Exploration of Item Selection in Dual-Purpose Cognitive Diagnostic Computerized Adaptive Testing: Based on the RRUM

Applied Psychological Measurement

2016, Vol. 40(8) 625–640

© The Author(s) 2016

Reprints and permissions:

sagepub.com/journalsPermissions.nav

DOI: 10.1177/0146621616666008

apm.sagepub.com

Buyun Dai^{1,2}, Minqiang Zhang^{1,3}, and Guangming Li^{1,3}

Abstract

Cognitive diagnostic computerized adaptive testing (CD-CAT) can be divided into two broad categories: (a) single-purpose tests, which are based on the subject's knowledge state (KS) alone, and (b) dual-purpose tests, which are based on both the subject's KS and traditional ability level (θ). This article seeks to identify the most efficient item selection method for the latter type of CD-CAT corresponding to various conditions and various evaluation criteria, respectively, based on the reduced reparameterized unified model (RRUM) and the two-parameter logistic model of item response theory (IRT-2PLM). The Shannon entropy (SHE) and Fisher information methods were combined to produce a new synthetic item selection index, that is, the “dapperness with information (DWI)” index, which concurrently considers both KS and θ within one step. The new method was compared with four other methods. The results showed that, in most conditions, the new method exhibited the best performance in terms of KS estimation and the second-best performance in terms of θ estimation. Item utilization uniformity and computing time are also considered for all the competing methods.

Keywords

cognitive diagnostic computerized adaptive testing, item selection method, knowledge state, “dapperness with information” index

Computerized adaptive testing (CAT) is typically engineered to tailor a test to each examinee's trait level, thus matching the difficulties of the items to the examinee being measured (Chang, 2014). In addition, the aim of cognitive diagnosis is to provide information about specific content areas in which an examinee needs help (McGlohen & Chang, 2008). “Adaptive” is the core idea behind CAT, whereas “efficiency” is an important objective of cognitive diagnostic testing

¹Guangdong Key Laboratory of Mental Health and Cognitive Science, Center for Studies of Psychological Application, School of Psychology, South China Normal University, Guangzhou, China

²School of Psychology, Jiangxi Normal University, Nanchang, China

³Guangdong Psychological Association, Guangzhou, China

Corresponding Author:

Minqiang Zhang, School of Psychology, South China Normal University, Zhongshan West Road, Guangzhou 510631, China.

Email: zhangmq117@qq.com

(CDT). The combination of these two ideas resulted in cognitive diagnostic computerized adaptive testing (CD-CAT).

According to measurement targets, CD-CAT can be divided into two broad categories: (a) single-purpose tests (e.g., Xu, Chang, & Douglas, 2003) that only estimate the subject's knowledge state (KS; that is, attribute mastery pattern), and (b) dual-purpose tests (e.g., McGlohen & Chang, 2008; Wang, Chang, & Douglas, 2012) that estimate both the subject's KS and traditional ability level (θ) in item response theory (IRT) framework. Dual-purpose tests are more complex and thus require more technically sophisticated methods.

The item selection algorithm is a critical component of CD-CAT. It affects whether items are adaptable to the subject's current KS and/or ability and the level to which an item can be adapted to the subject. In addition, the item selection method affects the efficiency and efficacy of a test. To date, many item selection methods have been proposed for the two types of CD-CAT, and the best method may vary according to experimental conditions. Based on the reduced reparameterized unified model (RRUM) and various numbers of attributes, several item selection methods for dual-purpose CD-CAT have been synthetically compared in this article to identify the most item-efficient method (the method which can get the best synthetic result with the same number of items).

Literature Review

Item Selection Based on KS Alone

This category of CD-CAT is a combination of the principles of CDT and CAT. Specifically, it combines the "adaptive" object of the traditional CAT with an estimation of the subject's KS. The Shannon entropy (SHE) method (Tatsuoka, 2002; Tatsuoka & Ferguson, 2003; Xu et al., 2003), the Kullback–Leibler (KL) information method (Xu et al., 2003), and its extensions are the key methods. In a broad sense, both KL and SHE are information indices. They are used to compute the information that each item can provide for the current subject given the current estimate of his or her KS; then the item best tailored to the current estimate of the subject is selected. These algorithms are introduced below.

The minimum expected predictive SHE algorithm. Assume that, in the Bayesian context, the prior distribution of the KS vector, α , has been specified, where α is a vector of length K and K is the number of attributes. In addition, binary attributes are considered in this article, whereby each element of α takes a value of 1 or 0, which represents the mastery or nonmastery of an attribute. After the subject has answered $n - 1$ items, selecting the next item by minimizing the expected predictive SHE of the posterior distribution of α would lead to minimal uncertainty and allow the estimated KS vector to better approximate the true KS vector. There are $L = 2^K$ possible KS vectors. The prior distribution of α is $P(\alpha = \alpha_l) = \lambda_l$, where $l = 1, \dots, 2^K$. When a subject has answered $n - 1$ items, the posterior probability of his KS, α_l , is denoted by $\pi_{n-1}(\alpha_l)$:

$$\pi_{n-1}(\alpha_l) \propto \lambda_l \prod_{t=1}^{n-1} (P_{lt})^{Y_t} (1 - P_{lt})^{1-Y_t} \quad (1)$$

where P_{lt} denotes $P(Y_{lt} = 1)$ and is equal to $P(Y_t = 1 | \alpha = \alpha_l)$, which indicates the probability of the subject providing a correct answer to his item t given that his KS is α_l .

