Article

Educational and Psychological Measurement
2019, Vol. 79(2) 335–357

DOI: 10.1177/0013164418790634
journals.sagepub.com/home/epm
SAGE

Item Selection Criteria With Practical Constraints in Cognitive Diagnostic Computerized Adaptive Testing

Chuan-Ju Lin[1] and Hua-Hua Chang[2]

Abstract

For item selection in cognitive diagnostic computerized adaptive testing (CD-CAT), ideally, a single item selection index should be created to simultaneously regulate precision, exposure status, and attribute balancing. For this purpose, in this study, we first proposed an attribute-balanced item selection criterion, namely, the standardized weighted deviation global discrimination index (SWDGDI), and subsequently formulated the constrained progressive index (CP_SWDGDI) by casting the SWDGDI in a progressive algorithm. A simulation study revealed that the SWDGDI method was effective in balancing attribute coverage and the CP_SWDGDI method was able to simultaneously balance attribute coverage and item pool usage while maintaining acceptable estimation precision. This research also demonstrates the advantage of a relatively low number of attributes in CD-CAT applications.

Keywords

cognitive diagnostic computerized adaptive test, cognitive diagnostic model, the reduced version of RUM, item selection criteria

Introduction

In recent years, cognitive diagnostic computerized adaptive testing (CD-CAT) has received much attention because it involves two popular psychometric aspects, namely, cognitive diagnosis and computerized adaptive testing. Interest in cognitive

[1]National University of Tainan, Tainan, Taiwan
[2]University of Illinois at Urbana-Champaign, Champaign, IL, USA

Corresponding Author:
Chuan-Ju Lin, Department of Education, National University of Tainan, 33, Section 2, Shu-Lin Street, Tainan 70005, Taiwan.
Email: cjulin@mail.nutn.edu.tw

diagnosis is motivated by the increasing frequency of requests for diagnostic feedback to students, parents, educators, and administrators. A cognitive diagnostic test generates an attribute profile rather than a summative score for each examinee; the attribute profile identifies the concepts and areas that the examinee in question has mastered and skills for which remedial instruction could facilitate improvement by generating data necessary for continuous improvement of teaching and learning. Computerized adaptive testing (CAT) has been a popular research topic for decades because of its individualized and efficient features. Under CAT, a test is tailored to an examinee's latent trait levels, and thus CAT may provide an efficient latent trait estimate compared with fixed form testing (Weiss, 1982). Thus, future administration of a cognitive diagnostic test tailored to an examinee's mastery status appears inevitable.

Because the objective of CAT is to sequentially select items matching a latent estimate, the optimal method of item selection is the core consideration. Although effective item selection methods for item response theory-based CAT have been developed, few have been developed for CD-CAT. Therefore, developing item selection criteria suitable for CD-CAT is a crucial goal to pursue. Several item selection indices have been proposed in CD-CAT studies. Criteria concerning an item's level of attractiveness in relation to its psychometric properties were initially proposed; examples include the Kullback–Leibler (KL)-based global discrimination index (GDI) and Shannon entropy procedure (Xu, Chang, & Douglas, 2003) and posterior-weighted KL information (PWKL) (Cheng, 2009). However, as detailed in this article, the focus of these criteria on maximizing psychometric information on tests without concern for attribute balancing or exposure control may lead to two problems. The first problem is unbalanced attribute coverage, which calls a test's validity into question. For example, valid or reliable inferences regarding whether a student has mastered converting imperial units into metric units cannot be drawn when a CD-CAT measurement procedure involves fewer items than required or even no items to assess the student's conversion skill. The second problem is highly uneven item pool usage, which refers to some items being administered to an excessive number of examinees; this endangers item pool security, and some items are never used, which causes economic inefficiency in test development. Although CD-CAT tends to be applied for classroom settings and low-stack settings, where test security is not a major concern, the concern of balancing the item exposure rate in CD-CAT is still crucial for test developers and practitioners. Specifically, the item pool must be maintained because the construction of CD-CAT items is tedious as well as time and money consuming given that item writing for CD-CAT must be based on a complicated blueprint of cognitive requirements. In addition, the practice or memorizing effect may generate invalid diagnostic information for repeated test takers when particular items are administered in every test. Under the limited pool size condition, which commonly appears in CD-CAT, how to use most items in the pool is a critical issue.

Some researchers have considered practical constraints along with psychometric appeal by incorporating, for example, an attribute-balancing index into the GDI, namely, the modified maximum global discrimination index (MMGDI) method (Cheng, 2010), or modifying an information index through a progressive exposure control technique, namely, the restrictive progressive posterior weighted Kullback–Leibler (RP_PWKL) information index method (Wang, Chang, & Huebner, 2011). However, these methods address one of the aforementioned two problems while ignoring the other, and thus, uneven item pool utilization is likely generated under the MMGDI and unbalanced attribute coverage with the RP_PWKL.

Few attempts to develop item selection criteria while considering attribute balancing and exposure control have been made. Thus, this article proposes a holistic item selection method for CD-CAT, namely, the constrained progressive index (CP_SWDGDI), which can simultaneously ensure adequate attribute coverage and a balanced item exposure rate. This holistic index is formulated by replacing the item information element in a progressive algorithm with an attribute-balanced item selection criterion, which is also proposed in this article and named the standardized weighted deviation GDI (SWDGDI).

The remainder of this article begins by describing the cognitive diagnostic model (CDM) applied in this study, namely, the reduced reparameterized unified model (reduced RUM; Roussos, DiBello, Stout, Hartz, Henson, & Templin, 2007). Subsequently, the KL-based GDI is reviewed before a detailed introduction of the proposed methods is provided; the proposed methods were evaluated in a simulation study where the GDI method was used as a baseline under various experimental conditions. This article concludes with a discussion, limitations of the proposed methods, and suggestions for application with a smaller number of attributes as well as for future studies.

Method

This section describes the two most crucial elements in CD-CAT development and implementation, namely, CDM selection and the item selection algorithm. Many CDMs have been developed and demonstrated as applicable for practical formative assessment (e.g., Fusion model applied by Romàn, 2009) and adaptive testing (e.g., deterministic input noisy ''and'' gate [DINA] model applied by Cheng, 2009, 2010). This study used the reduced RUM only to illustrate the application of our proposed methods for its relatively low computational demand and a potentially suitable candidate as the basis of a real-time CAT program. In addition, the RUM has been proven useful in real-data analysis (e.g., Roussos, Hartz, & Stout, 2003), which could provide parameter estimates based on which a simulation study can be constructed. The item selection methods proposed in this study can be adapted for application in any other CDM.

Table 1. *Q* Matrix for a Hypothetical Arithmetic Test.

Item	Attribute		
	Subtraction with parenthesis	Subtraction when $a < b$ [$a-b = -(b-a)$]	Addition with different sign numbers
$3-(-6)$	1	0	0
$3-9$	0	1	0
$-8+8$	0	0	1
$-8-(-4)$	1	1	1
$-5-(-7)$	1	0	1
$-7+3$	0	1	1

Reduced RUM

The goal of diagnostic classification testing is to identify strengths and weaknesses among multiple attributes rather than to assess an examinee's overall proficiency in a particular scholastic area. CDMs have been developed to fulfill such diagnostic purposes. Available CDMs include the rule space model (Tatsuoka, 1983), DINA model (Haertel, 1989; Junker & Sijtsa, 2001), noisy input deterministic ''and'' gate (NIDA) model (Maris, 1999), and RUM (Hartz, 2002), as well as the model used in this study, namely, the reduced RUM (Roussos et al., 2007).

