Increasing the Homogeneity of CAT's Item-Exposure Rates by Minimizing or Maximizing Varied Target Functions While Assembling Shadow Tests

by

Yuan H. Li

Prince Georges County Public Schools, Maryland

William D. Schafer

University of Maryland at College Park

Address:
Yuan H. Li
Prince George's County Public Schools
Test Admin. Department, Room, 202E
Upper Marlboro, MD 20772
e-mail:yuanhwangli@juno.com

Paper presented at the annual meeting of the American Educational Research Association April, 21-25, 2003, Chicago, IL.

Increasing the Homogeneity of CAT's Item-Exposure Rates by Minimizing or Maximizing Varied Target Functions While Assembling Shadow Tests

Abstract: A computerized adaptive testing (CAT) algorithm that has the potential to increase the homogeneity of CAT's item-exposure rates without significantly sacrificing the precision of ability estimates was proposed and assessed in the shadow-test (van der Linden & Reese, 1998) CAT context of . This CAT algorithm was formed by a combination of maximizing or minimizing varied target functions while assembling shadow tests. There were four target functions to be separately used in the first, second, third and fourth quarter test of CAT. The elements to be used in the four functions were associated with: (a) a random number assigned to each item, (b) the absolute difference between an examinee's current ability estimate and an item difficulty, (c) the absolute difference between an examinee's current ability estimate and an optimum item difficulty, and (d) item information.

The results indicated that this combined CAT fully utilized all the items in the pool, reduced the maximum exposure rates, and achieved more homogeneous exposure rates. Moreover, its precision in recovering ability estimates was similar to that of the maximum item information method. The combined CAT method resulted in the best overall results compared with the other individual CAT item-selection methods. The findings from the combined CAT are encouraging. Future uses are discussed.

Key Words: Computerized Adaptive Testing (CAT), Item Response Theory (IRT), Dimensionality, Zero-One Liner Programming, Constraints, Item Exposure

I. Introduction

A. Problems Associated with MIF Item-Selection Method

With the combination of advances in the computing power of personal computers and item response theory (IRT, Lord, 1980), computerized adaptive testing (CAT) has substantially increased the breadth of its applications in all areas. CAT involves the selection of test items during the process of administering a test so that each individual takes his/her appropriate difficulty-level items. In the past three decades, researchers have sought promising methods in ability estimation and in item selection for this technology-based assessment. For example, Warm (1989) proposed the weighted likelihood estimation (WLE) in estimating ability parameters. The WLE weights maximum likelihood estimation (MLE) in order to correct MLE biased trait (θ) estimates, especially when they are estimated from small numbers of items. Additionally, Chang and Ying (1996) recommended using the global Kullback-Leibler (KL) information instead of the most commonly used maximum Fisher information for the earlier-stage of CAT item selection.

Currently, there is growing interest focused on whether or not the maximum item-information function (MIF) method is appropriately employed from the beginning through to the end of CAT (e.g., Chang, Qian & Ying, 2001; Hau & Chang, 2001; Li & Schafer, 2003; Veerkamp & Berger, 1999). From a statistical perspective, CAT is based on seeking items with the maximum information for an examinee, which can rapidly improve this examinee's ability estimate**.** According to item response theory (IRT), an item with a high discrimination value will have a high information value while other parameters (e.g., difficulty) are held constant. In addition, an item's information value is dependent on the value of an examinee's ability parameter and consequently an item's information value for the same test-taker will vary because the test-taker's ability estimate is continually changing from early-stages to the end of CAT.

The value of any ability estimate in the early-stage CAT is, however, poorly estimated and is not as accurate as that estimated at the final stage. It thus seems reasonable that an early-stage item's information value is not as meaningful as those in later stages and should therefore have relatively less effect in the process of seeking the "true" ability estimate. Veerkamp and Berger (1999) further demonstrated that an item with the highest discrimination value is not necessarily the most informative item, especially when an examinee's ability parameter is distant from this item's difficulty parameter.