The SHE of the posterior distribution of the attribute pattern is

$$E(\pi_{n-1}) = - \sum_{l=1}^{2^K} [\pi_{n-1}(\alpha_l)] \times \log[\pi_{n-1}(\alpha_l)] \quad (2)$$

After answering n items, his answer pattern is $Y^n = (Y_1, Y_2, \dots, Y_{n-1}, Y_n)$, where the value of Y_n may be 0 or 1. Thus, the expected SHE is

$$\begin{aligned} \text{SHE}(\pi_n) &= \sum_{z=0}^1 [E(\pi_n | Y^{n-1}, Y_n = z) \times P(Y_n = z | Y^{n-1})] \\ &= \sum_{z=0}^1 \{E(\pi_n | Y^{n-1}, Y_n = z) \cdot \sum_{l=1}^{2^K} [P(Y_n = z | \alpha_l) \times \pi_{n-1}(\alpha_l)]\} \end{aligned} \quad (3)$$

The formula derivation above refers to the definition of conditional expectation. Then, in selecting the n th item for the subject, this algorithm attempts to minimize $\text{SHE}(\pi_n)$.

The maximum KL information algorithm and its advanced algorithms. The KL information (which is also called the KL distance or relative entropy) measures the distance or divergence between two probability distributions. The KL index for the i th item, given the subject's current estimate $\text{KS}(\hat{\alpha})$, is defined as the following sum of KL distances between $\hat{\alpha}$ and possible candidate attribute vectors α_l generated by the i th item:

$$\text{KL}_i(\hat{\alpha}) = \sum_{l=1}^L \left[\sum_{y=0}^1 \log \left(\frac{P(Y_i = y | \hat{\alpha})}{P(Y_i = y | \alpha_l)} \right) \times P(Y_i = y | \hat{\alpha}) \right]. \quad (4)$$

The inner sum represents the KL information for the distribution of the i th item depending on attribute vectors $\hat{\alpha}$ and α_l , where $\hat{\alpha}$ is regarded as a true value. Then, the item with the maximum KL index for the current subject is selected.

The KL index is also called the global discrimination index (GDI; Cheng, 2010). Many scholars have proposed improvements in the KL method. For example, Cheng (2009) proposed the posterior-weighted KL (PWKL) index and hybrid KL (HKL) index. If the prior is discrete uniform, then the PWKL index is equivalent to the likelihood-weighted KL (LWKL) information method. The HKL is the PWKL weighted by the inverse of the distance between the estimate of KS and any other possible KS. To ensure adequate coverage of all attributes/skills, Cheng (2010) imposed constraints on the number of times each attribute is represented in CD-CAT (for a review, see Chang, 2014), naming the new index as the Modified GDI (MGDI) and the new method as the Maximum Modified GDI (MMGDI) method.

Item Selection Based on Both KS and Ability

This category of CD-CAT provides cognitive diagnoses and estimates ability concurrently. During the test process, both a subject's KS (denoted as α) and traditional ability level (denoted as θ) are repeatedly estimated, and the item that is tailored to the current estimates of subject's current α and θ is selected. This category of CD-CAT is a combination of the main principles of CDT and CAT.

Based on the fusion model (Hartz, Roussos, & Stout, 2002) and the three-parameter logistic model of item response theory (IRT-3PLM), McGlohen & Chang (2008) proposed an item selection method for this category. This method involved construction of a shadow test (i.e., shadow item bank) that was optimized according to the subject's current ability estimate $\hat{\theta}$. Then, the best item for measuring the attribute vector $\hat{\alpha}$ was selected from the shadow test on the basis of the current $\hat{\alpha}$ estimate, using the SHE or the KL method. Based on the DINA model (Junker & Sijtsma, 2001) and the IRT-3PLM, Du (2010) proposed an item selection method that reversed the order: The shadow test was constructed using the minimized SHE according to $\hat{\alpha}$;

then, the best item for measuring $\hat{\theta}$ was selected from the shadow test on the basis of the current $\hat{\theta}$, using the Fisher information method.

Both of the above methods involve a shadow test; thus, they alternately select items according to ability and KS, and every item selection requires two steps. During the two steps, which aim to separately enhance the accuracy of ability and KS, the system selects the respective “local optimization” for each step; however, the combination of the two “local optimizations” does not necessarily guarantee a “good synthetic result.” The authors suggest that a more desirable item selection method should consider ability and KS concurrently within one step. To this end, a real “synthetic adaptive” algorithm must be developed.

The dual information (DI) method proposed by Cheng (2007) provides such a “synthetic adaptive” algorithm. The DI index was constructed as a weighted sum of the KL information of the $\hat{\theta}$ component and the KL information of the $\hat{\alpha}$ component, that is, $KL(\hat{\theta})$ and $KL(\hat{\alpha})$. According to the results, the effect of different weights on measurement precision was negligible unless the weights were extreme.

Wang et al. (2012) also proposed several “synthetic adaptive” algorithms. Among these algorithms, according to their study, the maximum information based method (MIinfor method) provided the highest KS and θ estimation accuracy. The method is described as follows: First, a priority index is proposed as

$$P_i = \sum_{k=1}^K (u_k - x_k) d_{ik}, \quad (5)$$

where P_i is the priority index of the i th item, d_{ik} summarizes the ability of the i th item to distinguish between examinees who had possessed the k th attribute and others who had not, and u_k and x_k are the upper bound and the accumulated information for the k th attribute, respectively. u_k is set by users and should be adjusted when one of the test conditions changes. The calculations of d_{ik} and x_k are introduced in Wang et al. (2012). $u_k - x_k$ serves as a weight for the attribute-level information and indicates the importance of the information for the k th attribute. Finally, the priority index is multiplied by the Fisher information, and the item with the maximum multiplication result would be selected.

The RRUM

To date, most CD-CAT item selection studies have utilized the DINA model. This model divides subjects into two groups: One group is composed of attribute vectors that contain all of the required attributes for item i , and the other group is composed of attribute vectors that lack at least one of the required attributes for the same item. Attribute vectors in the same group are assumed to have the same probability of answering the item correctly. However, the disadvantage of the DINA model is that this assumption may not always hold for the latter group because attribute vectors in this group have varying degrees of deficiencies with respect to the required attributes. Therefore, their probabilities of success may not be identical (de la Torre, 2011). Thus, the current article employs another cognitive diagnostic model, that is, the RRUM, which was proposed by Hartz (2002).