The aforementioned models intend to estimate examinees' latent attributes based on their item responses. Each examinee's attribute profile is defined by a vector $\tilde{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_k)$ of K skills, α_k being the examinee's cognitive state along attribute k. This study investigated the effectiveness of various item selection criteria in CD-CAT under the dichotomous attribute assumption. Accordingly, $\alpha_k = 1$ indicates mastery of attribute k, whereas $\alpha_k = 0$ indicates nonmastery. CDMs can be applied to map the skill structures of items and yield a Q matrix (Tatsuoka, 1983). Because an item pool contains J items, the Q matrix is a $(J \times K)$ matrix. The entry q_{jk} of a Q matrix equals 1 if item j measures attribute k, and zero otherwise. That is, the binary data for row j of a Q matrix specify whether attribute k is required for a test taker to complete item j or which attributes are measured by item j. Table 1 provides an illustration of a Q matrix for an arithmetic test. This Q matrix was designed to measure three attributes (or skills) with six items. For example, the first item, $3-(-6)$, measures only one skill, whereas Item 4, $-8-(-4)$, measures all three attributes.

To model the probability of a correct response for an examinee, the RUM defines slipping and guessing item parameters at the attribute level. This model originated in the NIDA model developed by Maris (1999), which, by separately estimating the slipping parameter s_{jk} and guessing parameter g_{jk}, defines the probability of a correct response to item j for examinee i as follows:

$$P\left(X_{ij}=1|\boldsymbol{\alpha}_i,\mathbf{s},\mathbf{g}\right)=\prod_{k=1}^{K}\left[\left(1-s_{jk}\right)^{\alpha_{ik}}g_{jk}^{1-\alpha_{ik}}\right]^{q_{jk}}. \quad (1)$$

To consider any attributes not included in the Q matrix, DiBello, Stout, and Roussos (1995) incorporated a continuous latent variable θ_i into Equation 1 to define the unified model. Subsequently, Hartz (2002) reparameterized the unified model by combining s_{jk} and g_{jk} to prevent unidentifiability in the unified model, and two new item parameters were yielded. The first of these parameters, π_j^*, defines the probability of a correct response to item j, given that the examinee has mastered all required attributes; this parameter is expressed as $\pi_j^* = \prod_{k=1}^{K} (1 - s_{jk})^{q_{jk}}$ The second parameter, r_{jk}^*, is a ''penalty'' for not mastering attribute k for item j and is expressed as $r_{jk}^* = \frac{g_{jk}}{1-s_{jk}}$ With these two parameters, the reduced version of the RUM (Roussos et al., 2007) can define the probability of a correct response as follows:

$$P(X_{ij} = 1 | \boldsymbol{\alpha}_i) = \pi_{\mathrm{ij}}^* \sum_{k=1}^{K} r_{jk}^{*(1-\alpha_{ik})q_{jk}}. \quad (2)$$

The reduced RUM omits the term of supplemental ability in the original RUM, thereby simplifying the model by assuming that no supplemental ability that may affect item responses is present.

Before our proposed methods for item selection is introduced, the KL-based GDI in CD-CAT (Xu et al., 2003) is reviewed in this article; our selection methods were constructed based on this index. The GDI method was chosen as a baseline in our study because KL divergence has been recognized as a legitimate concept applicable to cases where the latent traits are categorical (e.g., Eggen, 1999); many studies have used this concept as a basis for new criteria to improve item selection in estimation precision (e.g., Cheng, 2009) and to achieve balance between psychometric accuracy and nonpsychometric constraints (e.g., Cheng, 2010; Wang et al., 2011).

KL Information in CD-CAT

Fisher information (FI) and KL information are two of the most popular item selection criteria in CAT; however, FI is undefined in CD-CAT. FI defines the amount of information regarding unknown parameter θ carried by observable random variable X under the continuity requirement for the conditional distribution of X given θ. The discreteness of CD-CAT precludes the application of FI. By contrast, KL information is applicable to CD-CAT, where the latent structure involves categorical latent class.

KL information measures the divergence between two probability distributions $f(x)$ and $g(x)$ (Cover & Thomas, 1991) and is defined by

$$\mathrm{KL}(f \parallel g) = E_f \left[\log \frac{f(x)}{g(x)} \right] \quad (3)$$

so that the KL quantity increases as the two distributions diverge. Chang and Ying (1996) first suggested the use of the KL information in CAT research; ever since, many studies have used KL information in various computer-based testing contexts.

Chen, Ankenmann, and Chang (2000) used KL information in the early stage of CAT. Eggen (1999) and Lin (2011) have applied KL information to computerized classification testing. Henson and Douglas (2005) proposed KL-based divergence indices to assemble cognitive diagnostic tests. Wu et al. (2003) and Cheng (2009) have modified the KL index for CD-CAT applications.

In CD-CAT, information refers to an item's ability to discriminate between two attribute patterns in a pair. In this sense, the KL divergence measure in diagnostic classification should reflect the distance between two conditional distributions, namely, $f\left(X_{ij}|\hat{\boldsymbol{\alpha}}_i\right)$, the distribution of X_{ij} conditioning on the current estimated latent class, and $f\left(X_{ij}|\hat{\boldsymbol{\alpha}}_t\right)$, the conditional distribution of X_{ij} given the true state. This logic yields the following KL equation for CD-CAT:

$$\mathrm{KL}_j(\hat{\boldsymbol{\alpha}}_i \parallel \boldsymbol{\alpha}_t) = \sum_{x=0}^{1} \log\left(\frac{P\left(X_{ij}=x|\hat{\boldsymbol{\alpha}}_i\right)}{P\left(X_{ij}=x|\boldsymbol{\alpha}_t\right)}\right) P\left(X_{ij}=x|\hat{\boldsymbol{\alpha}}_i\right). \tag{4}$$

Considering that the true latent state is unknown and there are 2^k possible states, Xu et al. (2003) proposed a GDI formulated as follows:

$$\mathrm{GDI}_j(\hat{\boldsymbol{\alpha}}_i) = \sum_{c=1}^{2^K}\left[\sum_{x=0}^{1} \log\left(\frac{P\left(X_{ij}=x|\hat{\boldsymbol{\alpha}}_i\right)}{P\left(X_{ij}=x|\hat{\boldsymbol{\alpha}}_c\right)}\right) P\left(X_{ij}=x|\hat{\boldsymbol{\alpha}}_i\right)\right]. \tag{5}$$

This index is the sum of the KL distances between $P\left(X_{ij}|\hat{\boldsymbol{\alpha}}_i\right)$ and $P\left(X_{ij}|\hat{\boldsymbol{\alpha}}_c\right)$ for all possible latent states. Items with larger GDI values have a higher ability to discriminate between the estimated attribute pattern and all other possible cognitive profiles. In Xu et al. (2003), the GDI method performed well in recovering examinees' attribute profiles.