Empirical studies have shown that for the early-stage CAT (e.g., numbers of items less than 10), otherwise promising CAT algorithms still encounter problems resulting in imprecise ability estimates. For instance, Cheng and Liou (2000) evaluated a group of CAT algorithms that resulted from combining MLE and WLE ability estimations with the three item-selection criteria, including Fisher information, KL information and the optimal difficulty (Lord, 1980, to be discussed later). This study showed that ability estimates gradually converged to their true values only after 10 to 20 items were administered. Another comprehensive study (Chen, Ankenmann & Chang, 2000) studied which among a set of sophisticated item-selection methods introduced in that study was the best in facilitating EAP (expected a posteriori) ability estimates at the early-stage of CAT. That study also showed that stable ability estimates were not obtained for the less-than-10-item CAT no matter which selection methods were employed. If both studies also included the random-selection method (randomly selecting items from a pool) in their studies and compared this method with all methods used in both studies, we might speculate that this random-selection method would result in comparable results with those from the maximum-

information-based methods for a short test (e.g., 10) because those methods may not work well enough until the ability parameter is well estimated.

On the other hand, using a maximum-information selection criterion in the early stages of CAT usually leads CAT to overuse of items with high discrimination parameters and thus threatens test security. Further, it makes item exposure uneven in general. Other methods (Meijer and Nering, 1999, provide a literature review) that belong to the family of MIF method (e.g., weighted item information, Berger & Veerkamp, 1997; KL information method, Chang & Ying, 1996) take into account of the uncertainty of CAT's ability estimates, but they, like MIF, tend to select highly discriminating items.

Various remedies to these problems have been proposed (Revuelta & Ponsoda, 1998). The most popular method is the Sympson-Hetter procedure (1985). But because a typical Sympson-Hetter procedure involves maximum-information, either in the iterative process of obtaining the so-called item-exposure control parameter or in the operational stage of CAT, this item-exposure control approach ends up using the more discriminating items first as demonstrated by Hau and Chang (2001). Furthermore, although this procedure was able to successfully control each item's exposure rate to be less than a maximum desirable rate (e.g., 30 percent), CAT under this procedure still resulted in several of the original maximum-information problems (e.g., items in the pools unevenly used; a large proportion of items unused, Revuelta & Ponsoda, 1998). Apparently, the incorporation of Sympson-Hetter with the MIF CAT algorithm will only resolve part of MIF's problems.

B. Alternative Item-Selection Methods

Instead of using the maximum-information-based methods, the method of matching optimal difficulty (MOD) with test-taker's current ability estimate is another alternative. An item's optimal difficulty (OD) value is defined (refer to Equation 3) as the item's maximum information given the set of item parameter estimates (Lord, 1980, p.152). The MOD has been adopted in several CAT studies (e.g., Cheng & Liou, 2000; Li & Schafer, 2003; Warm, 1989; Weiss & McBride, 1984) and its capability to recover ability estimates comparable with those of MIF has been demonstrated (Cheng & Liou, 2000; Li & Schafer, 2003). Its performance on the criterion of item exposure rate is of primary interest in this study.

Another method, similar to MOD, is to match item difficulty (MID) with test-taker's current ability estimate. Dodd (1990) showed that MID increases the use of items in the pool, but Revuelta and Ponsoda (1998) found that it reduces the accuracy of ability estimates.

Chang and Ying (1999) as well as Chang, Qian and Ying (2001) proposed a different approach, called alpha-stratified adaptive testing. A rationale behind this item-selection method is that because the trait estimate is poorly estimated at the early-stage of CAT, using low-discriminating items in the early stages does not have a negative impact on the trait estimate, but increases the exposure rate of these low-discriminating items. Further, those high-discriminating items will be efficiently used at the later-stage CAT.