The RRUM is a compensatory model that allows the absence of one attribute to be remedied by the presence of other attributes. Its item parameters are as follows: The baseline parameter associated with item difficulty and the penalty parameter that describes the extent to which the mastery of a specific attribute would affect the chance of answering the item correctly (Feng, Habing, & Huebner, 2014). For each item, every attribute is assigned one

penalty parameter. Therefore, subjects with different KSs have different probabilities of successfully answering each item. Thus, the RRUM displays greater flexibility than the DINA model.

Consider an assessment of I items diagnosing K attributes. Let Y_i be the observed binary 1/0 response of a subject to item i , where 1 indicates that the subject provides a correct response to item i , and 0 indicates an incorrect response. The binary Q-matrix describes how the items are related to the attributes, with entry q_{ik} indicating whether attribute k is required by item i . Under the RRUM, the probability of a correct answer to item i given that a respondent has KS(α_i) is as follows (Feng et al., 2014; Rupp, Templin, & Henson, 2010):

$$p_{li} = P(Y_i = 1 | \alpha_i) = \pi_i^* \prod_{k=1}^K r_{ik}^{*(1-\alpha_{ik})q_{ik}}. \quad (6)$$

The baseline parameter π_i^* is the probability of a correct response to item i given that a respondent has mastered all the required attributes for the item. An item with a large π_i^* parameter indicates that the required attributes can effectively explain examinee responses to the item. The penalty parameter r_{ik}^* is the probability that a subject has not mastered attribute k but correctly answers item i divided by the probability that a subject has mastered attribute k and correctly answers item i .

In the remaining sections, a new index that could consider both KS and θ within one step without shadow tests is proposed, and is compared with four previous methods (Cheng, 2007; Du, 2010; McGlohen & Chang, 2008; Wang et al., 2012) under the RRUM and IRT-2PLM.

The Dapperness With Information (DWI) Index

As noted above, the most desirable item selection method should concurrently consider ability and KS within one step. To this end, a synthetic statistical index should be used. The authors decided to synthesize an index by combining one item selection method based on ability and one item selection method based on KS. In IRT-CAT, in which item selection is based on ability, the maximum Fisher information method is often used based on the current estimation of the subject's ability.

To compare the KS recoveries of the item selection methods for single-purpose CD-CAT, a pilot study was conducted. The following six methods were compared: random, SHE, KL (GDI), LWKL, HKL, and MMGDI. The independent variables are item selection methods (the six above-mentioned methods), test lengths (short test: length of $K \times 4$, long test: length of $K \times 5$), and attribute numbers ($K = 4, 6, 8$). The test length levels were set as above because a subject's KS requires more items be measured as the number of attributes increases. The sizes of item banks were set according to Stocking's (1994) recommendation. The experiments were repeated 20 times, using the MATLAB 2012b software package. Groups of examinees were simulated, and each group was composed of 1,000 examinees. The maximum a posteriori (MAP) method was used to estimate the KSs of the examinees. Then, the following results were obtained: for CD-CAT that estimates the subject's KS alone, the SHE method was the most accurate method, the methods of HKL and LWKL were slightly inferior, and the random method performed the worst. However, the SHE, HKL, and LWKL methods were almost identical, except for a small difference of no more than 1%. These results were similar to the findings of Cheng (2009) and Chen, Li, and Xin (2011) using the DINA model. Based on the above results, it was decided to include the SHE method in the synthetic index.

Thus, a synthetic index can be generated as a weighted sum as follows:

$$\omega f(I_i(\hat{\theta}_j)) + (1 - \omega)g(\text{SHE}_i(\hat{\alpha}_j)), \quad (7)$$

where $I_i(\hat{\theta}_j)$ is the Fisher information of item i with respect to the current estimation of the ability of subject j , $\text{SHE}_i(\hat{\alpha}_j)$ is the expected predictive SHE of the item to the same subject, f is a monotonically increasing function, g is a monotonically decreasing function, and $\omega \in (0, 1)$ represents the weight of the first part. Thus, the weight ω can be used to achieve a balance between the accuracy of the estimation of the ability θ and that of the KS(α). Because there is a considerable difference between the value range of $I_i(\hat{\theta}_j)$ and $\text{SHE}_i(\hat{\alpha}_j)$, it is not suitable to directly sum them with weights. Instead, $f(x) = \log(x)$ and $g(x) = -\log(x)$ are adopted, because the logarithmic function is often used for changing values with different orders of magnitude into values with the same order of magnitude.

The authors have tried several values of ω (0.1, 0.3, 0.5, 0.7, and 0.9) and found that $\omega = 0.5$ (i.e., $\omega = \frac{1}{2}$) provided the best synthetic results (See Appendix B). When $\omega = \frac{1}{2}$, the synthetic index is equal to

$$\frac{1}{2} \log(I_i(\hat{\theta}_j)) - (1 - \frac{1}{2}) \log(\text{SHE}_i(\hat{\alpha}_j)) = \frac{1}{2} \log \frac{I_i(\hat{\theta}_j)}{\text{SHE}_i(\hat{\alpha}_j)} \quad (8)$$

Due to the monotonicity of the logarithm function, choosing an item that maximizes $\frac{1}{2} \log \frac{I_i(\hat{\theta}_j)}{\text{SHE}_i(\hat{\alpha}_j)}$ is equivalent to that maximizes $\frac{I_i(\hat{\theta}_j)}{\text{SHE}_i(\hat{\alpha}_j)}$. Thus, the authors decide to use the ratio of the Fisher information and the expected predictive SHE as the synthetic item selection index for the dual-purpose CD-CAT.