A downside of the maximum GDI method is that it does not consider attribute balancing or exposure control. Thus, we proposed a holistic index to consider these crucial practical constraints in CD-CAT by using the GDI as a basis for developing the proposed item selection methods.

Holistic Item Selection Criterion in CD-CAT

In this study, the following two indices were applied to formulate the holistic item selection criterion: (1) an attribute-balanced item selection criterion and (2) a progressive control algorithm. The main objective was to initially form an attribute-balanced item selection criterion and subsequently incorporate it into a progressive algorithm for item exposure control; this idea is introduced as follows.

Weighted Deviation Balancing Index for Attribute-Balanced Item Selection. Attribute balancing in CD-CAT may appear analogous to content balancing in conventional CAT with respect to test validity. However, in CD-CAT, an item can measure more than one attribute simultaneously. Assuming that the content areas in question are

mutually exclusive, most content-balancing methods in traditional CAT (e.g., Chen & Ankenmann, 2004; Cheng, Chang, & Yi, 2007) cannot be adapted for attribute balancing in CD-CAT. Therefore, Cheng (2010) developed an attribute-balancing index and multiplied it by the GDI to create a modified GDI (MGDI) able to yield adequate attribute coverage for CD-CAT. The current study further proposed a more general item selection index with a weighing scheme for attribute balancing in cases where some attributes are more crucial than others or where constraints other than attribute balancing may be required in CD-CAT.

In this study, the proposed attribute-balancing item selection criterion, namely, the weighted deviation GDI (WDGDI) was formulated as

$$\mathrm{WDGDI}_j(\hat{\alpha}_i) = (-\mathrm{WD}_j) \times \mathrm{GDI}_j(\hat{\boldsymbol{\alpha}}_i). \tag{6}$$

In a process conceptually similar to that used to create the MGDI, we multiplied a native value of a weighted deviation index (WD) by the GDI. The WD index, namely, the attribute-balancing index of the WDGDI method, originated from the weighted deviations model (WDM) heuristic developed by Swanson and Stocking (1993). The WDM approach bases the evaluation of each item on the positive deviations of its nonpsychometric and psychometric properties from those required on the target test (i.e., constraints). The goal of the WDM heuristic is to seek the items whose inclusion in a test generates the smallest weighted sum of positive deviations. Accordingly, items are selected sequentially so that those selected first yield the maximum improvement while simultaneously meeting all constraints.

The WDM heuristic considers the upper and lower boundaries around all target values or constraints to gain some degree of flexibility in meeting each constraint. Decisions regarding distance between lower and upper bounds are made at the discretion of test developers and specialists based on their rationales and needs to be achieved. In this study, we constrained each test to contain at least 10 items (minimum) and no more than the number equivalent to the test length (maximum) to measure each attribute because we were mainly concerned about whether the minimum number of items could be selected. The upper bound was specified as the test length to ensure that the minimum requirement would be fulfilled. Although this upper bound specification seemed unnecessary for our study, we formulated the upper bound in our WD index, as shown in Equation (7), because we intended to develop an item selection index that can be applied to various testing situations. For example, the upper bound is needed because the requirement is constrained to be a specific target value. Finally, weights can be specified for each constraint by using the WDM heuristic so that some constraints can be emphasized over others. More details regarding the weight selection are provided in the section titled simulation study.

When the WDM with an attribute outline is applied in CD-CAT, the WD for each item candidate j in the pool is computed as

$$\mathrm{WD}_j = \sum_{k=1}^{K} (\mathrm{W}_k \mathrm{D}_{jL_k}) + \sum_{k=1}^{K} (\mathrm{W}_k \mathrm{D}_{jU_k}), \tag{7}$$

where W_k is the weight for the kth attribute, and D_{jL_k} and D_{jU_k}, respectively correspond to the positive deviations from the minimal (i.e., lower boundary) and maximal (i.e., upper boundary) numbers of items required to assess the k^{th} attribute when item j is included in the test. For each constraint k, D_{jL_k} is defined as $(L_k - q_k)$ and D_{jU_k} is defined as $(q_k - U_k)$, where L_k and U_k, respectively denote the lower and upper bounds for the k^{th}-attribute constraint. The term q_k represents the expected number of items measuring the k^{th} attribute that would have been obtained if item candidate j was included in the test, assuming that the remaining item selections were random. According to the WDM heuristic, when q_k is smaller than L_k, D_{jL_k} equals $(L_k - q_k)$ and D_{jU_k} equals 0; when q_k is greater than U_k, D_{jL_k} equals 0 and D_{jU_k} equals $(q_k - U_k)$; and when q_k is within the lower and upper bounds of the attribute coverage requirement, D_{jL_k} and D_{jU_k} both equal zero, and thus the minimum and maximum requirements are met.

The inclusion of an item candidate with the smallest WD (or greatest WD) value in the test can be expected to yield the greatest improvement toward fulfilling the attribute-balancing requirement. Consequently, the WDGDI, formulated as Equation (6), appears able to ensure attribute balancing and diagnostic precision. Although this study concerned only attribute constraints in the WDM, the model in Equation (7) provided the option of incorporating other nonpsychometric constraints such as content outline and answer choice specification.

To place the WD and the GDI metrics on an equal footing, we standardized the WD and GDI values, and the final attribute-balancing item selection index became

$$\mathrm{SWDGDI}_j(\hat{\boldsymbol{\alpha}}_i) = \left(\frac{\mathrm{Max}(\mathrm{WD}_j) - \mathrm{WD}_j}{\mathrm{Max}(\mathrm{WD}_j) - \mathrm{Min}(\mathrm{WD}_j)}\right) \times \left(\frac{\mathrm{GDI}_j(\hat{\boldsymbol{\alpha}}_i) - \mathrm{Min}\left(\mathrm{GDI}_j(\hat{\boldsymbol{\alpha}}_i)\right)}{\mathrm{Max}\left(\mathrm{GDI}_j(\hat{\boldsymbol{\alpha}}_i)\right) - \mathrm{Min}\left(\mathrm{GDI}_j(\hat{\boldsymbol{\alpha}}_i)\right)}\right) \tag{8}$$

Greater standardized GDI information represents more psychometric attractiveness for an item. Notably, the standardized WD is computed by $\mathrm{Max}(\mathrm{WD}_j) - \mathrm{WD}_j$ as opposed to $\mathrm{WD}_j - \mathrm{Min}(\mathrm{WD}_j)$, such that a greater standardized WD indicates a greater contribution made by an item to satisfy the attribute-balancing requirement. The product obtained from the standardized WDGDI (SWDGDI) synchronizes the nonpsychometric and psychometric attractiveness of an item. Under consideration for the attribute balancing, an item with the greatest SWDGDI is selected first in the test as opposed to one with the maximum GDI. Rather than the standard deviation, we used the simplest measure of variability, namely, range (i.e., maximum–minimum), for the standardization in Equation 8, mainly to render the mathematical form of the denominator similar to that of the numerator. Future studies may use the standard deviation for standardization.