Up this point, we have introduced several alterative item-selection methods which are purely grounded on statistical criteria, not considering other criteria such as content-balancing and other constraints (e.g., dependencies among items in a pool) into their algorithms. However, if those issues are simultaneously considered as is usually required in CAT, the stratification of both item discrimination and difficulty parameters, as illustrated in the alpha-stratified adaptive testing, together with these additional requirements might become very awkward. This problem

has been addressed by combining the alpha-stratified procedure with the shadow test approach (van der Linden & Chang, 2003).

C. Item-Selection Methods for Constrained CAT

Van der Linden (2000) reviewed several existing methods that incorporate additional constraints (e.g., test specifications) in CAT. The utilization of a shadow test approach (van der Linden and Reese, 1998) to CAT represented one of the more sophisticated item-selection methods that can accommodate various constraints (e.g., content specifications) without the risk of some of constraints that might be violated. Its capability to incorporate other item selection methods (e.g., alternative methods introduced above) as well other item-exposure control methods (e.g., the Sympson-Hetter procedure) into its algorithm is also desirable.

A more detailed description on the shadow-test approach is presented in the latter section of this paper. In short, there are two key components for the shadow-test approach. One is that users set up reasonable constraints and/or item-control methods for the proposed shadow test. The other component is to choose a target function to be maximized (or minimized) throughout the entire CAT testing. If the Fisher information is the function to be maximized, this component could be modified based on the following logic. While the shadow test is assembling at the beginning of the CAT, we are maximizing Fisher information that is not really as meaningful as we anticipate. This fact reminds us not to do so at the earlier-stage CAT, but to maximize (or minimize) other functions to prevent the problems recurred in the maximum-information methods (e.g. overuse of highly discriminating items). In doing so, the goal of increasing the homogeneity of CAT's item-exposure rates might be achieved. This study intended to address this issue of what other functions may be legitimate (or appropriate) to be maximized (or minimized) in the shadow-test CAT approach.

D. Research Purpose

The findings from past CAT studies suggest that pursuing the best combination of ability estimation and item-selection methods might slightly improve the precision of ability estimates, but is not the only solution for CAT. After all, other factors such as the content balance as well as the increase in homogeneity of CAT's item-exposure rates are also essential considerations in the CAT algorithms. This study was designed to find CAT algorithms that address these concerns. The proposed CAT algorithms are intended to limit the drawbacks introduced by existing maximum-information item selection methods but to retain the benefits from the use of maximum-information methods in CAT.

Implementing maximum information in item selection is supposed to work well once the trait estimate approaches its true value. It seems reasonable that this item-selection method should be efficiently employed at the final-stage shadow-test-constraint CAT. But seeking other legitimate functions to be maximized (or minimized) at the early stages of shadow-test CAT also seems reasonable.

In this study, a combination of different CAT algorithms being implemented at different stages of CAT was proposed under the umbrella of shadow-test CAT. There were four target functions to be separately used in the first, second, third and fourth quarter test of CAT. The elements to be used in the four functions were associated with: (a) a random number assigned to each item, (b) the absolute difference between an examinee's current ability estimate and an item difficulty, (c) the absolute difference between an examinee's current ability estimate and an optimum item difficulty, and (d) item information. A more detailed description on how to

incorporate these four functions in the context of shadow tests is provided later and the rationale behind the combination of these four functions is discussed in the conclusion. Of course, we might use the standard error of the ability estimate as a criterion to decide when it is suitable to start using one of these four functions, instead of using each quarter test as the transition point. However, this type of design, like a flexible-length CAT testing, is particularly difficult to implement in current fixed-test-length CATs.

It is therefore likely that the varied-targeted-function rather than the single-targeted-function shadow-test CAT will expand the use of all characteristics of items in the pool, guarantee meeting all constraints, and maintain the accuracy of the final-stage trait estimate. This COMBINED CAT algorithm was implemented in order to explore its use and to compare its performance with those using individual methods alone (e.g., MIF, MOD, MID, RANDOM).