In physics, entropy denotes the degree of confusion; thus, it is used to measure the degree of uncertainty (Shannon, 2001). Conversely, its inverse can be used to measure the degree of certainty, which is referred to as the degree of dapperness (or order). In CD-CAT, the SHE index selects the item that minimizes the expected SHE of the posterior distribution of the attribute vectors. Equivalently, a dapperness index can be constructed (the inverse of the SHE index), which selects the item with the maximum value. The new synthetic index is a combination of the dapperness index and the Fisher information quotient; thus, it is labeled as the DWI index. The formula can be written as

$$\text{DWI}_i = \frac{I_i(\hat{\theta}_j)}{\text{SHE}_i(\hat{\alpha}_j)}. \quad (9)$$

The Simulation Study

Simulation Design

Both McGlohen & Chang (2008) and Du (2010) used the IRT-3PLM. However, Baker and Kim (2004) noted that for the IRT-3PLM, 1,000 subjects and 60 items would be needed for accurately estimating the item and ability parameters. In the current study, all test lengths are less than 60 items; therefore, the IRT-2PLM is selected as Wang et al. (2012) did.

The purpose of this study is to compare the KS recovery, ability recovery, item utilization, and computing time of different item selection methods using the RRUM and the IRT-2PLM. The following five methods are compared: McGlohen & Chang (2008) method, Du's (2010) method, the DI method, the MInfor method, and the proposed DWI method.

The independent variables are the item selection methods (the five above-mentioned methods), attribute numbers ($K = 4, 6, 8$), and the type of relationship among the attributes (uncorrelated, correlated). The first independent variable is a within-group variable, and the latter two are between-group variables.

The evaluation criteria are the average values of the attribute recovery rate (ARR), the average value of the pattern correct classification rate (PCCR), the average values of the mean absolute error (ABSE), the average values of item utilization uniformity (χ^2), and the average values of computing time. The ARR quantifies the estimation accuracy of certain attributes, and the ARR of the k th attribute is computed as follows:

$$ARR_k = \frac{J_{kcorrect}}{J}, \quad (10)$$

where J is the total number of subjects and $J_{kcorrect}$ is the number of subjects whose the k th attribute are correctly classified.

The PCCR, which quantifies the estimation accuracy of the entire KS, is computed as follows:

$$PCCR = \frac{J_{patterncorrect}}{J}, \quad (11)$$

where $J_{patterncorrect}$ is the number of subjects whose KSs are correctly classified.

The ABSE quantifies the estimation accuracy of subjects' ability levels and is computed as follows:

$$ABSE = \frac{1}{J} \sum_{j=1}^J |\hat{\theta}_j - \theta_j|, \quad (12)$$

where θ_j is obtained through estimation involving the entire bank of items (see as the true IRT ability value of subject j), and $\hat{\theta}_j$ is its corresponding estimate.

Item utilization uniformity quantifies the difference between ideal item exposure distribution and real item exposure distribution, and is computed as follows (Chen et al., 2011):

$$\chi^2 = \frac{\sum_i (er_i - L/I)^2}{L/I}, \quad (13)$$

where L is the test length, L/I means ideal item exposure, er_i means the observed exposure of item i . The smaller the χ^2 value is, the more ideal the item utilization is.

Experiments were repeated 20 times using the MATLAB 2012b software package. Groups of examinees were simulated, and each group was composed of 1,000 examinees.

The number of items in the item bank is denoted as I . Stocking (1994) noted that a rule of thumb in item banking is that the pool must have at least 12 times as many items as the test length. Thus, the size of the item bank was set as $I = 300$ when $K = 4$, $I = 450$ when $K = 6$, and $I = 600$ when $K = 8$.

For the case in which all attributes were uncorrelated, each subject's true KS was randomly generated, with each attribute being independently generated and having an equal probability of being 0 or 1. For the case in which attributes were correlated, the correlation coefficients between any two different attributes were set as .5, and the probability of an attribute being mastered by a certain examinee remained .5, as before. The KS simulation of the latter case is identical to that of Henson and Douglas (2005), using multivariate normal K -dimensional vectors.

To ensure that every attribute was measured by at least one item, the first K items in the item bank were set as follows: the k th item measured the k th attribute only, where k is 1, 2, $\dots$, K . These items were used as the initial items in the CD-CAT simulation. In regard to the other items, Q-matrix generation requires that the probability that an attribute is measured by a given item is .5, and there should be no item that measures nothing. According to Feng et al. (2014), the baseline parameters and penalty parameters are generated from uniform distributions of (0.6, 1.0) and (0.05, 0.4), respectively.

The simulation involved 11 steps:

1. Q-matrix generation;
2. baseline parameter and penalty parameter generation, based on the RRUM;
3. true KS generation;
4. simulation of the subjects' responses on all of the items, generating the complete response matrix;
5. estimation of θ and item parameters a and b based on the IRT-2PLM, according to the complete response matrix;
6. selection of initial items for a subject;
7. search of the complete response matrix for the subject's responses to the initial items and presentation of the preliminary estimates of the subject's KS and θ ;
8. selection of the next item according to the item selection method;
9. search of the subject's scores for the selected item (based on the complete response matrix);
10. reestimation of the subject's KS and θ ; and
11. repetition of Steps 8 through 10 until the fixed test length has been reached.

The length of the shadow test would influence the results of McGlohen & Chang (2008) and Du's (2010) methods. After repeated trials, the best lengths of the shadow tests were found and selected in the design. The weights of $KL(\hat{\theta})$ and $KL(\hat{\alpha})$ in the DI method were set to equality. In addition, the upper bound information for each attribute employed in the Mlinfor method was obtained and selected after repeated testing.

The MAP method was used to estimate the KSs of the examinees, and the expected a posteriori (EAP) method was used to estimate the θ s.

A fixed-length stopping rule was employed. Because dual propose CD-CAT requires more items than single propose CD-CAT, the test lengths were fixed at 6 times the number of attributes.

Results

The results for the case in which all attributes were uncorrelated are provided in Tables 1 to 3.