Constrained Progressive Algorithm. A progressive exposure control algorithm was the second element—or more specifically, the framework of our holistic index—in charge of balancing the exposure rate. How to balance severely uneven item pool usage is always an issue when a maximum information method is applied in adaptive testing. Items with greater information may be administered excessively frequently and become overexposed, leading to a test security breach and compromising test validity. Less informative items are rarely chosen and become underexposed, which raises an economic concern. In previous decades, numerous exposure control techniques were proposed and categorized (Georgiadou, Triantafillou, & Economides, 2007). In general, such techniques address exposure control by aiming to suppress overexposure (e.g., Chen, Lei, & Liao, 2008), boost usage of the underexposed (e.g., Revuelta & Ponsoda, 1998), or both (e.g., Chang, Qian, & Ying, 2001; Wang et al., 2011).

The progressive method of Revuelta and Ponsoda (1998) was used as a template for our holistic item selection index because of its intention to increase the usage of barely used items. Progressive control involves a randomization scheme in the item selection criterion to diminish the possibility of selecting items with the greatest information. Under consideration of this aspect, we modified the progressive method to further suppress overexposure by adding one more stochastic component (R_{jI}) to the item selection criterion, expressed as follows:

$$\text{P\_INFO}_j = \left(1 - \frac{X}{L}\right) R_j + \frac{X}{L} \times R_{jI}, \tag{9}$$

where X equals the number of items already administered, L denotes the test length, and R_j is generated from uniform (0, maximum information). The term $\left(1 - \frac{X}{L}\right) R_j$ allows for a greater randomization impact in the early stage of testing, and this relaxes the precision demand. As the test progresses, the information progressively comes into play in the item selection process. The resultant item exposure can be controlled without seriously compromising estimation precision due to the initial lack of information regarding the examinee's proficiency.

The current index differs from the original progressive method in that we replaced the fixed information quantity with a random draw from an information interval. The term R_{jI} refers to a random draw from a uniform distribution bounded by the lower and upper limits of an information interval, or R_{jI} is generated from uniform (LB_j, UB_j). The information interval is computed from item j by defining the lower bound as $LB_j = \text{Info}_j - (\text{Info}_j - \text{Min})/s$ and the upper bound as $UB_j = \text{Info}_j + (\text{Max} - \text{Info}_j)/s$ with $1 \leq s \leq \infty$, where Info_j is the information calculated from item j and Min and Max are the maximum and minimum item information in the pool, respectively. The term s is an interval-adjusting factor. Notably, when $s = 1$, item selection is completely random. When $s = \infty$, denoting that the interval does not exist, $\text{P\_INFO}_j = \left(1 - \frac{X}{L}\right) R_j + \frac{X}{L} \times \text{Info}_j$, and our modified index becomes the index of Revuelta and Ponsoda (1998). This modified index renders item selection more flexible by adjusting the width of the information interval via s. The specification of s is

sometimes arbitrary, depending on the testing purpose. A lower *s* value results in a wider interval and a greater random effect on the information part, indicating that greater test security than estimation precision is demanded. By contrast, a higher *s* value results in a narrower interval and weaker random effect on the information part, indicating that greater estimation precision than test security is demanded.

Wang et al. (2011) proposed another modification of the progressive algorithm proposed by Revuelta and Ponsoda (1998) by adding an importance parameter β to adjust for the balance between exposure rate distribution and estimation accuracy. Despite the similarity between parameter β and the *s* factor, the proposal for the interval-adjusting factor *s* expresses an intention to use the randomization scheme throughout the item selection process for exposure control. Moreover, as described in the following section, the major difference between the proposed index of Wang et al. (2011) and our proposed index lies in the type of information imposed in the progressive algorithm.

Holistic Index. To simultaneously manage test security and validity, we incorporated the attribute-balanced information $\mathrm{SWDGDI_j}(\hat{\boldsymbol{\alpha}}_i)$instead of the purely psychometric information (e.g., Wang et al., 2011) into the progressive algorithm expressed as Equation 9. The resultant integrated item selection index, namely, the progressive SWDGDI, is expressed as

$$\mathrm{P\_SWDGDI}_j(\hat{\boldsymbol{\alpha}}_i) = \left(1 - \frac{X}{L}\right) R_j + \frac{X}{L} \times R_{jI}, \tag{10}$$

where in R_{jI}, $LB_j = \mathrm{SWDGDI}_j(\hat{\boldsymbol{\alpha}}_i) - \left(\mathrm{SWDGDI}_j(\hat{\boldsymbol{\alpha}}_i) - \mathrm{Min}\right)/s$ and $UB_j = \mathrm{SWDGDI}_j(\hat{\boldsymbol{\alpha}}_i) + \left(\mathrm{Max} - \mathrm{SWDGDI}_j(\hat{\boldsymbol{\alpha}}_i)\right)/s$.

Similar to the approach of Wang et al. (2011), our progressive SWDGDI approach substantially reduces the number of unexposed items, whereas the maximum exposure rate is high in our study (e.g., 0.89 in Table 5). To suppress overexposure and constrain the maximum exposure rate, or *r*_max, under a certain level, we multiplied$\mathrm{P_SWDGDI}_j(\hat{\boldsymbol{\alpha}}_i)$ by a dynamic exposure parameter. The final item selection criterion was named constrained $\mathrm{P\_SWDGDI}_j(\hat{\boldsymbol{\alpha}}_i)$ $\left(\text{i.e., } \mathrm{CP\_SWDGDI}_j(\hat{\boldsymbol{\alpha}}_i)\right)$ and became

$$\mathrm{CP\_SWDGDI}_j(\hat{\boldsymbol{\alpha}}_i) = \frac{r\_\max}{r_j} \times \mathrm{P\_SWDGDI_j}(\hat{\boldsymbol{\alpha}}_i) = \frac{r\_\max}{r_j} \times \left[\left(1 - \frac{X}{L}\right) R_j + \frac{X}{L} \times R_{jI}\right] \tag{11}$$

The exposure parameter $\frac{r\_\max}{r_j}$ is dynamic and requires no intensive simulation because the item exposure rate for item *j* (i.e., r_j) is updated as the test progresses (Chen et al., 2008). Specifically, if an item is frequently selected and its current exposure rate (i.e., r_j) rapidly approaches the constrained *r*_max, its $\frac{r\_\max}{r_j}$ value soon decreases toward 1 from a very high level, thereby reducing the likelihood of this item being selected. That is, an item with greater information of $\mathrm{P\_SWDGDI}_j(\hat{\boldsymbol{\alpha}}_i)$

tends to be frequently administered, resulting in a lower $\frac{r\_\max}{r_j}$ value; however, its overexposure could be suppressed by multiplying its $\text{P\_SWDGDI}_j(\hat{\boldsymbol{\alpha}}_i)$ with the corresponding lower $\frac{r\_\max}{r_j}$.

Simulation Study

This study ran a simulation to evaluate the proposed holistic item selection method, namely, comparing the effectiveness of the new criteria, CP_SWDGDI and SWDGDI, to that of the GDI method. Simulated item pools and examinees were generated in a manner consistent with or similar to those adopted in previous studies. The evaluation criteria included attribute recovery rates, attribute-balanced percentages, and exposure control statuses. The anticipated variables that could have influenced CD-CAT characteristics were the test length and number of attributes. The item selection methods were compared under these experimental conditions. Detailed descriptions of item pool construction and examinee generation are described in the following sections.

