Variable-Length Computerized Adaptive Testing Based on Cognitive Diagnosis Models

Chia-Ling Hsu, Wen-Chung Wang and Shu-Ying Chen

Applied Psychological Measurement published online 28 May 2013
DOI: 10.1177/0146621613488642

Article

Variable-Length Computerized Adaptive Testing Based on Cognitive Diagnosis Models

Applied Psychological Measurement
XX(X) 1–20

DOI: 10.1177/0146621613488642
apm.sagepub.com
SAGE

Chia-Ling Hsu[1], Wen-Chung Wang[2], and Shu-Ying Chen[1]

Abstract

Interest in developing computerized adaptive testing (CAT) under cognitive diagnosis models (CDMs) has increased recently. CAT algorithms that use a fixed-length termination rule frequently lead to different degrees of measurement precision for different examinees. Fixed precision, in which the examinees receive the same degree of measurement precision, is a major advantage of CAT over nonadaptive testing. In addition to the precision issue, test security is another important issue in practical CAT programs. In this study, the authors implemented two termination criteria for the fixed-precision rule and evaluated their performance under two popular CDMs using simulations. The results showed that using the two criteria with the posterior-weighted Kullback–Leibler information procedure for selecting items could achieve the prespecified measurement precision. A control procedure was developed to control item exposure and test overlap simultaneously among examinees. The simulation results indicated that in contrast to no method of controlling exposure, the control procedure developed in this study could maintain item exposure and test overlap at the prespecified level at the expense of only a few more items.

Keywords

computerized adaptive testing, cognitive diagnosis, fixed precision, DINA model, fusion model, test security

Computerized adaptive testing (CAT) is more efficient than nonadaptive testing in achieving the same degree of measurement precision in all examinees using fewer items. Moreover, previous studies have shown that CAT yields a higher degree of measurement precision with the same number of items as nonadaptive testing uses (van der Linden & Glas, 2000; Wainer, 2000). In CAT, the items are selected adaptively based on the provisional latent trait estimate according to the examinee's responses to administered items. After an item is administered, the latent trait estimate is updated, and the next item is selected adaptively. Most existing CAT

[1]National Chung Cheng University, Chia-Yi, Taiwan
[2]The Hong Kong Institute of Education, Tai Po, New Territories, Hong Kong

Corresponding Author:
Wen-Chung Wang, The Hong Kong Institute of Education, 10 Lo Ping Rd., Tai Po, New Territories, NA, Hong Kong.
Email: wcwang@ied.edu.hk

algorithms are based on item response theory (IRT) models, such as the one-, two-, and three-parameter logistic models (Birnbaum, 1968; Rasch, 1960/1980), which yield a summative ability estimate for each examinee. Occasionally, this summative score might not provide sufficient feedback information for examinees or their teachers to adjust their learning or teaching strategies accordingly. To meet this demand for more detailed feedback, cognitive diagnosis models (CDMs) were developed to provide a profile for each examinee, which specifies whether each element of the required tasks, skills, or attributes has been mastered (Junker & Sijtsma, 2001; Rupp, Templin, & Henson, 2010; K. K. Tatsuoka, 1983).

CAT algorithms on CDMs (denoted as CD-CAT) have been developed to increase the applicability of CDMs (Cheng, 2009; McGlohen & Chang, 2008; Xu, Chang, & Douglas, 2003). CD-CAT algorithms adopt the fixed-length rule for terminating CAT; that is, CAT stops when a prespecified fixed test length (e.g., 20 items) is reached. This fixed-length termination rule, although easy to implement, often yields different degrees of measurement precision in different examinees. In practice, it is often desirable that all examinees have the same degree of measurement precision, which is a major advantage of CAT over nonadaptive testing (Weiss & Kingsbury, 1984). In this study, the authors implement the fixed-precision termination rule in CD-CAT. The terms *fixed precision* and *variable length* are used interchangeably here because variable test lengths are required for different examinees to achieve a common fixed precision.

Because CAT has been widely implemented in real tests, operational issues should also be considered. Among them, test security is essential, especially in high-stakes CAT. If the items are compromised, an examinee may obtain an unfair test score because of his or her preknowledge of the items. Tests are not secure if items are exposed or there is overlap between the items of two examinees. Many control methods have been proposed to increase test security, which also can be implemented in CD-CAT. However, most control methods are based on fixed-length CAT. In this study, the authors developed an exposure-control method based on variable-length CD-CAT.

Two CDMs, the deterministic-input, noisy-and-gate (DINA) model (Junker & Sijtsma, 2001) and the fusion model (Hartz, 2002; Rupp et al., 2010) are briefly introduced. The posterior-weighted Kullback–Leibler (PWKL) information method (Cheng, 2009) that is used for selecting items in CD-CAT is described. Two criteria based on the fixed-precision termination rule are proposed. A control procedure that can simultaneously control item exposure and test overlap between examinees is outlined. The results of simulation studies conducted to evaluate the two termination criteria and the control procedures of the DINA and fusion models are summarized. Finally, several conclusions are drawn, and suggestions for future studies are provided.

CDM

The aim of CDMs is to detect the attributes (skills or tasks) that an examinee has mastered. In contrast to IRT models, which provide a summative score for a broadly defined latent trait, CDMs use a multidimensional latent binary vector to specify whether an examinee has mastered each element of a set of specific attributes. Many CDMs have been proposed (Rupp et al., 2010). In general, they can be divided into two types: compensatory models and noncompensatory models. The difference between these two types depends on whether an insufficiency in one attribute can be compensated by another attribute. In noncompensatory models, all the required attributes are needed to solve an item. Commonly used noncompensatory models include the DINA model, the noisy input, deterministic-and-gate (NIDA) model (Junker & Sijtsma, 2001), and the noncompensatory reparameterized unified model, which is also called the fusion model. Frequently used compensatory models include the deterministic-input, noisy-or-gate (DINO) model (Templin & Henson, 2006) and the noisy input, deterministic-or-gate

(NIDO) model (Maris, 1999). Recently, generalized CDMs, such as the generalized DINA model (de la Torre, 2011) and the general diagnostic model (von Davier, 2005), have been proposed to account for complexity in real data.

The key components of CDMs are a vector of latent binary variables relevant to examinees and a **Q** matrix of item-to-attribute mapping. The **Q** matrix is a matrix, representing K attributes specifically required by J items. The entry q_{jk} in the **Q** matrix is defined by

$$q_{jk} = \begin{cases} 1 \text{ if the correct response to item } j \text{ requires attribute } k \\ 0 \text{ otherwise} \end{cases}$$

For example, assume a 3 by 5 **Q** matrix as follows:

$$\mathbf{Q} = \begin{bmatrix} 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 \end{bmatrix}.$$

This **Q** matrix indicates that the first item requires attributes 2 and 4; the second item requires attributes 1, 3, and 5; and the third item requires attribute 2. The **Q** matrix is often specified by content experts, instead of inferred from the data. Let $\boldsymbol{\alpha}_i = (\alpha_{i1}, \alpha_{i2}, \ldots, \alpha_{iK})'$ be a vector of a latent binary variable (0 or 1) that represents examinee i's mastery status, where $\alpha_{ik} = 1$ indicates examinee i has mastered attribute k, and $\alpha_{ik} = 0$ indicates that examinee i has not mastered attribute k. $\boldsymbol{\alpha}_i$ is referred to as the latent class (profile of latent attributes) for examinee i. Let $\mathbf{X_i} = (X_{i1}, X_{i2}, \ldots, X_{iJ})'$ be a response vector of 0 (incorrect) and 1 (correct) to items 1 to J for examinee i.