As shown in Table 1, the following results were obtained when the attributes were uncorrelated and $K = 4$: With respect to PCCR, the priority of the five methods was the methods of DI, Du's (2010), DWI, McGlohen & Chang (2008), and Mlinfor. With respect to θ , the priority of the five methods was the methods of Mlinfor, DWI, McGlohen & Chang (2008), Du's (2010), and DI. The difference between Mlinfor and DWI was relatively small.

As shown in Table 2, when attributes were uncorrelated and $K = 6$, with respect to PCCR, the priority of the five methods was the methods of DWI, Du's (2010), McGlohen & Chang (2008), DI, and Mlinfor. With respect to θ , the priority of the five methods was the same as when $K = 4$.

Table 1. Results for the Five Methods When Attributes Were Uncorrelated and $K = 4$ (Item Bank = 300, Test Length = 24).

Method	ARR				PCCR		ABSE		χ^2	T
	At.1	At.2	At.3	At.4	M	SE	M	SE	M	M
McGlohen's	0.961	0.965	0.955	0.967	0.890	0.004	0.313	0.005	73.8	1.6
Du's	0.974	0.976	0.972	0.978	0.913	0.005	0.415	0.008	59.6	14.2
DI	0.976	0.981	0.980	0.981	0.924	0.008	0.681	0.012	78.4	89.0
Mlinfor	0.952	0.953	0.947	0.956	0.859	0.005	0.291	0.002	142.0	1.4
DWI	0.964	0.967	0.958	0.968	0.898	0.004	0.295	0.004	94.5	14.6

Note. χ^2 expresses item utilization uniformity. ARR = attribute recovery rate; PCCR = pattern correct classification rate; ABSE = absolute error; T = time (minutes); At. = Attribute; DI = dual information; Mlinfor = maximum information; DWI = dapperness with information.

Table 2. Results for the Five Methods When Attributes Were Uncorrelated and $K = 6$ (Item Bank = 450, Test Length = 36).

Method	ARR						PCCR		ABSE		χ^2	T
	At.1	At.2	At.3	At.4	At.5	At.6	M	SE	M	SE	M	M
McGlohen's	0.959	0.955	0.962	0.960	0.962	0.965	0.824	0.007	0.340	0.004	106.5	6.0
Du's	0.964	0.965	0.967	0.967	0.966	0.969	0.841	0.008	0.490	0.008	84.5	111.1
DI	0.952	0.971	0.958	0.964	0.975	0.974	0.821	0.008	1.043	0.012	105.4	203.0
Mlinfor	0.955	0.952	0.954	0.956	0.958	0.956	0.816	0.004	0.281	0.003	262.0	3.5
DWI	0.966	0.967	0.968	0.966	0.970	0.971	0.858	0.007	0.312	0.004	157.5	112.4

Note. ARR = attribute recovery rate; PCCR = pattern correct classification rate; ABSE = absolute error; T = time (minutes); At. = Attribute; DI = dual information; Mlinfor = maximum information; DWI = dapperness with information.

Table 3. Results for the Five Methods When Attributes Were Uncorrelated and $K = 8$ (Item Bank = 600, Test Length = 48).

Method	ARR								PCCR		ABSE		χ^2	T
	At.1	At.2	At.3	At.4	At.5	At.6	At.7	At.8	M	SE	M	SE	M	M
McGlohen's	0.962	0.955	0.962	0.956	0.960	0.959	0.965	0.955	0.754	0.007	0.657	0.029	146.3	30.7
Du's	0.970	0.970	0.969	0.967	0.970	0.971	0.972	0.969	0.799	0.007	0.803	0.023	145.2	868.9
DI	0.964	0.957	0.965	0.968	0.961	0.965	0.967	0.952	0.738	0.013	1.372	0.028	170.3	423.8
Mlinfor	0.968	0.965	0.966	0.967	0.967	0.966	0.971	0.966	0.799	0.006	0.507	0.023	283.9	14.9
DWI	0.970	0.969	0.970	0.970	0.971	0.970	0.974	0.970	0.814	0.007	0.584	0.025	198.3	884.1

Note. ARR = attribute recovery rate; PCCR = pattern correct classification rate; ABSE = absolute error; T = time (minutes); At. = Attribute; DI = dual information; Mlinfor = maximum information; DWI = dapperness with information.

As shown in Table 3, when attributes were uncorrelated and $K = 8$, with respect to PCCR, the priority of the five methods was the methods of DWI, Du's (2010), Mlinfor, McGlohen & Chang (2008), and DI. With respect to θ , the priority of the five methods was the same as when $K = 4$.

Table 4. Results for the Five Methods When Attributes Were Correlated and $K = 4$ (Item Bank = 300, Test Length = 24).

Method	ARR				PCCR		ABSE		χ^2	T
	At.1	At.2	At.3	At.4	M	SE	M	SE	M	M
McGlohen's	0.991	0.990	0.988	0.991	0.965	0.002	0.264	0.005	60.1	1.5
Du's	0.990	0.990	0.990	0.990	0.962	0.002	0.331	0.006	79.8	14.0
DI	0.965	0.964	0.931	0.964	0.832	0.015	0.388	0.005	76.7	89.3
Mlinfor	0.989	0.987	0.989	0.988	0.958	0.003	0.244	0.004	93.7	1.4
DWI	0.993	0.992	0.990	0.993	0.971	0.003	0.244	0.004	71.5	14.6

Note. ARR = attribute recovery rate; PCCR = pattern correct classification rate; ABSE = absolute error; T = time (minutes); At. = Attribute; DI = dual information; Mlinfor = maximum information; DWI = dapperness with information.

Table 5. Results for the Five Methods When Attributes Were Correlated and $K = 6$ (Item Bank = 450, Test Length = 36).