Item Pool Construction and Examinee Generation

This study simulated item pools (i.e., the Q matrices and item parameters) and examinees (i.e., $\boldsymbol{\alpha}$ matrices**)** in a manner consistent with those adopted in previous studies (e.g., Cheng, 2009; Finkelman, Kim, Roussos, & Verschoor, 2010; Henson & Douglas, 2005 Wang & Chang, 2011). We simulated two item pools, each containing 400 items, based on the reduced RUM model. First, the reduced RUM model requires a ($J\times K$) Q matrix for each item pool to specify which attributes are to be measured by each item. In this study, one pool measured four attributes and required a (400 $\times$ 4) Q matrix, whereas the other pool measured six attributes and required a (400 $\times$ 6) Q matrix. We assumed independence among the items and that the K attributes being diagnosed were uncorrelated. Thus, the entry q_{jk} of the Q matrix for each item pool was generated separately for each item under the constraint that each item was to assess 30% of the attributes on average and measure at least one attribute. The resultant 400 attribute patterns in each Q matrix were regarded as a random sample from a population of all possible attribute patterns conforming to the attribute-balancing constraint. Second, for the reduced RUM model, the item parameters for each pool were generated in a fashion very similar to those reported in previous studies such as Henson and Douglas (2005) and Finkelman et al. (2010); π_j^* parameters were simulated from a uniform (0.75, 0.95) distribution and r_{jk}^* parameters were simulated from another uniform (0.20, 0.95) distribution.

A 2000 $\times$ 4 matrix $\boldsymbol{\alpha}$ and 2000 $\times$ 6 matrix $\boldsymbol{\alpha}$ were generated with the $\tilde{\alpha}$ vectors representing the true attribute patterns of 2000 examinees. Specifically, 2000 multivariate normal k-dimensional vectors ($\tilde{\alpha}\sim\text{MVN}(0, \rho)$) were generated, where $\boldsymbol{\rho}$ was a $K\times K$ variance–covariance matrix with K = 4 or 6. Under the assumption of independence among attributes for simplicity, the off-diagonal elements in $\boldsymbol{\rho}$ equaled 0

Table 2. Number of Items Measuring (or Examinees Mastering) Each Attribute.

$K = 4$	1	2	3	4		
Number of items	158	172	160	161		
Number of examinees	983	1,002	1,008	1,001		
$K = 6$	1	2	3	4	5	6
Number of items	138	143	132	138	132	136
Number of examinees	982	980	1,001	956	1,025	1,018

and the diagonal elements equaled 1. Additionally, the probability of mastery was assumed to be 50% for each attribute, and thus, the i^{th} individual's mastery for attribute k was

$$\alpha_{ik} = \begin{cases} 1 & \text{if } \tilde{\alpha} \leq 0 \\ 0 & \text{otherwise} \end{cases}. \tag{12}$$

Tables 2 and 3 report the distributions of items and examinees over attributes for both attribute-number conditions. Table 2 shows that the number of items measuring each attribute was approximate over the attributes, as was the number of examinees who mastered each attribute. These findings resulted from the independence assumption among the attributes. Table 3 reveals that most items measured one or two attributes under both attribute-number conditions because of the constraint imposed on Q matrix generation that on average, each item was to measure 30% of the attributes; that is, $30\% \times 4 = 1.2$ attributes per item for $K = 4$ and $30\% \times 6 = 1.8$ attributes per item for $K = 6$. Table 3 also gives the numbers of examinees who mastered all possible numbers of attributes. Most examinees mastered one to three attributes with a mode of two attributes for $K = 4$ and two to four attributes with a mode of three attributes for $K = 6$. These findings confirmed our expectation because the examinee attribute patterns were derived from the multivariate normal distribution.

Estimation of Attribute Patterns

The current study applied the posterior mode to classify an individual's attribute pattern so that the prior distribution for each pool was constructed through generation of a further 10000 multivariate normal k-dimensional vectors($\tilde{\alpha} \sim \text{MVN}(0, \boldsymbol{\rho})$) with α_{ik} defined in the same manner as in Equation 12. The prior probabilities of all possible attribute patterns were estimated in this manner.

Item responses were simulated using the Monte Carlo method based on the reduced RUM. Notably, the initial attribute pattern estimate was randomly produced with approximately half 0s and half 1s. Based on the item responses, the posterior mode estimate of the cognitive profile was updated. Specifically, with the item

Table 3. Number of Items Measuring (or Examinees Mastering) Each Number of Attributes.

$K = 4$	0	1	2	3	4		
Number of items	0	199	154	44	3		
Number of examinees	126	508	731	516	119		
$K = 6$	0	1	2	3	4	5	6
Number of items	0	139	143	85	26	7	0
Number of examinees	35	191	477	613	468	190	26

parameters known, attribute pattern estimation was performed by calculating the likelihood of each possible attribute pattern based on the examinees' item responses and multiplying the results by the prior probabilities.

CD-CAT Simulation Condition

The independent variables in this study included the item selection method, test length, and number of attributes. Two test lengths were considered: a short test (32 items) and long test (60 items). Two numbers of attributes were specified: 4 and 6. Three item selection methods were applied: (1) GDI information, (2) the attribute-balanced information index (SWDGDI), and (3) the attribute-exposure-balanced information index (CP_SWDGDI). This design yielded 12 ($2 \times 2 \times 3$) experimental conditions. In method (2), this study constrained each test to have at least 10 items (minimum) and no more than the number equivalent to the test length (maximum) to enable measurement of each attribute under each experimental condition. Assuming that all attributes were of equal importance, a set of equal weights was applied and a value of 1 was assigned to each attribute. In method (3), the specifications for attribute balancing were identical to those in method (2), and the maximum exposure rate was set to 0.25. Although the s parameter in method (3) was adjustable according to various practical purposes, a value of 1.6 was assigned to it under all conditions to enable fair comparison; 1.6 was selected because it produced a reasonable trade-off between psychometric precision and nonpsychometric balancing under all conditions.

Weight Selection

The weight assigned to each test specification (e.g., the number of items required to measure attribute 1 was 10) was typically 1 when the specifications were identical or at least similar to the proportion of items in the pool. Moreover, so long as the pool contained sufficient items with relevant characteristics, it was expected that almost all the test specifications would be readily satisfied no matter how the weights were applied (i.e., either equal or unequal weights). However, some specifications were

difficult to satisfy when fewer items than expected in the pool had the relevant characteristics. Under such circumstances, a test specialist can prioritize these specifications by weighing them more when they are desired properties.

In addition to the constraints specified in the previous section, we simulated a condition characterized by difficulty fulfilling some specifications in a manner that enabled evaluation of the impact of weight selection on the results by using the CP_SWDGDI method. We evaluated only the effect of weight selection under the experimental condition of the short test with four attributes for illustration. Instead of constraining each test to obtain an equivalent minimum number of items across all attributes, the lower bounds for the number of items for attributes 1 to 4 were specified as 10, 10, 10, and 23, respectively. The weight assigned to each attribute was 1 under the equal weight condition. Under the unequal weight condition, the weight was 2 for attribute 4 and 1 for each of the other attributes to ensure that at least 23 items measured attribute 4. Weight selection is arbitrary, and test constructors may test several weights to obtain a satisfactory outcome.

