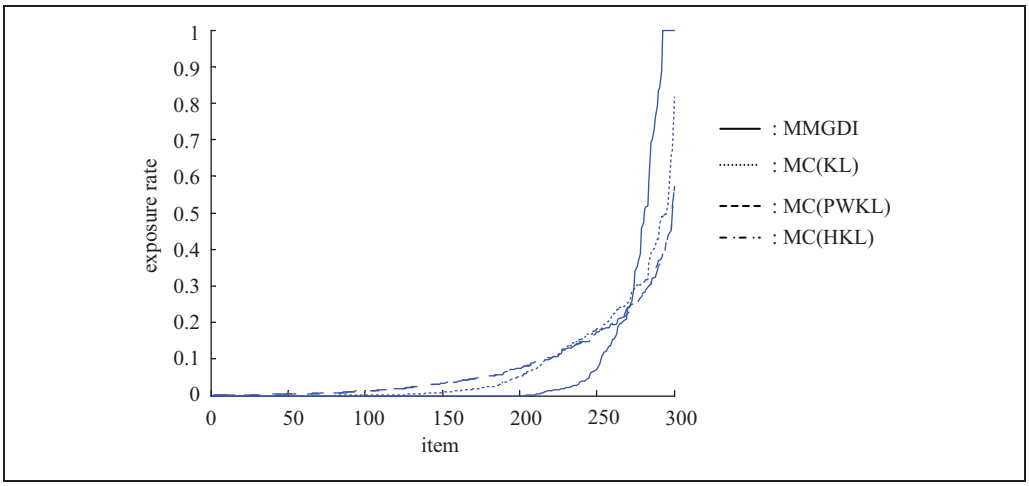

Figure 1. The distributions of exposure rates for different item selection methods.
Note: MMGDI = modified maximum global discrimination index; MC = Monte Carlo method; KL = Kullback–Leibler; PWKL = posterior-weighted Kullback–Leibler; HKL = hybrid Kullback–Leibler; MC(KL) = the Monte Carlo method using the KL information; MC(PWKL) = the Monte Carlo method using the PWKL algorithm; MC(HKL) = the Monte Carlo method using the HKL index.

MMGDI method was applied. Therefore, the Monte Carlo method performed remarkably better than the MMGDI method in measurement precision when the Monte Carlo method was used with either the PWKL algorithm or the HKL index.

It can be easily inferred from Table 6 that the Monte Carlo approach produced a more even distribution of item utilization rates compared with the distribution of item utilization rates produced by the MMGDI method. This advantage of the Monte Carlo method was even more evident when the PWKL algorithm or the HKL index was used. Furthermore, the Monte Carlo method applying the PWKL algorithm or the HKL index had comparable distributions of item utilization rates. The results of the Monte Carlo method using either the PWKL algorithm or the HKL index and the results of the MMGDI method illustrate the following differences between these methods: (a) The largest item exposure rate when the MMGDI method was applied was as high as 1.00 compared with the largest item exposure rate of 0.57 when the Monte Carlo method was applied; (b) the utilization rate of the item bank when the MMGDI method was applied was approximately 0.33, while that generated by the Monte Carlo method was more than 0.99; and

Table 7. The Number (Proportions) of Administered Items for Each Method Organized by the Possible Number of Attributes Per Item.

		Possible number of attributes per item					
Method	Index	1	2	3	4	5	6
MMGDI	MGDI	31249 (0.78)	6,139 (0.17)	1,885 (0.05)	27 (0.00)	0	0
MC	KL	17,273 (0.43)	17,525 (0.44)	4,954 (0.12)	245 (0.01)	3 (0.00)	0
MC	PWKL	17,448 (0.44)	16,815 (0.42)	5,431 (0.13)	283 (0.01)	23 (0.00)	0
MC	HKL	16,662 (0.42)	17,164 (0.43)	5,794 (0.15)	355 (0.01)	25 (0.00)	0

Note: MMGDI = modified maximum global discrimination index; MGDI = modified global discrimination index; MC = the Monte Carlo method; KL = Kullback–Leibler; PWKL = posterior-weighted Kullback–Leibler; HKL = hybrid Kullback–Leibler.