The DINA Model

Let a latent variable $\boldsymbol{\eta}_{ij}$ denote whether examinee i possesses all the attributes required for item j:

$$\boldsymbol{\eta}_{ij} = \prod_{k=1}^{K} \alpha_{ik}^{q_{jk}}, \tag{1}$$

where $\boldsymbol{\eta}_{ij} = 1$ only when examinee i has mastered all the required attributes for item j; otherwise, $\boldsymbol{\eta}_{ij} = 0$. There are two kinds of parameters in the DINA model: the slipping parameter (s_j), which indicates that the examinee has mastered all the required attributes but answered item j incorrectly, and the guessing parameter (g_j), which indicates that the examinee has not mastered all the required attributes but has answered item j correctly. The probability of a correct response to item j for examinee i is

$$P\left(X_{ij} = \mathbf{1} | \boldsymbol{\alpha}_i\right) = \left(1 - s_j\right)^{\eta_{ij}} g_j^{1-\eta_{ij}}. \tag{2}$$

Assuming local independence among items and examinees, the joint likelihood function of the DINA model is

$$L(\mathbf{s}, \mathbf{g}, \boldsymbol{\alpha} | \mathbf{X}) = \prod_{i=1}^{N} \prod_{j=1}^{J} \left[\left(1 - s_j\right)^{X_{ij}} s_j^{1-X_{ij}}\right]^{\eta_{ij}} \left[g_j^{X_{ij}} \left(1 - g_j\right)^{1-X_{ij}}\right]^{1-\eta_{ij}}, \tag{3}$$

where **s** and **g** are vectors that consist of all slipping parameters and all guessing parameters in the test, respectively; N is the sample size; and the others are defined as before.

The Fusion Model

The fusion model includes two types of parameters: (a) the baseline parameter π_j represents the probability of answering item j correctly, given that all the required attributes for item j are mastered, and (b) the penalty parameter r_{jk} represents the penalty for missing attribute k for item j. The probability of answering item j correction for examinee i in the fusion model is

$$P(X_{ij} = \mathbf{1}|\boldsymbol{\alpha}_i) = \left[\pi_j \prod_{k=1}^{K} r_{jk}^{(1-\alpha_{ik})q_{jk}}\right] P_{b_j}(\theta_i), \tag{4}$$

where $P_{b_j}(\theta_i)$ follows the Rasch model with item difficulty b_j, and θ_i is the latent trait for examinee i to account for the attributes that are not specified in the **Q** matrix. In the literature on fusion models (Henson & Douglas, 2005; McGlohen & Chang, 2008; Wang, Chang, & Huebner, 2011), $P_{b_j}(\theta_i)$ is often set at 1, which suggests that the specification of the **Q** matrix is complete. This practice is adopted in this study.

The DINA and fusion models are frequently used in CD-CAT (Cheng, 2009; Henson & Douglas, 2005; McGlohen & Chang, 2008; Wang et al., 2011; Xu et al., 2003). In practice and in simulation studies, investigators often prefer the DINA model mainly because of its simplicity in estimating, computing, and interpreting. The fusion model is sometimes used because it is more general than the DINA model is and often has a better fit to real data than the DINA model has (Wang et al., 2011). These two models were chosen to illustrate the utility of the proposed criteria for CAT termination and the control procedure for test security.

CAT Under CDM

Similar to other types of CAT, CD-CAT requires a model (e.g., the DINA and fusion models used in this study), a method for selecting the first item to administer, a scoring method for estimating the person measure, an item selection rule for the next item to administer, and a termination rule to stop the CAT. In addition to these major components, real CAT programs should consider operational issues, such as test security and content balancing. In this study, the authors focused on item selection, termination rule, and test security.

Item Selection

Most previous CD-CAT studies focused on developing algorithms for efficient item selection. Because of the nature of the latent class in CDMs, the commonly used Fisher information method is not applicable in CD-CAT. Instead, the Kullback -Leibler (KL) information method and the Shannon entropy method were developed and compared. Xu et al. (2003) showed that the entropy procedure outperforms the KL method. Cheng (2009) then developed the PWKL method, followed by the hybrid KL information method, which is a modification of the KL method. The simulation results showed that these two modified KL methods performed very similarly, and both outperformed the Shannon entropy method and the original KL method. Thus, the PWKL procedure was used in this study.

The KL information method is a measure of the discrepancy, or ''distance,'' between two probability distributions:

$$\mathrm{KL}(f||g) = E_f\left[\log\frac{f(x)}{g(x)}\right], \tag{5}$$

where $f(x)$ and $g(x)$ are two probability distributions. A critical feature of KL information is its asymmetry because $\mathrm{KL}(f \| g) \neq KL(g \| f)$. The larger $\mathrm{KL}(f \| g)$ is, the easier it is to distinguish the two distributions (Henson & Douglas, 2005). A cognitive diagnosis focuses on the conditional distribution of an examinee's responses given a latent class $\boldsymbol{\alpha}$. As in traditional CAT, the unknown true latent class in CD-CAT is substituted by the provisional estimate for the latent class. The KL information for item j is defined as

$$\mathrm{KL}_j(\hat{\boldsymbol{\alpha}} || \boldsymbol{\alpha}_c) = \sum_{x=0}^{1} \left[P\left(X_j = x | \hat{\boldsymbol{\alpha}}\right) \log \left(\frac{P\left(X_j = x | \hat{\boldsymbol{\alpha}}\right)}{P\left(X_j = x | \boldsymbol{\alpha}_c\right)} \right) \right], \tag{6}$$

where $\hat{\boldsymbol{\alpha}}$ is the provisional estimate for the latent class, and $\boldsymbol{\alpha}_c$ is a neighboring class of $\hat{\boldsymbol{\alpha}}$. Global KL information, which is based on the complete latent classes, is the sum of the KL information $P(\hat{\boldsymbol{\alpha}})$ and all $P(\boldsymbol{\alpha}_c)$s (there are 2^K possible latent classes, where K is the number of attributes). Equation 6 can be written as

$$\mathrm{KL}_j(\hat{\boldsymbol{\alpha}}) = \sum_{c=1}^{2^K} \sum_{x=0}^{1} \left[P\left(X_j = x | \hat{\boldsymbol{\alpha}}\right) \log \left(\frac{P\left(X_j = x | \hat{\boldsymbol{\alpha}}\right)}{P\left(X_j = x | \boldsymbol{\alpha}_c\right)} \right) \right]. \tag{7}$$

The item with the maximum value for KL, given the latent class of $\hat{\boldsymbol{\alpha}}$, will be administered (Xu et al., 2003).

As CAT proceeds, some latent classes have a higher probability than others of being the true latent class. Thus, the posterior probability of each latent class (PPLS) should be considered when items are selected. Let P_{0c} be the prior distribution for latent class c, $c = 1, 2, \ldots, 2^K$, $\sum_{c=1}^{2^K} P_{0c} = 1$. In this study, a uniform prior is used. Let $\mathbf{X}_i^L$ be examinee i's item responses in stage L (where L items have been administered) and $f(\mathbf{X}_i^L | \alpha_c)$ be the likelihood function of examinee i's item responses given latent class c in stage L. The posterior distribution of latent class, given L items, is defined as

$$P_L(\boldsymbol{\alpha}_c) = \frac{P_{0c} \cdot f\left(X_i^L | \boldsymbol{\alpha}_c\right)}{\sum_{c=1}^{2^K} P_{0c} \cdot f\left(X_i^L | \boldsymbol{\alpha}_c\right)}, \tag{8}$$

where $P_L(\boldsymbol{\alpha}_c)$ is the posterior distribution of latent class c, given L item responses, and the others are defined as before. The PWKL information can be written as

$$\mathrm{PWKL}_j\left(\hat{\boldsymbol{\alpha}}^L\right) = \sum_{c=1}^{2^K} \left\{ \left[\sum_{x=0}^{1} P\left(X_j = x | \hat{\boldsymbol{\alpha}}^L\right) \log \left(\frac{P\left(X_j = x | \hat{\boldsymbol{\alpha}}^L\right)}{P\left(X_j = x | \boldsymbol{\alpha}_c\right)} \right) \right] P_L(\boldsymbol{\alpha}_c) \right\}, \tag{9}$$

in which the KL information (Equation 7) is now weighted by the corresponding posterior distribution of a latent class. An item that has the maximum PWKL value, given the latent class of $\hat{\boldsymbol{\alpha}}$, will be administered.