Method	ARR						PCCR		ABSE		χ^2	T
	At.1	At.2	At.3	At.4	At.5	At.6	M	SE	M	SE	M	M
McGlohen's	0.986	0.975	0.977	0.987	0.984	0.985	0.901	0.006	0.343	0.010	54.5	6.0
Du's	0.982	0.978	0.979	0.982	0.980	0.981	0.892	0.006	0.383	0.007	84.6	111.0
DI	0.941	0.902	0.918	0.942	0.934	0.937	0.646	0.012	0.421	0.003	111.0	204.2
Mlinfor	0.981	0.972	0.975	0.981	0.976	0.981	0.885	0.006	0.287	0.005	139.5	3.5
DWI	0.986	0.973	0.975	0.987	0.983	0.984	0.904	0.007	0.308	0.006	86.2	112.5

Note. ARR = attribute recovery rate; PCCR = pattern correct classification rate; ABSE = absolute error; T = time (minutes); At. = Attribute; DI = dual information; Mlinfor = maximum information; DWI = dapperness with information.

In all, as shown in Tables 1 to 3, the following results were obtained when the attributes were uncorrelated: (a) with respect to KS estimation accuracy, the DWI method was the third best when $K = 4$, and it provided the best accuracy when $K = 6$ and $K = 8$. (b) With respect to θ estimation accuracy, regardless of the value of K , the Mlinfor method consistently provided the best results, while the DWI method always provided the second-best results; the difference between them was relatively small.

The results when "attributes were correlated" are provided in Tables 4 to 6.

As shown in Table 4, the following results were obtained when the attributes were correlated and $K = 4$: With respect to PCCR, the priority of the five methods was the methods of DWI, McGlohen & Chang (2008), Du's (2010), Mlinfor, and DI. With respect to θ , the priority of the five methods was the methods of DWI and Mlinfor, McGlohen & Chang (2008), Du's (2010), and DI. There was no difference between DWI and Mlinfor. Compared with Table 1, when attributes were correlated, all the methods had a better θ estimation, and all the methods had a better KS estimation except DI.

As shown in Table 5, when attributes were uncorrelated and $K = 6$, with respect to PCCR, the priority of the five methods was the same as in Table 4. With respect to θ , the priority of the five methods was almost the same as in Table 4, except that Mlinfor was slightly better than DWI. Compared with Table 2, when attributes were correlated, all the methods had a better θ estimation except McGlohen & Chang (2008) and Mlinfor, and all the methods had a better KS estimation except DI.

Table 6. Results for the Five Methods When Attributes Were Correlated and $K = 8$ (Item Bank = 600, Test Length = 48).

Method	ARR								PCCR		ABSE		χ^2	T
	At.1	At.2	At.3	At.4	At.5	At.6	At.7	At.8	M	SE	M	SE	M	M
McGlohen's	0.964	0.961	0.960	0.960	0.965	0.961	0.964	0.955	0.764	0.006	0.544	0.020	55.5	31.4
Du's	0.966	0.965	0.964	0.966	0.970	0.967	0.967	0.961	0.774	0.008	0.444	0.012	77.4	868.7
DI	0.921	0.911	0.906	0.904	0.901	0.917	0.919	0.896	0.470	0.013	0.414	0.004	145.0	423.7
MIinfor	0.963	0.961	0.958	0.962	0.963	0.963	0.962	0.952	0.747	0.007	0.457	0.020	185.0	14.9
DWI	0.966	0.962	0.959	0.962	0.967	0.963	0.966	0.956	0.780	0.005	0.396	0.006	82.5	883.9

Note. ARR = attribute recovery rate; PCCR = pattern correct classification rate; ABSE = absolute error; T = time (minutes); At. = Attribute; DI = dual information; MIinfor = maximum information; DWI = dapperness with information.

As shown in Table 6, when attributes were uncorrelated and $K = 8$, with respect to PCCR, the priority of the five methods was the methods of DWI, Du's (2010), McGlohen & Chang (2008), MIinfor, and DI. With respect to θ , the priority of the five methods was the methods of DWI, DI, Du's (2010), MIinfor, and McGlohen & Chang (2008). Compared with Table 3, when attributes were correlated, all the methods had a better θ estimation, and all the methods had a worse KS estimation except McGlohen & Chang (2008).

As shown in Tables 4 to 6, the following results were obtained when attributes were correlated: (a) In terms of KS estimation accuracy, regardless of the value of K , the DWI method consistently provided the best results. (2) In terms of θ estimation accuracy, the DWI method provided the best accuracy when $K = 4$ and $K = 8$, and it was the second best when $K = 6$.

With respect to item utilization uniformity, (a) as shown in Tables 1 to 3, when attributes were uncorrelated, the first three methods provided the best uniformity, following DWI, and MIinfor provided the worst uniformity. (2) As shown in Tables 4 to 6, when attributes were correlated, McGlohen & Chang (2008) method provided the best uniformity, MIinfor provided the worst uniformity, and the other three methods were in the middle.

With respect to computing time, (a) as shown in Tables 1 and 4, when $K = 4$, the methods of MIinfor and McGlohen & Chang (2008) were the fastest, the methods of Du's (2010) and DWI were in the middle, and DI was the slowest. (b) As shown in Tables 2 and 5, when $K = 6$, the results were the same as when $K = 4$. (c) As shown in Tables 3 and 6, when $K = 8$, the methods of MIinfor and McGlohen & Chang (2008) were the fastest, DI was in the middle, and Du's (2010) and DWI were the slowest. For this setting, DWI cost 884.1 or 883.9 min, it means every subject cost about 0.884 min (about 53 s) for all the 48 items, and selecting one item for one subject cost about 1.1 s. (d) In all, no matter what K was, MIinfor consistently cost the least times, and McGlohen & Chang (2008) was always the second fastest. DWI was the second slowest when $K = 4$ and 6, and was the slowest when $K = 8$.

The authors have also tested a few different test lengths including Test Length = $4 \times K$, $5 \times K$, and $7 \times K$, and found that across various practical test lengths, the DWI method was robust and the orders of the five methods were almost the same. The data are presented in Appendix A.