II. Overview of CAT Techniques

A. 3PL Model

The commonly-used three-parameter (3PL) logistic IRT model was used to model the dichotomous scored items in this study. Under the 3PL model, the probability, P_{ji}, of the correct response on an item i for an examinee with ability θ_j is given by the following function (Lord, 1980).

$$P_{ji} = c_i + (1 - c_i)\frac{\exp(Da_i(\theta_j - b_i))}{1 + \exp(Da_i(\theta_j - b_i))} \quad (1)$$

where
the symbol of "exp" stands for the mathematical function of the natural logarithm exponential,
a_i is the item discrimination,
b_i is the item difficulty,
c_i is the lower asymptote parameter (also known as the guessing parameter), and
D is a scaling factor (usually equal to 1.702).

The scaling factor D is included in the model to make the logistic function as close as possible to the normal ogive function (Baker, 1992).

B. Computations of FI and OD Values

1.Fisher Information (FI)

A Fisher information is computed at the current ability estimate, that is:

$$I_i(\theta) = \frac{[P_i'(\theta)]^2}{P_i(\theta)Q_i(\theta)} \quad (2)$$

where, $P_i'(\theta)$ is the first partial derivative of $P_i(\theta)$ with respect to θ.

2. Optimal Difficulty (OD)

For the 3PL, an item's maximum information is located, based on the item-pool item's difficulty scale, at (Lord, 1980, p.152):

$$OD = b_i + \frac{1}{Da_i} \ln\left(\frac{1+\sqrt{1+8c_i}}{2}\right) \tag{3}$$

For the case of 3PL data modeling, both FI and OD values are derived from the 3PL item's parameters, a, b, and c. A difference between them is that an FI value depends on a location on the ability (θ) scale; in contrast, the OD does not. An FI value may be misleading, as discussed previously, if it is computed from an ability estimate that is far away from its true value.

C. CAT Ability Estimates

1. Maximum Likelihood Estimator (MLE)

Assuming that the local independence assumption holds, then given an examinee with an ability, θ, who responds to a set of n items with the response pattern $\underline{u}$, the probability (or likelihood) of obtaining this response pattern $\underline{u}$ can be modeled by:

$$L = L(\underline{u} \mid \theta) = \prod_{i=1}^{n} P_i(\theta)^{u_i} Q_i(\theta)^{1-u_i} \quad , i \in S_n \tag{4}$$

where $Q_i(\theta) = 1 - P_i(\theta)$ and S_n connote the n items that have been administered (or selected) to the examinee during the CAT testing process. The log of this likelihood function is given by:

$$\ln(L) = \sum_{i=1}^{n} \left[u_i \log(P_i(\theta)) + (1-u_i)\log(Q_i(\theta))\right] \tag{5}$$

Several methods of CAT ability estimation exist. For the maximum likelihood estimate of θ, MLE (θ), the log likelihood function (Equation 5) should be partially differentiated with respect to θ, then set to equal zero and finally used to solve this equation 6 for θ using the Newton-Raphson method or some other suitable numerical strategy. This equation is given below (Lord, 1980) :

$$\frac{\partial \ln(L)}{\partial \theta} = \sum_{i=1}^{n} \frac{P_i(\theta)' (u_i - P_i(\theta))}{P_i(\theta) Q_i(\theta)} \tag{6}$$

where ∂ denotes partial differentiation. A problem for the MLE is that it is unable to estimate examinees' abilities when they get items all right or all wrong. This has become an issue especially at the early stage of CAT. Thus, this estimator was not considered in this research.

2. Bayesian Modal or Mean Estimator

The parameter θ needs to be estimated. If the prior information f(θ) for the distribution (or probability density) of θ together with the observed response pattern $\underline{u}$, are available, we are then able to approximate the posterior distribution of θ according to Bayes' rule. The posterior density of θ is:

$$f(\theta \mid \underline{u}) = \frac{L(\underline{u} \mid \theta) f(\theta)}{f(\underline{u})} \tag{7}$$

Where $f(\underline{u})$ is the marginal probability of $\underline{u}$ given by Bock and Lieberman (1970) and Bock and Aiken (1981):

$$f(\underline{u}) = \int_{-\infty}^{\infty} L(\underline{u} \mid \theta) f(\theta) d\theta \tag{8}$$