(c) the test overlap rate and the χ^2 statistic when the MMGDI method was applied were much larger than the corresponding values when the Monte Carlo method was applied. In summary, the item exposure rates were more evenly distributed when the Monte Carlo method was applied compared with when the MMGDI method was applied. This conclusion can also be found easily from Figure 1, which plotted the distributions of all item exposure rates in ascending order.

In addition, Table 7 illustrates that when the MMGDI method was used to select items, 78% of the administered items measured only one attribute, 17% of the administered items measured only two attributes and approximately 5% of the administered items measured more than two attributes. However, for the Monte Carlo method, the proportion of items that measured one attribute and the proportion of items that measured two attributes were both approximately 44%. Therefore, the MMGDI method was likely to select items that tested fewer attributes, which resulted in an uneven distribution of item exposures. The result was consistent with the inferences that were made in the prior section. In a word, the Monte Carlo method produced a more even distribution of item exposures, a higher utilization rate of the item bank, and a lower test overlap rate compared with the MMGDI method.

Conclusion and Discussion

CD-CAT is a promising research area that has gained much attention because it integrates both the cognitive diagnostic method and adaptive testing. The item selection procedure that is used in CD-CAT needs to consider both statistical optimality and nonstatistical constraints. Prior item selection methods that have been implemented in CD-CAT have not performed very well with respect to nonstatistical constraints such as content constraints. Hence, this study applied the Monte Carlo method to CD-CAT because the Monte Carlo method was shown to perform better than other methods when nonstatistical constraints were present in traditional CAT. The two simulation studies that were conducted in this study showed that the Monte Carlo method not only handles content constraints but also produces satisfactory results about the item exposure and the recovery rate of the cognitive profile. The simulations also showed that the Monte Carlo method using either the PWKL algorithm or the HKL index performed better than did the MMGDI method in terms of both the measurement precision and the item exposure.

As noted earlier, items are chosen from a large number of shadow tests assembled uniformly. So there are only a few content constraints to guarantee that the Monte Carlo method has a solution. Moreover, the Monte Carlo method is not restricted to these constraints and can also be used to manage other requirements, such as balancing the response time among all of

the examinees and controlling the maximum item exposure rate. Factually, there may be hundreds or thousands of content specifications that are required to assemble a test (van der Linden, 2005). Increasing the number of test requirements usually leads to decreasing the number of eligible shadow tests that can be constructed from the item bank. Accordingly, if test developers increase the test constraints, it will usually lengthen the reaction time and even make the Monte Carlo method unsolvable. Therefore, the Monte Carlo method imposes stronger requirements on the quality of the item bank compared with the MMGDI method. To counteract this problem, it may be worthwhile to explore the effects of different values for m because there is currently no specific rule for m. Furthermore, it may be a feasible way to advance the Monte Carlo method by reducing the number of shadow tests needed to be assembled before the selection of each item.

Although the Monte Carlo method is not restricted to the DINA model and some meaningful results were obtained, the experimental design still needs to be improved further. For example, it supposes that all of the attributes are independent from each other and that all the knowledge states are equally likely. In fact, the attributes are always related and hierarchically organized. For example, Leighton, Gierl, and Hunka (2004) pointed out four major hierarchical structures among attributes: linear, convergent, divergent, and unstructured. Hence, the robustness of the Monte Carlo method should be examined under different attribute correlations and hierarchical structures.

There are several remaining open issues with the current methods that deserve further investigation. For example, in the current methods, shadow-test items are selected simultaneously. Hence, shadow test can be assembled by using algorithms for the optimal assembling test, such as binary programming (Finkelman et al., 2010) and genetic algorithm (Finkelman, Kim, & Roussos, 2009). So, it is worthy to consider how to apply these algorithms to CD-CAT. So is researching algorithms for the optimal assembling test.

Second, the results of two simulation studies show that the Monte Carlo method significantly decreases the maximum item exposure rate. However, the Monte Carlo method cannot increase the utilization rates of most items and the test overlap rates are still very high. Therefore, this method does not have perfect control over item exposures. In general, the CD-CAT is a relatively low-stake test and test security is not a major concern. However, the control of item exposure is potentially important when the CD-CAT is a high-stake test (Cheng, 2009). Moreover, item exposure not only affects the test security but also is a significant consideration in constructing the item bank. Therefore, it is necessary to explore mechanisms that control item exposure in CD-CAT. Finally, although the current study demonstrates that the Monte Carlo method works successfully in CD-CAT under idealistic situations, these simulated results will need to be confirmed in a real item bank. Actually, it is hard to find an available item bank with substantial number of items and a valid **Q**-matrix. Therefore, it is expected that a new cognitive diagnostic item bank will need to be constructed to investigate the performances of these item selection methods under more realistic circumstances.