Termination Rules

There are two major termination rules in CAT. Whereas in the fixed-length rule, CAT stops when a prespecified test length has been reached, in the fixed-precision rule, CAT stops when a prespecified precision has been reached. There are other termination rules. For example, CAT stops when there are no more available items capable of providing prespecified minimum

information or when the predicted gain in measurement precision brought on by administering an additional item is below a prespecified value (Choi, Grady, & Dodd, 2011).

One of the most important advantages of CAT over nonadaptive testing is that all test takers can be measured with the same degree of precision. However, most CD-CAT studies have used the fixed-length termination rule. For example, Wang et al. (2011), Cheng (2009), and Xu et al. (2003) compared several item selection algorithms in CD-CAT based on the DINA or fusion models using the fixed-length termination rule. To the best of knowledge, the variable-length termination rule has never been implemented in CD-CAT. In a relevant study, C. Tatsuoka (2002) classified examinees into one of a set of partially ordered states. The test stopped when the posterior probability the examinee belonged to a given latent class exceeded .80. The rationale was that the more peaked the posterior distribution was, the more reliable the classification (Huebner, 2010). In other words, posterior probability can be treated as a measure of precision.

Inspired by C. Tatsuoka (2002), the authors implemented the variable-length termination rule using the following two termination criteria:

Criterion 1: CAT stops when the largest PPLS is not smaller than a prespecified value (e.g., .70).
Criterion 2: CAT stops when not only the largest PPLS is not smaller than a prespecified value (e.g., .70) but also the second largest PPLS is not greater than a prespecified value (e.g., .10).

The second largest PPLS is considered in Criterion 2 to avoid two competing latent classes. For example, assume the prespecified value for the largest PPLS is set at .60. Using Criterion 1, a CAT stops as long as the largest PPLS reaches .60. At that time, the second largest PPLS might still be very high (e.g., near .20). If so, there would be two competing latent classes. One way to resolve this problem is to increase the prespecified value for the largest PPLS from .60 to .90, for example. Thus, a CAT would stop only when one is very confident of the outstanding latent class. Occasionally, this level of confidence might not be necessary. Another way to resolve this problem is to set a criterion for the second largest PPLS. For example, only when the largest PPLS is not smaller than .60 and the second largest PPLS is not greater than .20 can CAT stop. Although in theory another criterion can be set on the third or other PPLS, in the authors' experience, considering the first and second largest PPLS suffices.

In this study, the authors use the PPLS rather than the posterior probability of the latent attribute to terminate a CAT. In the CDM literature, classifying an examinee into one of all possible latent classes is of greater interest than is specifying individual latent traits as belonging to one of the two statuses (mastery or nonmastery). In fact, with the PPLSs, one can easily obtain the posterior probability of the latent attribute by marginalizing the PPLSs. Although CAT can stop when the posterior probability of the latent attribute reaches a prespecified value, no gain will be generated, which is in contrast to using the PPLS to terminate CAT. That is, requiring the PPLS to reach a prespecified level guarantees that the posterior probabilities of individual latent attributes will reach the same level. In contrast, requiring the posterior probability of the latent attribute to reach a prespecified level cannot guarantee that the posterior probability of the corresponding latent class will reach that level.

Setting an appropriate level for the two criteria is critical. For the largest PPLS, the prespecified value depends on the intended use of the test results. In high-stakes tests, the value should be set higher (e.g., .90 or .95); for low-stakes tests, the value can be lower (e.g., .70). Setting an appropriate level for the second largest PPLS is slightly complicated. Assume that the prespecified value for the largest PPLS (p_{1st}) has already been set at .60, and the number of latent attributes (K) is two. There will then be a total of four latent classes ($2^K = 4$). In CAT, when the

largest PPLS of an examinee has reached .60, the remaining three latent classes share the remaining probability of .40. In the best scenario, the three latent classes share .40 equally, which means that they are equally probable. In the worst scenario, one of the three latent classes takes .40 entirely, which means that this latent class can compete with the latent class that has a PPLS of .60. Thus, the seemingly lower bound of the prespecified value of the second largest PPLS (p_{2nd}) is $(1 - p_{1st})/(2^K - 1)$, and the upper bound is $1 - p_{1st}$ In this example, the lower and upper bounds are .13 and .40, respectively. The following formula can be used to set the level of p_{2nd}:

$$p_{2nd} = \frac{1 - p_{1st}}{2^K - 1} + \frac{(2^K - 2)\times(1 - p_{1st})\times d}{2^K - 1}, 0 \leq d \leq 1, \quad (10)$$

If $d = 0$, the lower bound is used; if $d = 1$, the upper bound is used. A small d prevents two competing latent classes.

The lower bound is derived under the assumption that the largest PPLS of an examinee is exactly equal to p_{1st}, and the remaining latent classes share the remaining probability equally. In reality, this assumption is seldom realized. When an examinee's largest PPLS is not smaller than p_{1st}, the actual value can be much larger than p_{1st}, such that the remaining probability for the other latent classes can be much smaller than $1 - p_{1st}$. For example, when an examinee has the largest PPLS, .67, the remaining posterior probability for the three latent classes is only .33, not .40. Thus, the lower bound for this particular examinee will be .11 instead of .13. Setting p_{2nd} at its lower bound (.13 in this example) may be attainable. However, as noted by Cheng (2009), when a latent class has the highest probability, the other latent classes will not share the remaining probability equally because their probabilities are related to their distances from the latent class with the largest probability. For example, when the latent class of ''nonmaster, nonmaster'' has the largest PPLS, the latent class of ''master, master'' will have a smaller probability than that of the latent classes of ''master, nonmaster'' or ''nonmaster, master.'' Although the assumption is not realized in reality, Equation 10 is a rule of thumb. The smaller the value of d, the higher the classification accuracy would be, at the cost of a longer test length.

Test Security

Many operational issues need to be considered in practical CAT programs. Test security deserves special attention. Particularly in high-stakes tests, item exposure and test overlap among examinees are two major concerns of test security. Many control methods have been developed to control either one or both by applying a set of exposure-control parameters (Chen & Lei, 2005; Chen, Lei, & Liao, 2008; Stocking & Lewis, 1995; Sympson & Hetter, 1985; van der Linden & Veldkamp, 2004). Intensive simulations are often needed to obtain stable exposure-control parameters before real CAT are held (Chen & Lei, 2005; Stocking & Lewis, 1995; Sympson & Hetter, 1985). Moreover, the tedious and iterative simulations must be conducted again whenever the CAT settings, such as the distribution of the examinee population, item pool, and item selection rules, change. Online exposure-control methods have been proposed to resolve these problems by updating the exposure-control parameters on the fly whenever an examinee finishes the CAT (Chen et al., 2008; van der Linden & Veldkamp, 2004).