The function $f(\underline{u})$ is irrelevant while finding the solution of the θ parameter. Hence, the posterior function can simply be proportional to a prior function times a likelihood function. That is:

$$f(\theta \mid \underline{u}) \propto L(\underline{u} \mid \theta) f(\theta) \tag{9}$$

The relative influence of observed data (the input for the likelihood function) and prior information on the posterior function (related to the updated belief) depends on test-lengths, item-pool characteristics, and the magnitude of prior dispersion. As the prior becomes vague or diffuse, the posterior function is closely approximated by the likelihood function and consequently the Bayesian approach will result in the same solution as the likelihood approach. In contrast, if the prior is very informative or specific, then it would have a relatively greater influence on the posterior function.

For the maximum a posteriori (MAP) estimator , the estimate is the value that maximizes the posterior density function of $f(\theta \mid \underline{u})$. The $\hat{\theta}$ can be derived by partially differentiating the log-posterior density function with respect to θ , setting this equation equal to zero (Equation 10), and solving this non-linear equation:

$$\frac{\partial}{\partial \theta} \ln f(\theta \mid \underline{u}) = 0 \tag{10}$$

MAP is the mode of the posterior distribution. Another method to solve equation 10 is to find the mean of the posterior distribution of θ. This method is called the expected a posteriori (EAP) estimator. The mathematical expression for this estimator can be found in Bock and Aitkin (1981) and its features in CAT has been well documented by DeAyala, Schafer, and

Sava-Bolesta (1995). Compared with MAP estimation (Wang, Hanson & Lau, 1999), EAP had slightly lower standard errors, but was slightly more biased.

C. The Shadow-Test CAT

1. Zero-One Liner Programming

If factors such as content balance and item-exposure control are needed to be taken into account during the process of selecting items for examinees, the technique of zero-one linear programming (Theunissen, 1985, 1986; van der Linden & Boekkooi-Timminga, 1989) is a suitable method that can be easily and effectively adapted into the CAT process. A description of zero-one linear programming is presented below, before the shadow test CAT, that utilizes the zero-one linear programming (van der Linden & Reese, 1998), is illustrated.

Linear programming is designed to seek the maximum value for a linear function such as Equation 11 while the required constraints formalized in Equation 12 are imposed.

$$\text{Maximize } \sum_{i=1}^{L} \text{Information}_i(\hat{\underline{\theta}})x_i, \tag{11}$$

subject to

$$\mathbf{A} \cdot \underline{x} \square \underline{b} \tag{12}$$

where $\square$ could be <, =, or > ,

and for i to L,

$$x_i \in \{1,0\} \tag{13}$$

The usual target function to be maximized in Equation 11 is the item information. The target function can be replaced with others as illustrated later.

In Equation 11, the items in the bank are indexed by i=1,…,L and the values in the variable x_i are parameters that will be estimated. For zero-one linear programming, the x values are constrained to be either one or zero as indicated in Equation 13 to identify whether an item is a qualified candidate item. The value of one or zero in the decision variables, x, indicates whether the items are selected or not for the test.

A vector will be created as shown below before seeking the solution---the binary decision values for the x-vector. The matrix of **A** together with vector of $\underline{b}$ are used to specify what constraints are specified. In the present example we have an item pool with 10 items; the first five items belong to the first domain and the rest of the items belong to the second domain. In addition, there are two constraints to be imposed, namely 2 and 3 items from these two domains, respectively, which are expected to be administered to each examinee. Under this testing scenario, the **A** matrix and $\underline{b}$ vector will be created as shown below before seeking the solution of the vector parameter $\underline{x}$.