Discussion and Conclusion

Some tests, such as achievement tests in primary school and middle school, require both macro ability estimation and micro KS estimation. This study proposes a practical method, DWI, that attempts to achieve this goal via CD-CAT at a minimum cost.

Under the RRUM, five item selection methods based on both the subject's KS and the subject's ability were synthetically compared across various numbers and correlations of knowledge attributes in this study. The results indicated that the DWI method and the MIinfor method were more item-efficient in practice than the other methods. In general, the DWI method was found to be better than the MIinfor method with respect to KS estimation and item utilization uniformity, and the latter was found to be better than the former with respect to θ estimation and computing time. The DWI method is more effective in situations that require better KS estimation compared with θ estimation.

In addition, both McGlohen & Chang (2008) method and Du's (2010) method involve a shadow test. Therefore, their methods have the following disadvantages: (a) the selection of one item requires two steps at each iteration. As previously illustrated, the system determines the "local optimization" for each step, but the combination of two "local optimizations" does not necessarily guarantee a "good synthetic result." (b) The length of the shadow test influences KS and θ estimation results. "Best length" of the shadow test may be related to multiple factors, such as the number of attributes, the size of the item bank, and the total length of the test. When one of these conditions changes, the length of the shadow test should be reset, which creates difficulties for research. Because they do not employ shadow tests, the MIinfor method, the DI method, and the proposed DWI method have the following two advantages: (a) within one step, "good synthetic result" is pursued throughout the entire item selection process; and (b) there is no need to blindly search for "the best length." Therefore, this novel method is easier to implement.

In addition, the upper bound information for each attribute in the MIinfor method influences the estimation results. "Best upper bound" is related to multiple factors, such as the quality of the item bank and the test length, and should be adjusted when one of these conditions changes. By contrast, the DWI method is quite simple because it does not encounter this problem.

However, when the number of cognitive attributes involved in the test was relatively large (e.g., eight attributes), none of the five item selection methods that were based on both the KS and the θ provided a satisfactory θ estimation accuracy, especially when attributes were uncorrelated. This finding implies that, given current technical conditions, it is advisable to avoid including too many attributes in a dual-purpose CD-CAT.

Furthermore, two certain forms among the attributes was assumed in this article, while a Q-matrix constructed from a real test blueprint could take several complicated forms, such as linear, convergent, and divergent (Leighton, Gierl, & Hunka, 2004). Determination of the best method under each type of structure warrants future investigation.

In this article, the RRUM and the IRT-2PLM are both used. Hence, there are some limitations. In future studies, a new model named the High-Order RRUM, similar to the High-Order DINA model (de la Torre & Douglas, 2004), may be proposed. In the High-Order RRUM, attributes are modeled as arising from a broadly defined latent trait that resembles θ in the item response models. Thus, KS estimation and general ability estimation could be modeled in the same analysis, that is, using one model rather than two models.

The estimation time given in this article can be considered conservative. This is because (a) the estimation code was written by a nonprofessional programmer in a certain programming language (i.e., MATLAB), and (b) the programs were operated by a server with low capacity (only two cores, 2.67 GHz, 4GB). Additional efficiency can be gained if the estimation code can be written by a more experienced programmer using another language and high-capacity servers; therefore, the implementation of the DWI method with larger data sets can be more practicable.

Appendix A

When $\omega = \frac{1}{2}$, Test Length = $5 \times K$, and attributes were correlated, the results were as follows:

Table A1. Mean PCCR and Mean ABSE for the Five Methods When Attributes Were Correlated and $K = 4$ (Item Bank = 300, Test Length = 20).

Item selection method	PCCR	ABSE
	<i>M</i>	<i>M</i>
McGlohen's	0.942	0.280
Du's	0.922	0.333
DI	0.689	0.422
Mlinfor	0.928	0.267
DWI	0.949	0.269

Note. PCCR = pattern correct classification rate; ABSE = absolute error; DI = dual information; Mlinfor = maximum information; DWI = dapperness with information.

Table A2. Mean PCCR and Mean ABSE for the Five Methods When Attributes Were Correlated and $K = 6$ (Item Bank = 450, Test Length = 30).

Item selection method	PCCR	ABSE
	<i>M</i>	<i>M</i>
McGlohen's	0.852	0.371
Du's	0.847	0.404
DI	0.539	0.434
Mlinfor	0.850	0.299
DWI	0.864	0.324

Note. PCCR = pattern correct classification rate; ABSE = absolute error; DI = dual information; Mlinfor = maximum information; DWI = dapperness with information.

Table A3. Mean PCCR and Mean ABSE for the Five Methods When Attributes Were Correlated and $K = 8$ (Item Bank = 600, Test Length = 40).

Item selection method	PCCR	ABSE
	<i>M</i>	<i>M</i>
McGlohen's	0.745	0.706
Du's	0.745	0.494
DI	0.452	0.451
Mlinfor	0.726	0.504
DWI	0.747	0.447

Note. PCCR = pattern correct classification rate; ABSE = absolute error; DI = dual information; Mlinfor = maximum information; DWI = dapperness with information.

Appendix B

Results of different ω values in DWI when Test Length = $6 \times K$ are as follows:

Table B1. Mean PCCR and Mean ABSE for the DWI Method When Attributes Were Uncorrelated and $K = 4$.

	PCCR	ABSE
$\omega = 0.1$	0.982	0.347
$\omega = 0.3$	0.928	0.304
$\omega = 0.5$	0.898	0.295
$\omega = 0.7$	0.873	0.300
$\omega = 0.9$	0.857	0.309

Note. PCCR = pattern correct classification rate; ABSE = absolute error; DWI = dapperness with information.

Table B2. Mean PCCR and Mean ABSE for the DWI Method When Attributes Were Uncorrelated and $K = 6$.

	PCCR	ABSE
$\omega = 0.1$	0.970	0.500
$\omega = 0.3$	0.906	0.348
$\omega = 0.5$	0.858	0.312
$\omega = 0.7$	0.803	0.322
$\omega = 0.9$	0.752	0.326

Note. PCCR = pattern correct classification rate; ABSE = absolute error; DWI = dapperness with information.