In this study, the authors developed a control procedure called the Sympson–Hetter method, which comprises test overlap control, variable length, online update, and restricted maximum information (SHTVOR). Based on Sympson and Hetter's (1985) method, this procedure is capable of controlling test overlap for variable-length termination (Chen & Lei, 2005; Hsu & Chen,

2007), updating the exposure-control parameters online (Chen et al., 2008), and using restricted maximum information to freeze items with an exposure rate greater than the prespecified maximum until their exposure rate decreased (Revuelta & Ponsoda, 1998). SHTVOR consists of the following steps:

1. Specify the number of items in the item bank (e.g., $J = 300$); the number of examinees (e.g., $N = 2{,}000$); the maximum test length (e.g., $L_{\max} = 40$ items); the target maximum item exposure rate (e.g., $r_{\max} = .20$ indicates that every item has an exposure rate no greater than .20); the target test overlap rate ($\bar{T}_{\max}$), which is the target maximum value of the test overlap among examinees (e.g., $\bar{T}_{\max} = .20$ indicates that two examinees have an average of 20% of items overlapping); and the fixed-precision criterion (e.g., $p_{1\text{st}} = .90$). Let pk denote the exposure-control parameter of an item, which is set initially at 1 for all items.
2. Administer CAT to an examinee by selecting an item from the item pool and comparing the item's pk value with a random number drawn from $U(0, 1)$. If pk is larger, then administer the item; otherwise, select another item from the item pool and compare again to determine whether the item can be administered. Repeat this procedure until an item is administered. Remove the items selected for this examinee from the item pool.
3. After an item is administered, update the examinee's latent class estimate and select another item for administration according to the comparison of pk and random numbers. Repeat this step until the examinee has reached the prespecified fixed-precision criterion or until $L_{\max}$ is reached.
4. Compute $P(A)$ and $P(S)$ for item j ($j = 1, \ldots, J$) as the percentage an item has been administered and selected, respectively. The variance and mean of the item exposure rate across items are denoted as S^2 and $\bar{r}$ ($\bar{r} = \bar{L}/J$ for variable-length CAT, $\bar{L}$ is the mean test length across all examinees), respectively. The test overlap rate $\bar{T}$ is computed as (Hsu & Chen, 2007).

$$\bar{T} = \frac{N \times \sum_{j=1}^{J} P^2(A_j)}{\bar{L} \times (N-1)} - \frac{1}{(N-1)}, \tag{11}$$

where N is the total number of examinees who have undergone CAT thus far. The detailed rationale of Equation 11 is shown below. Update pk as follows (Chen et al., 2008; Revuelta & Ponsoda, 1998):

$$\begin{cases} pk = 0, & \text{if } P(\text{A}) \geq r_{\max}; \\ pk = \frac{r_{\max}}{P(S)}, & \text{if } P(\text{A}) \leq r_{\max} \text{and} P(S) > r_{\max}; \\ pk = 1, & \text{if } P(\text{A}) \leq r_{\max} \text{and} P(S) \leq r_{\max}. \end{cases} \tag{12}$$

5. If $\bar{T} > \bar{T}_{\max}$, then.
   (1). Calculate the target variance of the item exposure rate across items (S_0^2) based on Equation 11 while $\bar{T} = \bar{T}_{\max}$.
   (2). Set $P'(A)$ as $S_0[\frac{P(A)-\bar{r}}{S}] + \bar{r}$ for each item and set pk' as $P'(A)/P(S)$, where $P'(A)$ is the adjusted proportion of times an item has been administered based on S_0 and pk' is the adjusted exposure-control parameter according to $P'(A)$.
   (3). If $pk' > 1$, let $pk' = 1$. If $pk > pk'$, let $pk = pk'$.

6. To guarantee that all examinees will complete the CAT before exhausting the item pool, set the L_{max} largest *pks* as 1.
7. With the updated *pks*, repeat Steps 2 to 6 to administer the CAT again until all examinees have finished the CAT.

The computation of $\bar{T}$ in variable-length CAT deserves special attention. Assume that Examinees A and B have answered 10 and 20 items, respectively, and they share 5 common items. For Examinee A, 5 of the 10 items are shared so that the test overlap rate is 50% (= 5/10). For Examinee B, 5 of the 20 items are shared so that the test overlap rate is 25% (= 5/20). Thus, each pair of examinees has two overlap rates. For N examinees, there are $\binom{N}{2}$ pairs, and $2\binom{N}{2}$ overlap rates. $\bar{T}$ can be defined as the mean of the $2\binom{N}{2}$ overlap rates. In the case of two examinees, $\bar{T}$ is 37.5% (= (50% + 25%)/2). Alternatively, $\bar{T}$ can be defined as the mean number of shared items across examinees divided by the mean test length across examinees (Hsu & Chen, 2007). In this case, the mean number of shared items across examinees is 5, and the mean test length across examinees is 15 (= (10 + 20)/2), and $\bar{T}$ is 33.3% (= 5/15). In fixed-length CAT, these two definitions are mathematically equivalent.

The latter definition of $\bar{T}$ was adopted in this study. Specifically, let L_n denote the test length of examinee n (n = 1, . . . , N). There are $\binom{N}{2}$ pairs of examinees. For each pair of examinees, one can compute the number of shared items, denoted as O_m (m = 1, . . . , $\binom{N}{2}$). Let $\bar{O}$ denote the mean of the shared items across all pairs, and let $\bar{L}$ denote the mean test length across examinees:

$$\bar{O} = \sum_{m=1}^{\binom{N}{2}} \frac{O_m}{\binom{N}{2}}, \tag{13}$$

$$\bar{L} = \sum_{n=1}^{N} \frac{L_n}{N}. \tag{14}$$

The test overlap rate is

$$\bar{T} = \frac{\bar{O}}{\bar{L}} = \frac{\sum_{m=1}^{\binom{N}{2}} O_m / \binom{N}{2}}{\sum_{n=1}^{N} L_n / N}. \tag{15}$$

Because the test overlap rate is defined as the ratio of the mean shared items over the mean test length, different mean shared items and different mean test lengths may generate the same test overlap rate. Thus, the test overlap rate across conditions should be interpreted cautiously. When the mean test length is fixed, the greater the number of shared items is, the larger is the test overlap rate.

The computation of Equation 15 is very tedious and can be simplified as follows (Chen & Lei, 2005). Let h_j (j = 1, . . . , J) be the number of times that item j was administered. Then, $\sum_{m=1}^{\binom{N}{2}} O_m$ can be computed by $\sum_{j=1}^{J} \binom{h_j}{2}$. Thus,

$$\begin{aligned}\bar{T} &= \frac{\sum_{j=1}^{J}\binom{h_j}{2}}{\bar{L}\binom{N}{2}} = \frac{\sum_{j=1}^{J} h_j(h_j-1)}{\bar{L}N(N-1)} = \frac{\sum_{j=1}^{J} P(A_j)\left(NP(A_j)-1\right)}{\bar{L}(N-1)} && \because P(A_j) = \frac{h_j}{N} \\ &= \frac{N\sum_{j=1}^{J} P^2(A_j) - \sum_{j=1}^{J} P(A_j)}{\bar{L}(N-1)} = \frac{N\sum_{j=1}^{J} P^2(A_j) - \bar{L}}{\bar{L}(N-1)} && \because \sum_{j=1}^{J} P(A_j) = \frac{\sum_{j=1}^{J} h_j}{N} = \bar{L} \\ &= \frac{N\sum_{j=1}^{J} P^2(A_j)}{\bar{L}(N-1)} - \frac{1}{(N-1)}. \end{aligned} \quad (16)$$

Setting an appropriate level for $r_{\max}$ and $\bar{T}_{\max}$ is important. Setting $r_{\max}$ and $\bar{T}_{\max}$ at 1 indicates that there is no control of item exposure and test overlap, respectively. In practice, $r_{\max}$ is often set at .20 or .30, and $\bar{T}_{\max}$ can be set at a value slightly lower than $r_{\max}$ (Hsu & Chen, 2007). A large value of $\bar{T}_{\max}$ can lead to no additional control of test overlap other than setting $r_{\max}$ alone.