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

$$\underline{b} = \begin{bmatrix} 2 \\ 3 \end{bmatrix}$$

The number of columns in the **A** matrix should equal the number of items in the pool. In addition, each row in the **A** matrix together with the corresponding row in the $\underline{b}$ vector expresses

a single constraint. The first constraint as described above is expressed in the first row in the **A** matrix together with the first row in the $\underline{b}$ vector. The series of five 1's indicate that the first five of the 10 items are from Domain 1 and the last series of five 0's indicate that the last five of the 10 items are not from Domain 1. Finally, the condition of 2 items in Domain 1 to be picked is specified as "2" in the first row in the $\underline{b}$ vector.

2. CAT with Shadow Tests

The shadow-test CAT was proposed by van der Linden and Reese (1998) using zero-one linear programming to impose constraints. Various types of constraints that can be employed in CAT are enumerated in van der Linden and Reese (1998) and van der Linden (2000). The algorithms (refer to van der Linden & Reese, 1998; van der Linden, 2000) for this constraint under CAT are:

(1). Set an initial ability estimate, for example, a value generated from the uniform distribution with range [0,1].

(2). Assemble an on-line shadow test that: (a) has met all the constraints as specified by Equation 12, (b) maximizes information value on the Equation 11 at the provisional ability estimate, and (c) includes the previously administered item(s).

(3). Administer the item with maximum information among items from the on-line shadow test that have not yet been administered.

(4). Re-estimate abilities based on the examinee's response(s) to the items that have been administered.

(5). Release all unused items that have been previously included in the shadow test to the item pool.

(6). Add an additional constraint to the constraints that have been imposed in the zero-one linear model to ensure that the item being recently selected and administered to the examinee "must" be included in the next updated on-line shadow test. For matrix expression in the above example, if Item 2 is selected first, the **A** matrix and the $\underline{b}$ vector should be updated in the following way:

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

$$\underline{b}=\begin{bmatrix} 2 \\ 3 \\ 1 \end{bmatrix}$$

The last row in the A matrix combined with the last row in the $\underline{b}$ vector indicates that Item 2 is part of the updated shadow test.

(7). Repeat the procedures 2-6 until the fixed n items (or other criteria) have been administered (or met).

In short, this shadow-test CAT begins with assembling an on-line "full-shadow" test that meets all the desirable constraints and has maximum information (or other statistical criteria) at a provisional ability estimate. An item with maximum information (or other statistical criteria) is then selected from this shadow test, instead of directly from the full item pool. Each of the next serial shadow tests includes all previously administered items and consequently the last shadow test is an actual CAT test which always meets all constraints.

D. CAT Item Selection In the Context of Shadow Test

The item selection methods used in this study are defined in the context of the shadow-test constraint:

(1) MIF: Maximize the summation of all items' information values while assembling shadow tests, then select the item with the maximum information from the shadow test.

(2) MOD: Minimize the summation of all items' absolute difference values between a test-taker's current ability estimate and items' OD values while assembling shadow tests; then select the item with the minimum absolute difference from the shadow test.

(3) MID: Minimize the summation of all items' absolute difference values between the test-taker's current ability estimate and items' difficulty values while assembling shadow tests; then select the item with the minimum absolute difference from the shadow test.

(4) RANDOM: Each item is assigned a random number (RN). This RANDOM method used a similar procedure as MIF method did for selecting an item. The difference between both methods is that instead of maximizing the summation of all items' information values in the MIF method, the RANDOM method maximizes the summation of all items' RN values while assembling shadow tests. There is no operational value in maximizing the summation of all items' RN, but this tricks the zero-one linear programming to assemble shadow tests that meet constraints. Accordingly, this RANDOM method is similar to randomly select an item from the pool; the difference here is that this RANDOM method also takes into account the required constraints.