Evaluation Criteria

The evaluation criteria included psychometric precision and the degree of nonpsychometric balancing in CD-CAT. Psychometric precision is evaluated based on the entire pattern recovery rate and recovery rate for each attribute. The entire pattern recovery rate refers to the proportion of examinees in the sample with estimated attribute pattern $\hat{\boldsymbol{\alpha}}$ identical to the true pattern $\boldsymbol{\alpha}$ for all attributes. The attribute recovery rate for individual attribute k is defined as the proportion of examinees with estimated mastery status matching true status for attribute k in the sample.

This study examined nonpsychometric balancing in relation to attribute coverage and item exposure status. The degree of attribute balancing was measured by the percentage of tests that fulfilled the attribute coverage constraints at the attribute level and entire test level. To assess item exposure balance, the maximum exposure rate and number of unused items were recorded, and the chi-square statistic was computed as $\chi^2 = \sum_{j=1}^{n} \frac{\left(r_j - \bar{r}\right)^2}{\bar{r}}$ (Chang & Ying, 1999), where r_j is the item exposure rate for item j and $\bar{r}$ is the average exposure rate. This index measures the evenness of item exposure rates over items; the lower the chi-square value, the more similar are the exposure rates.

Results

To simplify the interpretation of the results, special emphasis was placed on comparisons among item selection methods across various experimental conditions. Tables 4 to 6 present the recovery rates, item exposure statuses, and percentages of attribute-balanced tests for three item selection methods for various test-length and attribute-number conditions, respectively. Of the six rightmost columns in each of these tables,

Table 4. Recovery Rate for Each Attribute and the Entire Cognitive Pattern.

	Short test (32 items)			Long test (60 items)		
Method	GDI	SWDGDI	CP_SWDGDI	GDI	SWDGDI	CP_SWDGDI
$K = 4$						
Attribute 1	0.98	0.99	0.96	1.00	1.00	0.98
Attribute 2	0.97	0.98	0.96	0.99	1.00	0.98
Attribute 3	0.99	0.98	0.96	1.00	1.00	0.98
Attribute 4	0.99	0.99	0.97	1.00	1.00	0.98
Entire pattern	0.93	0.95	0.87	0.98	0.99	0.93
$K = 6$						
Attribute 1	0.93	0.94	0.90	0.98	0.98	0.95
Attribute 2	0.96	0.96	0.93	0.99	0.99	0.97
Attribute 3	0.98	0.96	0.93	0.99	0.99	0.96
Attribute 4	0.96	0.96	0.92	0.99	0.99	0.95
Attribute 5	0.96	0.96	0.92	0.98	0.99	0.97
Attribute 6	0.95	0.95	0.92	0.99	0.99	0.96
Entire pattern	0.77	0.78	0.64	0.92	0.94	0.80

Note. GDI = global discrimination index; SWDGDI = standardized weighted deviation global discrimination index; CP_SWDGDI = constrained progressive standardized weighted deviation global discrimination index.

the first three display the results for the short test condition, whereas the final three show the outcomes for the long test condition.

The results in Table 4 show that when an attribute-balancing index was added to the solely information-based criterion, the SWDGDI method slightly outperformed the GDI method by gaining an individual attribute and overall precision over the GDI method. The difference in the entire pattern recovery rate is more evident than that in the individual attribute recovery rate because entire pattern recovery results from correct recovery for all attributes and a gain at the attribute level aggregates; this result is in line with the corresponding result of Cheng (2010).

Table 4 reveals that the CP_SWDGDI method performed worst in terms of recovering the individual attributes and the entire pattern, mainly because of its item-exposure-control mechanism. As shown in Table 5, the CP_SWDGDI method controls the maximum exposure rate at the prespecified 0.25 level, uses all items in the pool (number of unused items = 0), and yields a much lower chi-square value (e.g., 34.28 vs. 163.81 and 164.66). However, the GDI and SWDGDI methods yield maximum exposure rates at least 0.89, substantial numbers of unused items (more than 200), and relatively high chi-square values. Taken together, the results for psychometric precision and the exposure control indices show that the CPI_SWDGDI method successfully evens item usage alongside the ''side effect'' of precision loss under most conditions; this represents a trade-off between exposure control and measurement precision that occurs in most CAT methods. These results are generalizable across various test-length and attribute-number conditions. The attribute

Table 5. Exposure Balance Measures.

	Short test (32 items)			Long test (60 items)		
Method	GDI	SWDGDI	CP_SWDGDI	GDI	SWDGDI	CP_SWDGDI
$K = 4$						
Maximum r	0.97	0.91	0.25	1.00	1.00	0.25
No. of unused	289	288	0	251	251	0
χ^2	163.81	164.66	34.28	182.10	182.46	23.42
$K = 6$						
Maximum r	0.97	0.89	0.25	0.99	0.99	0.25
No. of unused	240	233	0	204	204	0
χ^2	157.94	149.01	26.54	157.72	161.20	22.03

Note. GDI = global discrimination index; SWDGDI = standardized weighted deviation global discrimination index; CP_SWDGDI = constrained progressive standardized weighted deviation global discrimination index.

classification precision performance of the CP_SWDGDI method may be improved by applying less stringent exposure control; a demonstration of such application is described subsequently.

Considerably larger differences in psychometric precision were observed among the attribute-number conditions, whereas item exposure statuses were similar. Taking the short length as an example, as the number of attributes increased, the GDI, SWDGDI, and CP_SWDGDI reduced the entire recovery rate by 0.16 (from 0.93 to 0.77), 0.17 (from 0.95 to 0.78), and 0.23 (from 0.87 to 0.64), respectively. These results indicated that the CP_SWDGDI method produced the greatest difference of all the methods. Although this difference pattern is generalizable across all test-length conditions, the degree of precision loss decreased as the test length increased. For example, for the long test length, the GDI, SWDGDI, and CP_SWDGDI reduced the entire recovery rate by 0.06 (from 0.98 to 0.92), 0.05 (from 0.99 to 0.94), and 0.13 (from 0.93 to 0.80), respectively. As such, given that measurement precision can be improved by adding more items to the test, this advantage of an increasing test length becomes more evident as the number of attributes increases.

Table 6 compares the percentages of tests that fulfilled the attribute coverage constraints at the attribute level and entire test level as obtained from the three item selection methods. For example, the first entry for the GDI in Table 6, namely, 0.85 denotes that 85% of the tests under the GDI method fulfilled the coverage constraints of the first attribute, or 85% of the tests have at least 10 items measuring the first attribute. The SWDGDI and CP_SWDGDI methods yielded perfect attribute balancing, with 100% of the tests under these conditions fulfilling all attribute coverage constraints, or 100% of these tests having 10 or more items measuring each of the four attributes. The gain of these two methods over the GDI method in terms of attribute balancing was greater at the entire test level. Under the four-attribute condition, only 31% of the short tests were attribute balanced under the GDI, whereas the other

Table 6. Percentages of Attribute-Balanced Tests.