To acquire additional control of test overlap when $r_{\max}$ has already been specified, $\bar{T}_{\max}$ can be set as follows: Conduct a simulation with a prespecified value of $r_{\max}$ (e.g., .20) and $\bar{T}_{\max} = 1$ and obtain an empirical result of $\bar{T}$ (e.g., .18). In this case, the empirical result of $\bar{T}$ reflects the average test overlap among examinees when there is control of item exposure, but no control of test overlap. In this example, setting $r_{\max}$ at .20 alone leads to an average test overlap of .18. To acquire additional control of test overlap, $\bar{T}_{\max}$ must be set at a value smaller than .18 (e.g., $\bar{T}_{\max} = .15$); otherwise, nothing will change.

Demonstration 1: The DINA Model

Condition 1: Without Test Security Control

Design and Analysis. The performances of Criteria 1 and 2 were evaluated using the DINA model. No method for controlling test security was implemented. The item bank consisted of 300 items in six attributes, which resulted in 64 latent classes. Each examinee had a 50% chance of mastering each attribute, which was generated independently from a Bernoulli distribution with a probability of .5. That is, each of the 64 latent classes had a probability of 1/64. A total of 2,000 examinees were generated. The settings were similar to those in Cheng (2009). Although latent classes can be generated from a nonuniform distribution, as in Henson and Douglas (2005), the feasibility of Criteria 1 and 2 would not change. Each attribute was measured by 20% of the items, and each item measured at least one attribute. The item parameters g_j and s_j were generated from $U(0.05, 0.25)$. The PWKL method was used to select items. The maximum a posteriori estimate, with a prior substituted by 1/64, was used to update the provisional and final estimate of the latent class. After an item was administered, the PPLS was computed using Equation 8.

Criteria 1 and 2 were used to terminate the CAT. In Criterion 1, p_{1st} was set at .60, .70, .80, .90, and .95, respectively. In Criterion 2, in addition to p_{1st}, p_{2nd} was set as the value computed from Equation 10 with d set at 0, .25, .50, and .75, respectively. No constraint was imposed on the maximum test length. The CAT stopped when either Criterion 1 or 2 was satisfied or when all 300 items in the item bank were exhausted. When the CAT stopped, an examinee was assigned to the latent class with the largest PPLS.

Table 1. Classification Accuracy for Latent Class and Test Length Under the DINA Model Without Test Security Control.

			Test length			
p_{1st}	p_{2nd}	Accuracy of latent class	*M*	*SD*	Maximum	Minimum
.60	—	.67	6.7	1.2	13	4
.60	.292 (*d* = .75)	.67	6.7	1.2	13	4
.60	.197 (*d* = .50)	.67	6.7	1.2	14	4
.60	.102 (*d* = .25)	.67	6.7	1.2	15	4
.60	.006 (*d* = 0)	.99	15.6	3.9	40	7
.70	—	.77	8.1	1.9	19	4
.70	.219 (*d* = .75)	.77	8.1	1.9	19	4
.70	.148 (*d* = .50)	.77	8.1	1.9	19	4
.70	.076 (*d* = .25)	.78	8.4	2.0	21	4
.70	.005 (*d* = 0)	.99	16.5	4.0	41	7
.80	—	.84	9.6	2.4	23	5
.80	.146 (*d* = .75)	.84	9.6	2.4	23	5
.80	.098 (*d* = .50)	.84	9.6	2.4	23	5
.80	.051 (*d* = .25)	.94	12.2	2.7	27	5
.80	.003 (*d* = 0)	.00	18.7	4.3	43	8
.90	—	.93	11.8	2.9	26	5
.90	.073 (*d* = .75)	.93	11.8	2.9	26	5
.90	.049 (*d* = .50)	.95	12.4	2.8	28	5
.90	.025 (*d* = .25)	.97	13.0	3.0	28	6
.90	.002 (*d* = 0)	.00	18.9	4.5	45	8
.95	—	.97	13.1	3.1	28	6
.95	.037 (*d* = .75)	.97	13.1	3.1	28	6
.95	.025 (*d* = .50)	.97	13.1	3.1	28	6
.95	.013 (*d* = .25)	.97	13.3	3.4	34	6
.95	.001 (*d* = 0)	.01	19.5	4.7	46	8

Note: DINA = deterministic-input, noisy-and-gate; p_{1st} and p_{2nd} = the first and second largest posterior probabilities of latent class, respectively; — = not implemented; *d* = the lower bound of the prespecified value for p_{2nd}.

The major dependent variables were (a) the classification accuracy of the latent class, which was defined as the percentage of examinees whose latent classes were classified correctly, and (b) the test length required to finish the CAT. The authors expected that Criteria 1 and 2 would achieve their goals successfully.

Results. Table 1 shows the classification accuracy of the latent class and the summary statistics (mean, standard deviation [SD], maximum, and minimum) of the test length required to finish the CAT across examinees. The accuracy of all classifications of the latent class was larger than p_{1st}. To check validity (detailed results not shown), the minimum value of the largest PPLS across examinees was not smaller than p_{1st}, and the maximum value of the second largest PPLS across examinees was not greater than p_{2nd}. In other words, Criteria 1 and 2 achieved their goals successfully. With Criterion 1, the second largest PPLS across examinees could be as large as .23 when p_{1st} = .60; .20 when p_{1st} = .70; .12 when p_{1st} = .80; .08 when p_{1st} = .90; .04 when p_{1st} = .95. Such large values justified the use of Criterion 2 when p_{1st} was set at .70 or lower.

With Criterion 2, when p_{2nd} was set at its lower bound (d = 0), the accuracy of the latent class was increased to as high as .99 or 1.00, at the cost of a substantially longer test. When d was set at .25, Criterion 2 required a slightly longer test (fewer than three items on average) than Criterion 1 did. When d was set at .50 or higher, there were almost no differences in

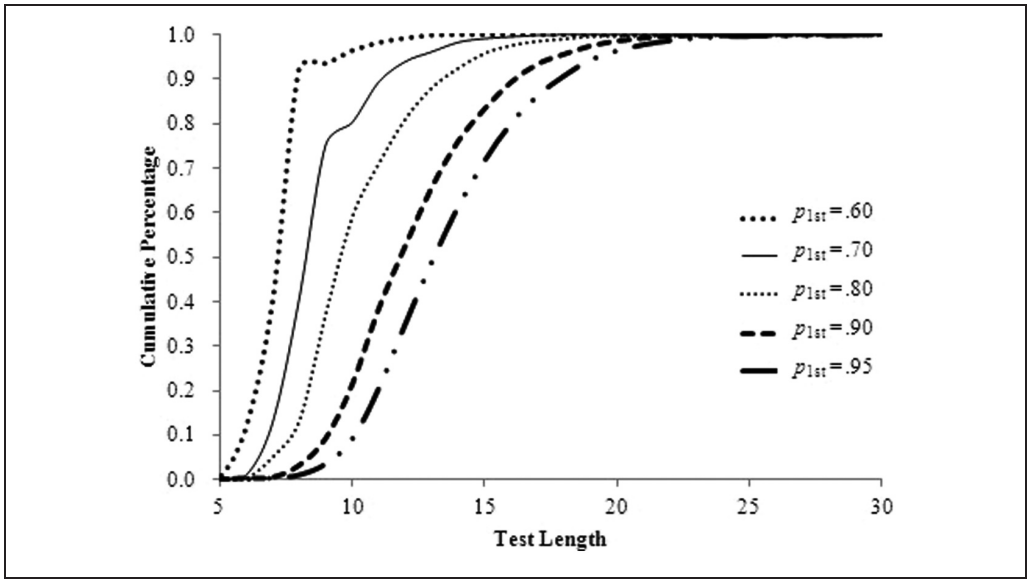

Figure 1. Cumulative percentages of examinees finishing the CAT in different test lengths under the DINA model with Criterion 1.

accuracy and test length between Criteria 1 and 2. In other words, compared with Criterion 1, Criterion 2 controlled the second largest PPLS and achieved a similar degree of accuracy with little cost to test length when d was .25 or above. In light of the simulation results and their feasibility, the following two suggestions seem useful:

1. For high-stakes tests, p_{1st} can be set as high as .90 or .95. If so, Criterion 1 suffices, and Criterion 2 is not necessary.
2. For low-stakes tests, if p_{1st} is set at .70 or lower, Criterion 2 can be adopted with d between .25 and .50, or simply p_{2nd} = .10.