III. Methodology

A. Simulated Item Bank

An item response matrix from 9351 examinees on the 45 Reading/Language Arts items on the CTBS second-grade test (CTB/McGraw-Hill, 1997) was used to generate a simulated item bank. Forty five items are not enough to form an item pool. In order to simulate a large item pool that covers a variety of possible combinations of the 3PL item parameters, a, b and c, and to maintain the intercorelations of item parameters for the items in the pool, we relied on the principle that the numerical value of each item parameter estimate is dependent on the metric of the underlying latent trait (θ). This fact allows us to obtain another numerical value for the same item when we intentionally change the metric of the latent trait. Accordingly, the 9351 examinees were first grouped into six subgroups by the level of the examinee's abilities. We then purposely combined some subgroups together (e.g., the combination of groups 1, 3 and 5 or groups 2, 4 and 6, or groups 1, 4 and 5, etc.) to form a sample with about 3500 examinees. Afterwards, the computer program of BILOG (Mislevy & Bock, 1990) was used for item calibrations with this sample's test data to obtain a set of item parameters. The above procedure was repeated 13 times and a 585-item bank was created, whose summary of item characteristics is presented in Table 1.

It is important to note that these numerical values of parameters calibrated from different-ability samples are assumed to lie on the same scale even though they do not. We used this process of generating an item pool in order to simulate a larger item pool having similar diversity as that which would be obtained from existing test data in terms of the variance of the items'

parameters, a_i, b_i, and c_i. In addition, the existing intercorrelation among item parameters in the real data will continue to exist in the simulated item pool.

Table 1:
Summary of Item Characteristics for the Simulated Item Pool
(Number of Items = 585)

Item Parameter	Mean	SD	Minimum	Maximum
Discrimination	.88	.35	.35	2.4
Difficulty	-.40	1.18	-4.44	4.15
Guessing	.27	.10	.05	.45

B. True Ability Levels and Test Starting Points

We included 13 points on the true ability or θ scale, ranging from –3.0 to 3.0 in increments of 0.5. The initial ability of all simulated subjects at the beginning of the test was found as a random drawn from the uniform distribution with range [0,1]. The EAP ability estimator was used to estimate abilities, assuming the prior distribution of examinee's abilities to be distributed as a standard normal N(0,1).

C. Item Selections under the Varied-Function Shadow Test Constraints

Table 2 presents the five CAT algorithms explored. The first is a typical shadow-test CAT and is expected to result in the most precise ability estimate among five, but may result in poorest homogeneity of item-exposure rates. The result from this condition served as a basis for comparisons with other CAT algorithms when the shadow test approach was used to maintain content balance in CAT.

The MID and MOD algorithms were employed in the second and third simulation conditions, respectively. The fourth was the RANDOM shadow test algorithm, which was expected to result in the poorest ability estimates, but the greatest homogeneity of item-exposure rates.

The above four shadow-test CAT algorithms employed a unique function to be maximized or minimized from the beginning throughout the end of CAT. The fifth simulation condition was created to combine these four algorithms together (called COMBINED) and implemented them for each quarter test when a CAT was divided into four-quarter-length subtests. As seen in Table 2, RANDOM was used for the first quarter, MID for the second, MOD for the third, and MIF for the last. This algorithm was motivated in part by an assumption that an ability estimate at the fourth-quarter of a CAT is accurate and stable enough to efficiently and effectively implement the maximum-information algorithm.

Table 2: Simulated CAT Algorithms

Condition	Q1	Q2	Q3	Q4	**Dependent Variable**
MIF	MIF	MIF	MIF	MIF	**Item-exposure Rates** **Ability Estimates**
MID	MID	MID	MID	MID	
MOD	MOD	MOD	MOD	MOD	
RANDOM	RANDOM	RANDOM	RANDOM	RANDOM	
COMBINED	RANDOM	MID	MOD	MIF	

D. Test Constraint and Test Lengths

There were four constraint conditions imposed to assemble an on-line shadow test. Table 3 lists the item classification for the CTBS, Grade 2 Test. The first condition required one item from each of the 8 domains and created a 8-item test. The second required one item from each of 15 objectives and created a 15-item test. The third required two items from each of 15 objectives and created a 30-item test. These three test lengths were used in the first four CAT algorithms (MIF, MID, MOD, and RANDOM).