	Short test (32 items)			Long test (60 items)		
Method	GDI	SWDGDI	CP_SWDGDI	GDI	SWDGDI	CP_SWDGDI
K = 4						
Attribute 1	0.85	1.00	1.00	1.00	1.00	1.00
Attribute 2	0.66	1.00	1.00	0.92	1.00	1.00
Attribute 3	0.84	1.00	1.00	1.00	1.00	1.00
Attribute 4	0.87	1.00	1.00	0.99	1.00	1.00
Entire pattern	0.31	1.00	1.00	0.91	1.00	1.00
K = 6						
Attribute 1	0.37	1.00	1.00	0.90	1.00	1.00
Attribute 2	0.60	1.00	1.00	0.95	1.00	1.00
Attribute 3	0.69	1.00	1.00	0.96	1.00	1.00
Attribute 4	0.53	1.00	1.00	0.93	1.00	1.00
Attribute 5	0.44	1.00	1.00	0.91	1.00	1.00
Attribute 6	0.59	1.00	1.00	0.94	1.00	1.00
Entire pattern	0.01	1.00	1.00	0.65	1.00	1.00

Note. GDI = global discrimination index; SWDGDI = standardized weighted deviation global discrimination index; CP_SWDGDI = constrained progressive standardized weighted deviation global discrimination index.

two methods ensured 100% attribute balancing. These findings are generalizable to all test-length and attribute-number conditions.

Table 6 shows that the GDI method produced the worst attribute-balancing statuses under the conditions of a shorter test length and larger number of attributes. Under the GDI, 91% of the long tests but only 31% of the short tests exhibited adequate coverage of all four attributes. When the number of attributes equaled six, the decreasing trend of test length was almost parallel to that under the four-attribute condition; however, the percentage of balanced tests was consistently considerably smaller across all test-length conditions—65% of the long tests but only 1% of the short tests attained the required level of attribute coverage. Therefore, the advantage of attribute balancing in the SWDGDI and CP_SWDGDI methods became more pronounced as the test length decreased and the number of attributes increased.

Table 7 provides information regarding the effect of weight selection on the attribute recovery rate, exposure balance status, and percentage of attribute-balanced tests by using the CP_SWDGDI method. Given that the specification of at least 23 items measuring attribute 4 was prioritized and weighted more, the unequal weight condition generated 100% congruence in this specification, whereas with 97%, the equal weight condition did not. The differences in the other evaluation criteria between the equal and unequal weight conditions were negligible. This article provides only preliminary information regarding the weighing effect. In-depth investigation may be conducted by considering weight selection as an independent variable when more nonpsychometric constraints are incorporated into CD-CAT.

Table 7. Recovery Rates, Exposure Balance Measures, and Percentages of Attribute-Balanced Tests (PABT) of Various Weighing Schemes Under the Condition of a Short Test With Four Attributes by Using the CP_SWDGDI Method.

	Recovery rate and exposure balance		PABT	
	Equal weight	Unequal weight	Equal weight	Unequal weight
Attribute 1	0.92	0.92	1.00	1.00
Attribute 2	0.93	0.93	1.00	0.98
Attribute 3	0.94	0.94	1.00	0.99
Attribute 4	0.98	0.97	0.97	1.00
Entire pattern	0.81	0.80	0.97	0.97
Maximum *r*	0.25	0.25		
No. of unused	0.00	0.00		
χ^2	40.18	40.43		

Taken as a whole, the results for attribute balancing highlight the effectiveness of the SWDGDI and CP_SWDGDI methods, and the outcomes for the exposure balance indices highlight the effectiveness of the CP_SWDGDI method. The SWDGDI method was developed to balance attribute coverage and realized this goal by achieving attribute balancing in 100% of tests, with attribute classification rates as high as those of the GDI method. Although the CP_SWDGDI method also proved highly effective in balancing attribute coverage by achieving attribute balancing in 100% of tests, this method did not demonstrate a high attribute classification rate. The lowest recovery rate was yielded by the CP_SWDGDI method under the test-assembly conditions of a shorter test and greater number of attributes; this phenomenon may have been a result of the stringent exposure constraint, given that the value of *s* was specified as 1.6 and the maximum item exposure rate (i.e., *r*_max) was constrained to 0.25.

CD-CAT is usually employed for a low-stack setting, and thus a relatively high maximum item exposure rate is allowed. To improve the attribute classification precision of the CP_SWDGDI method and demonstrate that this method can simultaneously balance attribute coverage and item pool usage while maintaining acceptable estimation precision, we relaxed the exposure specifications in the CP_SWDGDI algorithm by imposing a greater *r*_max (0.60) and *s* (6) under the condition of a short test with six attributes. Table 8 contrasts the outcome of the original exposure specification with that of the less stringent exposure specification in terms of attribute recovery rate, exposure balance status, and percentage of attribute-balanced tests. The results shown in Table 8 indicated that as the exposure specifications were relaxed, the recovery rates of the CP_SWDGDI method increased uniformly across each attribute and the entire pattern were almost as high as those of the GDI method. Moreover, the CP_SWDGDI method with a less stringent exposure specification achieved attribute balancing in 100% of tests, used all the items in the pool (number

Table 8. Recovery Rates, Exposure Balance Measures, and Percentages of Attribute-Balanced Tests (PABT) for the GDI and CP_SWDGDI Methods With Two Specifications Under the Condition of a Short Test With Six Attributes.

	Recovery rate and exposure balance			PABT		
Method	GDI	CP_SWDGDI	CP_SWDGDI	GDI	CP_SWDGDI	CP_SWDGDI
Specification	$s = 1.6$	$s = 1.6$, $r = 0.25$	$s = 6$, $r = 0.6$	$s = 1.6$	$s = 1.6$, $r = 0.25$	$s = 6$, $r = 0.6$
Attribute 1	0.93	0.90	0.97	0.37	1.00	1.00
Attribute 2	0.96	0.93	0.95	0.60	1.00	1.00
Attribute 3	0.98	0.93	0.94	0.69	1.00	1.00
Attribute 4	0.96	0.92	0.95	0.53	1.00	1.00
Attribute 5	0.96	0.92	0.95	0.44	1.00	1.00
Attribute 6	0.95	0.92	0.94	0.59	1.00	1.00
Entire pattern	0.77	0.64	0.76	0.01	1.00	1.00
r	0.97	0.25	0.60			
No. of unused	240	0	0			
χ^2	157.94	26.54	89.31			

Note. GDI = global discrimination index; SWDGDI = standardized weighted deviation global discrimination index; CP_SWDGDI = constrained progressive standardized weighted deviation global discrimination index. No maximum r is specified under the GDI method; "r" denotes maximum r.

of unused items = 0), and yielded a considerably lower chi-square value than did the GDI method (i.e., 89.31 vs. 157.94).

Conclusion and Discussion

This research supports our proposal of a single item selection criterion that balances attribute coverage and the exposure rate without a severe loss in estimation accuracy. The results revealed that the SWDGDI method successfully balanced attribute coverage in CD-CAT. Furthermore, the CP_SWDGDI method simultaneously achieved balance over attribute coverage and ensured test security by reducing the maximum exposure rate and number of unused items while maintaining acceptable measurement precision.