Figure 1 shows the cumulative percentage of examinees finishing the CAT in different test lengths under different levels of p_{1st} when Criterion 1 was adopted. The larger the value of p_{1st}, the longer the test required to finish the CAT. For example, approximately 90% of examinees finished the CAT in eight items at p_{1st} = .60, in 11 items at p_{1st} = .70, in 13 items at p_{1st} = .80, in 16 items at p_{1st} = .90, and in 18 items at p_{1st} = .95. However, when the test maximum length was fixed at a certain level, the lower the value of p_{1st} was, the higher the number of examinees who finished the CAT. For example, if the maximum test length was set at 10 items, then approximately 96% of examinees finished the CAT at p_{1st} = .60, 80% at p_{1st} = .70, 59% at p_{1st} = .80, 21% at p_{1st} = .90, and 9% at p_{1st} = .95. When Criterion 2 was adopted, the general pattern was similar to that shown in Figure 1, except that finishing the CAT was much more difficult when d = 0.

Condition 2: With Test Security Control

Design and Analysis. The SHTVOR was implemented under the DINA model, together with PWKL for item selection, Criterion 1 for CAT termination, and p_{1st} = .90. There were three levels of r_{max} (1, .30, and .20) and four levels of $\bar{T}_{max}$ (1, .30, .20, and .15). An r_{max} of 1 indicated

no control of item exposure, and a $\bar{T}_{max}$ of 1 indicated no control of test overlap. In other words, when $r_{max} = \bar{T}_{max} = 1$, no control procedure for test security was implemented. Random item selection was also implemented for comparison. In real CAT programs, a maximum test length is often imposed; therefore, a maximum test length was also imposed here. According to Figure 1, approximately 99% of examinees finished the CAT in 21 items under the DINA model when $p_{1st} = .90$. Thus, the maximum test length was set at 21 items. That is, an examinee would be classified into a latent class if either Criterion 1 was met or the maximum test length was reached. The item and person parameter settings were identical to those in Condition 1.

The dependent variables included the following: (a) the maximum item exposure rate across items, denoted as max(r); (b) the test overlap rate $\bar{T}$; (c) the percentage of examinees who received the maximum test length; (d) the classification accuracy of the latent class; (e) the mean test length across examinees; and (f) bank usage, defined as the number of items in the 300-item pool that had been administered as least once. SHTVOR was expected to maintain control of item exposure and test overlap at the prespecified levels and increase bank usage at the expense of longer tests, compared with no exposure control (i.e., $r_{max} = \bar{T}_{max} = 1$). The random item selection was expected to yield the best test security and bank usage but the worst classification accuracy.

Results. The results are shown in Table 2. Several conclusions can be drawn. First, the empirical max(r) and empirical $\bar{T}$ were well controlled, which suggests that SHTVOR had good control of item exposure and test overlap, simultaneously. For example, when r_{max} was set at .20, the empirical max(r) was between .1985 and .2000; when $\bar{T}_{max}$ was set at .20, the empirical $\bar{T}$ was between .1651 and .2022. Second, setting $\bar{T}_{max}$ at a value higher than the empirical $\bar{T}$ obtained from the condition where $\bar{T}_{max} = 1$ did not bring any additional control of test overlap than setting r_{max} alone did. For example, when $r_{max} = .30$ and $\bar{T}_{max} = 1$, the empirical $\bar{T}$ was .2364. Thus, when r_{max} was set at .30, $\bar{T}_{max}$ must be set at a value smaller than .2364 to acquire additional control of test overlap. This was evident in comparing the results under the condition of $r_{max} = .30$ and $\bar{T}_{max} = .30$, and the results under the condition of $r_{max} = .30$ and $\bar{T}_{max} = 1$. In contrast, once $\bar{T}_{max}$ was set at a value smaller than .2364, additional control of the test overlap was required, as shown in the results under the condition of $r_{max} = .30$ and $\bar{T}_{max} = .20$, or $r_{max} = .30$ and $\bar{T}_{max} = .15$.

Third, a more stringent value for r_{max} or $\bar{T}_{max}$ led to better bank usage at the slight cost of a longer test and a higher percentage of examinees receiving the maximum test length. Compared with the mean test length of 11.8 items when SHTVOR was not implemented ($r_{max} = \bar{T}_{max} = 1$), implementing SHTVOR cost approximately only one additional item. Fourth, examinees finishing the CAT before reaching the maximum test length had a classification accuracy rate of .93, which was slightly higher than p_{1st} (.90), which indicated that PWKL with Criterion 1 was successful. Examinees receiving the maximum test length had a classification accuracy rate between .58 and .75. Fifth, the random item selection yielded the best test security (smallest max(r) and $\bar{T}$) and bank usage, but the worst classification efficiency. It required 20.6 items to achieve a classification accuracy rate of .91; whereas the other methods required only 12 or 13 items to achieve a rate of .93. In addition, the random item selection had the largest percentage of examinees receiving the maximum test length.

Demonstration 2: The Fusion Model

Condition 1: Without Test Security Control

Design and Analysis. The fusion model was adopted. The item parameters π_j and r_{jk} were generated from $U(0.75, 0.95)$ and $U(0.2, 0.95)$, respectively. The other settings, including the number

Table 2. Classification Accuracy of Latent Class and Other Summary Statistics Under the DINA Model With Test Security Control.

r_{max}	$\bar{T}_{max}$	max(r)	$\bar{T}$	% (max length)	Class accuracy (p_{1st})	Class accuracy (max length)	Mean test length	Bank usage
SHTVOR								
1.00	1.00	.9775	.5928	0.3	.93	.50	11.8	108
.30	1.00	.2995	.2364	2.0	.93	.60	12.4	134
.30	0.30	.2995	.2364	2.0	.93	.60	12.4	134
.30	0.20	.2950	.2022	2.1	.93	.68	12.7	146
.30	0.15	.2875	.1668	3.8	.93	.71	13.2	168
.20	1.00	.2000	.1651	3.4	.93	.75	12.9	152
.20	0.30	.2000	.1651	3.4	.93	.75	12.9	152
.20	0.20	.2000	.1651	3.4	.93	.75	12.9	152
.20	0.15	.1985	.1522	3.9	.93	.58	13.2	161
Random selection		.0860	.0688	88.9	.91	.53	20.6	300

Note: DINA = deterministic-input, noisy-and-gate; SHTVOR = Sympson–Hetter method, which comprises test overlap control, variable length, online update, and restricted maximum information; r_{max} = the prespecified maximum item exposure rate; $\bar{T}_{max}$ = the prespecified maximum test overlap rate; max(r) = the maximum item exposure rate in the item bank; $\bar{T}$ = the test overlap rate; % (max length) = the percentage of examinees reaching the maximum test length of 21 items under the DINA model and 47 items under the fusion model; class accuracy (p_{1st}) = the classification accuracy rate for examinees who finished the computerized adaptive testing (CAT) using Criterion 1; class accuracy (max length) = the classification accuracy rate for examinees reaching the maximum test length; mean test length = the mean test length required to terminate the CAT across examinees; bank usage = the number of items used in the 300-item pool.