Acknowledgment

The authors are grateful for the valuable comments on earlier versions of the manuscript from three anonymous reviewers.

Declaration of Conflicting Interests

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

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

References

Belov, D. I., Armstrong, R. D., & Weissman, A. (2008). A Monte Carlo approach for adaptive testing with content constraints. *Applied Psychological Measurement*, *32*, 431-446.

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

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

Chen, P., Xin, T., Wang, C., & Chang, H. H. (2012). Online calibration methods for the DINA model with independent attributes in CD-CAT. *Psychometrika*, *77*, 201-222.

Chen, S. Y., & Ankenmann, R. D. (1999, April). *Effects of practical constraints on item selection rules at the early stages of computerized adaptive testing*. Paper presented at the Annual Meeting of the American Educational Research Association, Montreal, Canada.

Chen, S. Y., Ankenmann, R. D., & Spray, J. A. (2003). The relationship between item exposure and test overlap in computerized adaptive testing. *Journal of Educational Measurement*, *40*, 129-145.

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

Cheng, Y. (2010). Improving cognitive diagnostic computerized adaptive testing by balancing attribute coverage: The modified maximum global discrimination index method. *Educational and Psychological Measurement*, *70*, 902-913.

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

de, la Torre, J. (2008). An empirically based method of Q-matrix validation for the DINA model: Development and applications. *Journal of Educational Measurement*, *45*, 343-362.

de, la Torre, J. (2009a). A cognitive diagnosis model for cognitively based multiple-choice options. *Applied Psychological Measurement*, *33*, 163-183.

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

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

de, la, Torre, J., & Douglas, J. (2008). Model evaluation and multiple strategies in cognitive diagnosis: An analysis of fraction subtraction data. *Psychometrika*, *73*, 595-624.

de, la, Torre, J., & Karelitz, T. M. (2009). Impact of diagnosticity on the adequacy of models for cognitive diagnosis under a linear attribute structure: A simulation study. *Journal of Educational Measurement*, *46*, 450-469.

Finkelman, M. D., Kim, W., & Roussos, L. (2009). Automated test assembly for cognitive diagnosis models using a genetic algorithm. *Journal of Educational Measurement*, *46*, 273-292.

Finkelman, M. D., Kim, W., Roussos, L., & Verschoor, A. (2010). A binary programming approach to automated test assembly for cognitive diagnosis models. *Applied Psychological Measurement*, *34*, 310-326.

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

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. (1991). A comparison of procedures for content-sensitive item selection in computerized adaptive testing. *Applied Measurement in Education*, *4*, 241-261.

Leighton, J. P., Gierl, M. J., & Hunka, S. M. (2004). The attribute hierarchy method for cognitive assessment: A variation on Tatsuoka's rule-space approach. *Journal of Educational Measurement*, *41*, 205-237.

Leung, C. K., Chang, H. H., & Hau, K. T. (2003). Incorporation of content balancing requirements in stratification designs for computerized adaptive testing. *Educational and Psychological Measurement*, *63*, 257-270.

Macready, G. B., & Dayton, C. M. (1977). The use of probabilistic models in the assessment of mastery. *Journal of Educational Statistics*, *2*, 99-120.

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

Moyer, E. L., Galindo, J. L., & Dodd, B. G. (2012). Balancing flexible constraints and measurement precision in computerized adaptive testing. *Educational and Psychological Measurement*, *72*, 629-648. doi:10.1177/0013164411431838

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

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

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). Norwell, MA: Kluwer.

van der Linden, W. J. (2005). *Linear models for optimal test design*. New York, NY: Springer-Verlag.

van der Linden, W. J., & Reese, L. M. (1998). A model for optimal constrained adaptive testing. *Applied Psychological Measurement*, *22*, 259-270.

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.

Xu, X. L., Chang, H. H., & Douglas, J. (2003, April). *A simulation study to compare CAT strategies for cognitive diagnosis*. Paper presented at the Annual Meeting of National Council on Measurement in Education, Montreal, Canada.