The advantageous features of the SWDGDI method include its weighing scheme and the capability to incorporate a variety of nonpsychometric constraints. The SWDGDI index was subsequently incorporated into the progressive algorithm to create the CP_SWDGDI index. To characterize the CP_SWDGDI method as an attribute-exposure-balanced item selection criterion, the SWDGDI does not incorporate additional nonpsychometric requirements other than attribute balancing in this study, despite its ability to serve multiple purposes. However, the success of the CP_SWDGDI method eventually depends on whether the item pool can support a test with the required properties. Notably, CD-CAT is usually applied for a low-stack setting. When an item pool can barely support CD-CAT, a greater maximum

exposure rate may be specified or greater weights may be placed on the prioritized specifications within the CP_SWDGDI index to produce acceptable psychometric precision and nonpsychometric balancing in CD-CAT (see Tables 7 and 8).

This study showed that the SWDGDI and CP_SWDGDI methods are more crucial when the test in question is short or designed for a large number of attributes. This is because without attribute-balancing constraints, lower percentages of attribute-balanced tests tend to be associated with short tests and tests measuring more attributes.

The CP_SWDGDI method incorporates two crucial parameters. The *s* parameter adjusts the information interval and the exposure parameter specifies the desired maximum exposure rate according to the specific testing purpose. Consequently, the CP_SWDGDI method successfully controls the maximum exposure rate, which benefits from the SH-based dynamic exposure parameter, and uses all items for CD-CAT, which benefits from the stochastic progressive technique that we currently proposed in this article.

The results of our preliminary analyses revealed that the number of unused items became zero when the CP_SWDGDI method was applied, no matter which value of *s* was specified. As the value of *s* increased, so did the estimation precision, albeit with more uneven item pool usage, and vice versa. The preliminary results also revealed that the maximum exposure rate can be controlled at any desired level through adjustment of the dynamic exposure parameter and that larger maximum exposure rates yield greater estimation precision, and vice versa. These results of adjusting the two parameters confirm the flexibility of the CP_SWDGDI method in balancing the exposure control requirement and psychometric precision. Practitioners may select specific *s* values and maximum exposure rates to serve their purposes.

Particularly of note is the significant decrease in the entire recovery rate alongside the increase in the number of attributes. This occurred because with independent attributes, the entire recovery rate is the product of the individual attribute recovery rates based on the probability multiplicative rule. Incorporating more items into a test may to some extent compensate for the precision loss resulting from an increase in the number of attributes in CD-CAT. However, longer tests may compromise the measurement efficiency promoted by adaptive testing. The trade-off between measurement precision and test length should be carefully considered.

This research recommends that for psychometric accuracy, CD-CAT should not cover a broad scope of knowledge or learning materials. Classroom assessment may be suitable for CD-CAT application; for example, similar to formative assessment, CD-CAT can be imbedded within the instructional process. CD-CAT with relatively few attributes can still collect compressive information for classroom learning and teaching if tests are administered periodically. CD-CAT enables students to learn what they specifically do well and obtain specific suggestions for improvement so that they may reach higher levels of learning. Educators can use CD-CAT to adjust learning objectives and instructional strategies through information concerning what students have learned and in which areas they are struggling.

This study could be expanded in future research. The current study shows that the SWDGDI and CP_SWDGDI methods can be good candidates for item selection in CD-CAT based on a simulation study. However, the simulation may have confined the generation of results to specific conditions. Therefore, to further justify the effectiveness of the methods proposed herein, future research should involve real application of the proposed two methods of CD-CAT by using real item pools. Second, this study assumed independence among attributes and defined the same cut-off point across all attributes for the sake of simplicity and interpretability. Future studies may examine the effectiveness of the currently proposed item selection criteria under more realistic conditions such as correlated attributes and different cutoff points across all attributes. One possible investigation is how correlated attributes and various mastery difficulty levels affect attribute-number selection in CD-CAT. Third, another expansion of this simulation study could incorporate various nonpsychometric constraints in the attribute-balancing index with various weighing schemes (i.e., equal vs. unequal) to comprehensively evaluate the SWDGDI and the CP_SWDGDI methods' effectiveness for regulating balance among all requirements. Fourth, the logic of forming the SWDGDI and CPI_SWDGDI indices could be adapted to other item selection criteria such as the method based on expected Shannon entropy or that based on posterior-weighted KL information. Fifth, this study focused on fixed-length CD-CAT. Application of the proposed item selection methods in variable-length CD-CAT is a key consideration because inquiries are rarely made into variable-length testing despite its vivid ''tailored'' nature. Finally, the distance between the upper and lower bounds was fairly large in the current study because each test was constrained to have at least 10 items measuring each attribute. To further evaluate the performance of our proposed methods, future studies may consider a smaller distance between the upper and lower bounds or a fixed number of items under the condition of the constraint being a specific target value.

Declaration of Conflicting Interests

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

Funding

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

References

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

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

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

Chen, S., & Ankenmann, R. D. (2004). Effects of practical constraints on item selection rules at the early stages of computerized adaptive testing. *Journal of Educational Measurement*, *41*, 149-174.

Chen, S. Y., Ankenmann, R. D., & Chang, H. (2000). A comparison of item selection rules at the early stages of computerized adaptive testing. *Applied Psychological Measurement*, *24*, 241-255.

Chen, S. Y., Lei, P. W., & Liao, W. H. (2008). Controlling item exposure and test overlap on the fly in computerized adaptive testing. *British Journal of Mathematical and Statistical Psychology*, *61*, 471-492.

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

Cover, T. M., & Thomas, J. A. (1991). *Elements of information theory*. New York, NY: Wiley.

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

Eggen, T. (1999). Item selection in adaptive testing with the sequential probability ratio test. *Applied Psychological Measurement*, *23*, 249-261.

Finkelman, M., 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.

Georgiadou, E., Triantafillou, E., & Economides, A. (2007). A review of item exposure control strategies for computerized adaptive testing developed from 1983 to 2005. *Journal of Technology, Learning, and Assessment*, *5*, 4-38.

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

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

Henson, R. A., & 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.

Lin, C.-J. (2011). Item selection criteria with practical constraints for computerized classification testing. *Educational and Psychological Measurement*, *71*, 20-36.

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

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

Romàn, A. I. S. (2009). *Fitting cognitive diagnostic assessment to the Concept Assessment Tool for Statics (CATS)* (Unpublished doctoral dissertation). Purdue University, Lafayette, IN.

Roussos, L. A., DiBello, L. V., Stout, W., Hartz, S. M., Henson, R. A., & Templin, J. L. (2007). The fusion model skills diagnosis system. In J. P. Leighton & M. J. Gierl (Eds.), *Cognitive diagnostic assessment for education: Theory and applications* (pp. 275-318). Cambridge, England: Cambridge University Press.

Roussos, L. A., Hartz, S. M., & Stout, W. M. (2003, April). Real data applications of the Fusion Model skills diagnostic system. *Paper presented at the annual meeting of the National Council on Measurement in Education*, Chicago.

Swanson, L., & Stocking, M. L. (1993). A model and heuristic for solving very large item selection problems. *Applied Psychological Measurement*, *17*, 151-166.

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

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

Weiss, D. (1982). Improving measurement quality and efficiency with adaptive testing. *Applied Psychological Measurement*, *6*, 473-492.

Xu, X., Chang, H., & Douglas, J. (2003, April). *Computerized adaptive testing strategies for cognitive diagnosis.* Paper presented at the annual meeting of National Council on Measurement in Education*,* Montreal, Canada.