Table 3. Classification Accuracy for Latent Class and Test Length Under the Fusion Model Without Test Security Control.

			Test length			
p_{1st}	p_{2nd}	Accuracy of latent class	*M*	*SD*	Maximum	Minimum
.60	—	.66	12.0	4.0	36	6
.60	.292 (*d* = .75)	.66	12.0	4.0	36	6
.60	.197 (*d* = .50)	.66	12.0	4.0	36	6
.60	.102 (*d* = .25)	.68	12.4	4.6	67	6
.60	.006 (*d* = 0)	.98	33.7	14.5	195	14
.70	—	.75	14.7	4.8	39	7
.70	.219 (*d* = .75)	.75	14.7	4.8	39	7
.70	.148 (*d* = .50)	.75	14.7	4.9	39	7
.70	.076 (*d* = .25)	.78	15.6	5.6	76	7
.70	.005 (*d* = 0)	.98	34.9	15.4	195	15
.80	—	.84	17.7	5.7	50	8
.80	.146 (*d* = .75)	.84	17.7	5.7	50	8
.80	.098 (*d* = .50)	.84	17.7	5.8	67	8
.80	.051 (*d* = .25)	.87	19.1	6.6	83	9
.80	.003 (*d* = 0)	.99	38.5	17.0	195	16
.90	—	.93	22.1	7.5	76	11
.90	.073 (*d* = .75)	.93	22.1	7.5	76	11
.90	.049 (*d* = .50)	.93	22.1	7.6	83	11
.90	.025 (*d* = .25)	.94	23.2	8.3	83	11
.90	.002 (*d* = 0)	.99	41.8	18.7	195	18
.95	—	.96	26.9	9.6	156	13
.95	.037 (*d* = .75)	.96	26.9	9.6	157	13
.95	.025 (*d* = .50)	.96	26.9	9.7	157	13
.95	.013 (*d* = .25)	.96	28.4	11.5	195	13
.95	.001 (*d* = 0)	.99	47.9	23.4	213	19

Note: p_{1st} and p_{2nd} = the first and second largest posterior probabilities of latent class, respectively; — = not implemented; *d* = the lower bound of the prespecified value for p_{2nd}.

of items, attributes and examinees, item selection rules, scoring methods, termination criteria, and dependent variables, were identical to those in Condition 1 of Demonstration 1. The item bank was different from that in Demonstration 1 because different models were used to construct the item banks in these two demonstrations.

Results. Table 3 shows the classification accuracy of the latent class and the summary statistics of the test length required to finish the CAT across examinees under the fusion model. The conclusions drawn from the DINA model apply to the fusion model. In every case, the classification accuracy of the latent class was larger than p_{1st}. Criteria 1 and 2 were successful because the minimum value of the largest PPLS across the examinees was never smaller than p_{1st}, and the maximum value of the second largest PPLS across examinees was never greater than p_{2nd}. With Criterion 1, the second largest PPLS across examinees could be as large as .18 when p_{1st} = .60; .17 when p_{1st} = .70; .13 when p_{1st} = .80; .07 when p_{1st} = .90; .05 when p_{1st} = .95. These large values justified the use of Criterion 2 when p_{1st} was set at .70 or lower.

With Criterion 2, setting p_{2nd} at its lower bound (d = 0) substantially increased the classification accuracy of the latent class and the test length. When d was set at .25, Criterion 2 cost approximately one more item than Criterion 1 did. When d was set at .50 or .75, there was almost no cost in test length. Given that the great similarity in the conclusions between the

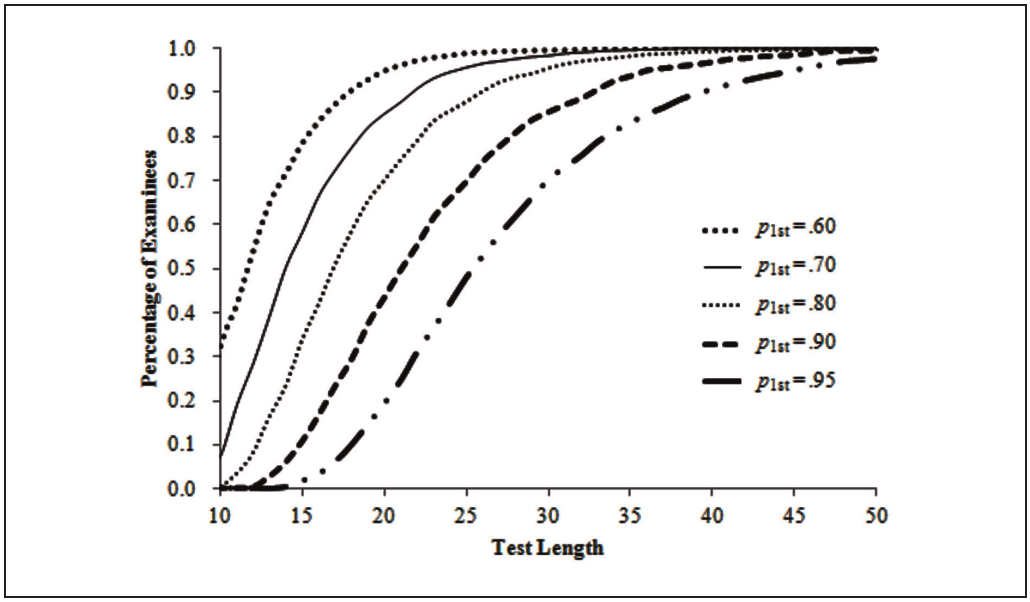

Figure 2. Cumulative percentages of examinees finishing the CAT in different test lengths under the fusion model with Criterion 1.

fusion model and the DINA model, the two suggestions for the DINA model are applicable to the fusion model: (a) For high-stakes tests, p_{1st} can be set as high as .90 or .95, and Criterion 2 is not necessary; (b) for low-stakes tests, if p_{1st} is set at .70 or lower, Criterion 2 can be adopted with d between .25 and .50, or simply $p_{2nd} = .10$.

Figure 2 shows the cumulative percentage of examinees finishing the CAT in different test lengths under different levels of p_{1st} when Criterion 1 was adopted. The same conclusions drawn from Figure 1 apply here. A smaller p_{1st} resulted in a shorter test length. Approximately 90% of examinees finished the CAT in 18 items at $p_{1st} = .60$, in 22 items at $p_{1st} = .70$, in 26 items at $p_{1st} = .80$, in 33 items at $p_{1st} = .90$, and 39 items at $p_{1st} = .95$. When the maximum test length was fixed, the lower the value of p_{1st}, the higher the number of examinees who finished the CAT. For example, in a maximum test length of 20 items, approximately 95% of examinees finished the CAT at $p_{1st} = .60$, 85% at $p_{1st} = .70$, 70% at $p_{1st} = .80$, 44% at $p_{1st} = .90$, and 19% at $p_{1st} = .95$. When Criterion 2 was adopted, the same conclusions that were drawn for Criterion 1 applied, except that the test was much longer at $d = 0$.

Condition 2: With Test Security Control

Design and Analysis. The SHTVOR procedure was implemented in the fusion model. The settings were identical to those in Condition 2 of Demonstration 1, except the maximum test length was set at 47 items.