Table B3. Mean PCCR and Mean ABSE for the DWI Method When Attributes Were Uncorrelated and $K = 8$.

	PCCR	ABSE
$\omega = 0.1$	0.968	0.901
$\omega = 0.3$	0.883	0.700
$\omega = 0.5$	0.814	0.584
$\omega = 0.7$	0.748	0.525
$\omega = 0.9$	0.690	0.489

Note. PCCR = pattern correct classification rate; ABSE = absolute error; DWI = dapperness with information.

Table B4. Mean PCCR and Mean ABSE for the DWI Method When Attributes Were Correlated and $K = 4$.

	PCCR	ABSE
$\omega = 0.1$	0.995	0.264
$\omega = 0.3$	0.979	0.246
$\omega = 0.5$	0.971	0.244
$\omega = 0.7$	0.959	0.248
$\omega = 0.9$	0.921	0.251

Note. PCCR = pattern correct classification rate; ABSE = absolute error; DWI = dapperness with information.

Table B5. Mean PCCR and Mean ABSE for the DWI Method When Attributes Were Correlated and $K = 6$.

	PCCR	ABSE
$\omega = 0.1$	0.974	0.322
$\omega = 0.3$	0.927	0.313
$\omega = 0.5$	0.904	0.308
$\omega = 0.7$	0.883	0.307
$\omega = 0.9$	0.865	0.304

Note. PCCR = pattern correct classification rate; ABSE = absolute error; DWI = dapperness with information.

Table B6. Mean PCCR and Mean ABSE for the DWI Method When Attributes Were Correlated and $K = 8$.

	PCCR	ABSE
$\omega = 0.1$	0.841	0.534
$\omega = 0.3$	0.800	0.414
$\omega = 0.5$	0.780	0.396
$\omega = 0.7$	0.763	0.391
$\omega = 0.9$	0.736	0.387

Note. PCCR = pattern correct classification rate; ABSE = absolute error; DWI = dapperness with information.

Acknowledgments

The authors thank the editors and the anonymous reviewers for their constructive comments on the earlier drafts. They would also like to thank Dr. Jimmy de la Torre, Dr. Chun Wang, Dr. Ying Cheng, Dr. Yiqun Gan, Dr. Zhaosheng Luo, Mr. Shuliang Ding, Dr. Dongbo Tu, Dr. Ping Chen, Dr. Wenyi Wang, and Dr. Lei Guo for their suggestions.

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) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This project was supported by Guangzhou education scientific planning subject (Grant No. 1201411413), by the National Social Science Fund of China during the Twelfth Five-Year Plan Period (Grant No. BHA130053), by Guangzhou primary and secondary school education quality comprehensive evaluation project (2) (Grant No. GZJY2051S/YD16G0510), and by the National Natural Science Foundation of China (Grant No. 31470050).

References

- Baker, F. B., & Kim, S. H. (Eds.). (2004). *Item response theory: Parameter estimation techniques*. New York, NY: Marcel Dekker.
- Chang, H. H. (2014). Psychometrics behind computerized adaptive testing. *Psychometrika*, 79, 1-20.

- Chen, P., Li, Z., & Xin, T. (2011). A note on the uniformity of item bank usage in cognitive diagnostic computerized adaptive testing. *Studies of Psychology and Behavior*, 9, 125-132.
- Cheng, Y. (2007). *The dual information method for item selection in cognitive diagnostic computerized adaptive testing* (Master's thesis). University of Illinois at Urbana-Champaign.
- 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.
- de la Torre, J. (2011). The generalized DINA model framework. *Psychometrika*, 76, 179-199.
- de la Torre, J., & Douglas, J. (2004). Higher-order latent trait models for cognitive diagnosis. *Psychometrika*, 69, 333-353.
- Du, X. X. (2010). *A new strategy of item selection of cognitive diagnosis computerized adaptive testing* (Master's thesis). Jiangxi Normal University, Nanchang, China.
- Feng, Y., Habing, B. T., & Huebner, A. (2014). Parameter estimation of the reduced RUM using the EM algorithm. *Applied Psychological Measurement*, 38, 137-150.
- Hartz, S. M., Roussos, L., & Stout, W. (2002). *Skills diagnosis: Theory and practice* (User manual for Arpeggio software). Educational Testing Service.
- Hartz, S. M. (2002). *A Bayesian framework for the unified model for assessing cognitive abilities: Blending theory with practicality* (Doctoral dissertation). University of Illinois at Urbana-Champaign.
- Henson, R., & Douglas, J. (2005). Test construction for cognitive diagnosis. *Applied Psychological Measurement*, 29, 262-277.
- Junker, B. W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement*, 25, 258-272.
- 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, 205-236.
- McGlohen, M. K., & Chang, H. H. (2008). Combining computer adaptive testing technology with cognitively diagnostic assessment. *Behavior Research Methods*, 40, 808-821.
- Rupp, A., Templin, J., & Henson, R. (2010). *Diagnostic measurement: Theory, methods, and applications*. New York, NY: Guilford Press.
- Shannon, C. E. (2001). A mathematical theory of communication. *ACM SIGMOBILE Mobile Computing and Communications Review*, 5, 3-55.
- Stocking, M. L. (1994). *Three practical issues for modern adaptive testing item pools* (ETS Research Report ERT-RR-94-5). Princeton, NJ: Educational Testing Service.
- 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 Royal Statistics, Series B: Statistical Methodology*, 65, 143-157.
- Wang, C., Chang, H. H., & Douglas, J. (2012). Combining CAT with cognitive diagnosis: A weighted item selection approach. *Behavior Research Methods*, 44, 95-109.
- Xu, X., Chang, H. H., & Douglas, J. (2003, April). *A simulation study to compare CAT strategies for cognitive diagnosis*. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.