Results. The results are shown in Table 4. Five conclusions were drawn. First, SHTVOR maintained good control of both item exposure and test overlap at their prespecified levels. Second, $\bar{T}_{max}$ should be set at a value smaller than the empirical $\bar{T}$ under the condition of $\bar{T}_{max} = 1$ to acquire more control of test overlap than setting r_{max} alone would. Third, compared with the mean test length of 22.0 items without SHTVOR, the implementation of SHTVOR cost approximately 4 to 14 more items. Fourth, PWKL together with Criterion 1 was successful because

Table 4. Classification Accuracy of Latent Class and Other Summary Statistics Under the Fusion Model With Test Security Control.

r_{max}	$\bar{T}_{max}$	max(r)	$\bar{T}$	% (max length)	Class accuracy (p_{1st})	Class accuracy (max length)	Mean test length	Bank usage
SHTVOR								
1.00	1.00	.9815	.4503	0.6	.92	.42	22.0	195
.30	1.00	.2995	.2389	3.3	.90	.69	25.8	223
.30	0.30	.2995	.2389	3.3	.90	.69	25.8	223
.30	0.20	.2995	.2115	6.1	.90	.70	27.4	238
.30	0.15	.2740	.1748	19.6	.90	.69	32.9	263
.20	1.00	.2000	.1644	16.6	.91	.69	31.9	272
.20	0.30	.2000	.1644	16.6	.91	.69	31.9	272
.20	0.20	.2000	.1644	16.6	.91	.69	31.9	272
.20	0.15	.1995	.1533	34.8	.91	.68	35.8	282
Random selection		.1755	.1555	93.7	.90	.57	46.7	300

Note: SHTVOR = Sympson–Hetter method, which comprises test overlap control, variable length, online update, and restricted maximum information; r_{max} = the prespecified maximum item exposure rate; $\bar{T}_{max}$ = the prespecified maximum test overlap rate; max(r) = the maximum item exposure rate in the item bank; $\bar{T}$ = the test overlap rate; % (max length) = the percentage of examinees reaching the maximum test length of 21 items under the DINA model and 47 items under the fusion model; class accuracy (p_{1st}) = the classification accuracy rate for examinees who finished the computerized adaptive testing (CAT) using Criterion 1; class accuracy (max length) = the classification accuracy rate for examinees reaching the maximum test length; mean test length = the mean test length required to terminate the CAT across examinees; bank usage = the number of items used in the 300-item pool.

examinees finishing the CAT before the maximum test length had a classification accuracy rate between .90 and .91. Fifth, the random item selection yielded the best test security and bank usage, but the worst classification efficiency and the largest percentage of examinees receiving the maximum test length.

Discussion and Conclusion

In CAT practice, it is often desirable that all examinees are measured with similar precision. In this study, the authors developed two criteria for terminating CD-CAT based on the fixed-precision rule. Criterion 1 requires that the largest PPLS not be smaller than a prespecified value (p_{1st}), whereas Criterion 2 has the additional requirement that the second largest PPLS is not greater than a prespecified value (p_{2nd}). These two criteria were demonstrated under the DINA and fusion models. The lower and upper bounds of p_{2nd} were identified for the setting of an appropriate level of p_{2nd}. The simulation results indicated that both criteria achieved their goals successfully: the more stringent the prespecified level of p_{1st} and p_{2nd}, the longer the test that was required; and a large value of d (e.g., .50 or .75) made Criterion 2 equivalent to Criterion 1. The following recommendations are based on these results: for high-stakes tests, p_{1st} should be set at .90 or higher; for low-stakes tests, p_{1st} can be set at .80 or lower and p_{2nd} as .10.

Test security is essential in high-stakes tests. High item exposure and test overlap threaten test security. SHTVOR was developed and implemented in CD-CAT to control item exposure and test overlap simultaneously. Using simulations, the performance of SHTVOR was evaluated under the DINA and fusion models. The results indicated that SHTVOR maintained control of item exposure and test overlap at the prespecified levels: the more stringent the levels, the longer the test that was required to terminate the CAT, and the greater number of examinees that received the maximum test length.

Several issues require further investigation. Similar to other simulation studies, the simulation settings in this study are not comprehensive. Evaluating the performance of the two termination criteria and SHTVOR under other conditions, such as different item selection methods, ability estimation methods, item banks, Q matrices, and number of attributes, would be of great value. Unlike in the fixed-length CAT, the test overlap rate in variable-length CAT is not uniquely defined. More studies are needed to derive a unique definition or to compare different definitions. This study used the DINA and fusion models as examples. However, in theory, the two termination criteria and SHTVOR are independent of models and could be implemented on other CDMs, which could be verified by future studies. Although SHTVOR was implemented in the CD-CAT scenario, this method can be easily implemented in IRT-based CAT programs.

As with Fisher information, using PWKL to select items often leads to uneven item usage. For example, as shown in Table 2, when no control method for test security was implemented, the empirical max(r) was as high as .98, and the bank usage was uneven. This study thus implemented SHTVOR to control item exposure and test overlap directly at a prespecified level. In effect, restrictive item selection methods, in which randomness or item exposure rate plays a role in item selection, can be helpful in increasing bank usage and reducing item exposure, at little cost to measurement precision (Barrada, Olea, Ponsoda, & Abad, 2008; Revuelta & Ponsoda, 1998; Wang et al., 2011). The establishment and effectiveness in CD-CAT of using these restrictive item selection methods with a fixed-precision termination rule need further investigation. Finally, whereas content balancing is a significant issue in operational CAT, attribute balancing is important in CD-CAT (Cheng, 2010). An interesting topic for future research would be the implementation of attribute balancing with a fixed-precision termination rule.

Declaration of Conflicting Interests

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

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The second author was sponsored by the Research Grants Council, Hong Kong (Grant HKIEd 8012-PPR-10).

References

Barrada, J. R., Olea, J., Ponsoda, V., & Abad, F. J. (2008). Incorporating randomness in the Fisher information for improving item exposure control in CATs. *British Journal of Mathematical and Statistical Psychology*, *61*, 493-513.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397-479). Reading, MA: Addison-Wesley.

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

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.

Choi, S.-W., Grady, M. W., & Dodd, B. G. (2011). A new stopping rule for computerized adaptive testing. *Educational and Psychological Measurement*, *71*, 37-53.

de la, & Torre, J. (2011). The generalized DINA model framework. *Psychometrika*, *76*, 179-199.

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., & Douglas, J. (2005). Test construction for cognitive diagnosis. *Applied Psychological Measurement*, *29*, 262-277.

Hsu, C.-L., & Chen, S.-Y. (2007). Controlling item exposure and test overlap in variable length computerized adaptive testing (in Chinese). *Psychological Testing*, *54*, 403-428.

Huebner, A. (2010). An overview of recent developments in cognitive diagnostic computer adaptive assessments. *Practical Assessment Research & Evaluation*, *15*(3). Retrieved from http://pareonline.net/getvn.asp?v=15&n=3

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.

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

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

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests*. Chicago, IL: The University of Chicago Press. (Original work published 1960)

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

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

Stocking, M. L., & Lewis, C. (1995). *A new method of controlling item exposure in computerized adaptive testing* (Research Report 95–25). Princeton, NJ: Educational Testing Service.

Sympson, J. B., & Hetter, R. D. (1985). Controlling item-exposure rates in computerized adaptive testing. In *Proceedings of the 27th annual meeting of the Military Testing Association* (pp. 973-977). San Diego, CA: Navy Personnel Research and Development Centre.

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

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

van der Linden, W. J., & Glas, C. A. W. (Eds.). (2000). *Computerized adaptive testing: Theory and practice*. Boston, MA: Kluwer.

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

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

Wainer, H. (Ed.). (2000). *Computerized adaptive testing: A primer* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum.

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

Weiss, D. J., & Kingsbury, G. G. (1984). Application of computerized adaptive testing to educational problems. *Journal of Educational Measurement*, *21*, 361-375.

Xu, X., Chang, H., & Douglas, J. (2003, April). *A simulation study to compare CAT strategies for cognitive diagnosis*. Paper presented at the Annual Meeting of the American Education Research Association, Chicago, IL.