The fourth constraint condition corresponded to the specifications listed in Table 3 that were used by the publisher in creating the CTBS Reading/Language Arts. Hence, the number of items for each objective on the shadow test was constrained as in the original test consisting of 45 items. The fifth CAT algorithm of Combined was used for this test length.

Table 3
Item Classification Table for CTBS, Grade 2 Test

Domain	Objective and Skills	Number of Items
1 Basic Understanding	vocabulary	3
	stated information	5
2 Analyze Text	conclusions	4
	story elements/character	3
3 Evaluate and Extend Meaning	predict/hypothesize	1
	extend/apply meaning	3
4 Identify Reading Strategies	make connections	2
	vocabulary strategies	4
5 Introduction to Print	word analysis	7
6 Sentence Structure	complete/fragment/run-on	2
7 Writing Strategies	sequence	2
	relevance	2
8 Editing Skills	usage	3
	punctuation	1
	capitalization	3

E. Computer Program

The computer program CAT was used for running the simulation conditions. The CAT was coded using the MATLAB matrix language (The MathWorks, 2001), in which the 0-1 linear programming was resolved from the callable library of LINDO API (LINDO Systems, Inc. 2001). Technically, the LINDO API was called into the CAT to seek the solution of the vector of $\underline{x}$ in Equation 11.

F. Data Analyses and Evaluation

One hundred replications for each condition were conducted. Afterward, the BIAS and RMSE (root mean squared error) for each of the ability estimates were calculated by the formulas shown below.

$$BIAS(\theta_j) = \frac{\sum_{j=1}^{r}(\hat{\theta}_j - \theta_j)}{r} \quad \text{and} \quad (14)$$

$$RMSE(\theta_j) = \sqrt{\frac{\sum_{j=1}^{r}(\hat{\theta}_j - \theta_j)^2}{r}} \quad (15)$$

where θ_j is the true ability parameter, $\hat{\theta}_j$ is the corresponding estimated ability parameter, and r is the number of replications, which was 100 in this study.

RMSE is a measure of total error of estimation that consists of systematic error (BIAS) and random error (SE). These three indexes relate to each other as follows (Rao, 2000):

$$RMSE(\theta_j)^2 \cong SE(\theta_j)^2 + BIAS(\theta_j)^2 \quad (16)$$

As can be seen from Equation 16 either a large variance (SE^2) or a large BIAS will produce a large RMSE. The accuracy of an estimator is inversely proportional to its RMSE so that the RMSE index is the best criterion for accuracy of an estimator (Rao, 2000). Accordingly, this index was primarily used to compare the accuracy of ability estimates when they were estimated under various simulation conditions. We also provided the BIAS results for reference if needed.

The item-exposure rate refers to the ratio of the number of times an item has been administered to the total number of test-takers. The following indices (refer to Revuelta & Ponsoda, 1998) were used to compare the five CAT algorithms: (a) the percentage of items never administered in the population, (b) the standard deviation (SD) of the variable of the item-exposure rate, (c) the minimum and maximum values of this variable. The distribution of the item-exposure rates, grouped in several intervals, were also computed for each CAT condition.

IV. Results

A. The Effect of CAT Algorithms on the Ability Estimate

1. Test Length =8

Figure 1a shows BIAS as a function of true θ for MIF, MID, Random and MOD methods for test length (TL) = 8. As test length was small (e.g., TL= 8), this condition can be analogous to the early-stage of CAT. All algorithms show similar patterns. Large BIAS (>1.5 or <-1.5) of CAT ability estimates were produced for the highest and lowest abilities either using the statistically sound method, MIF, or the unreasonable method, RANDOM.

Among these four item-selection methods, MIF was the best method in producing the least BIAS of CAT's ability estimates, RANDOM was the poorest, and MOD and MID were ranked between the best and the poorest. This is because MOD tended to perform slightly better in the high abilities than MID did, but it performed slightly worse in the low abilities than MID did.

In terms of accuracy (or RMSE) of CAT ability estimates as shown in Figure 1b, the same rank ordering of these four methods as that found for the BIAS measure was observed.

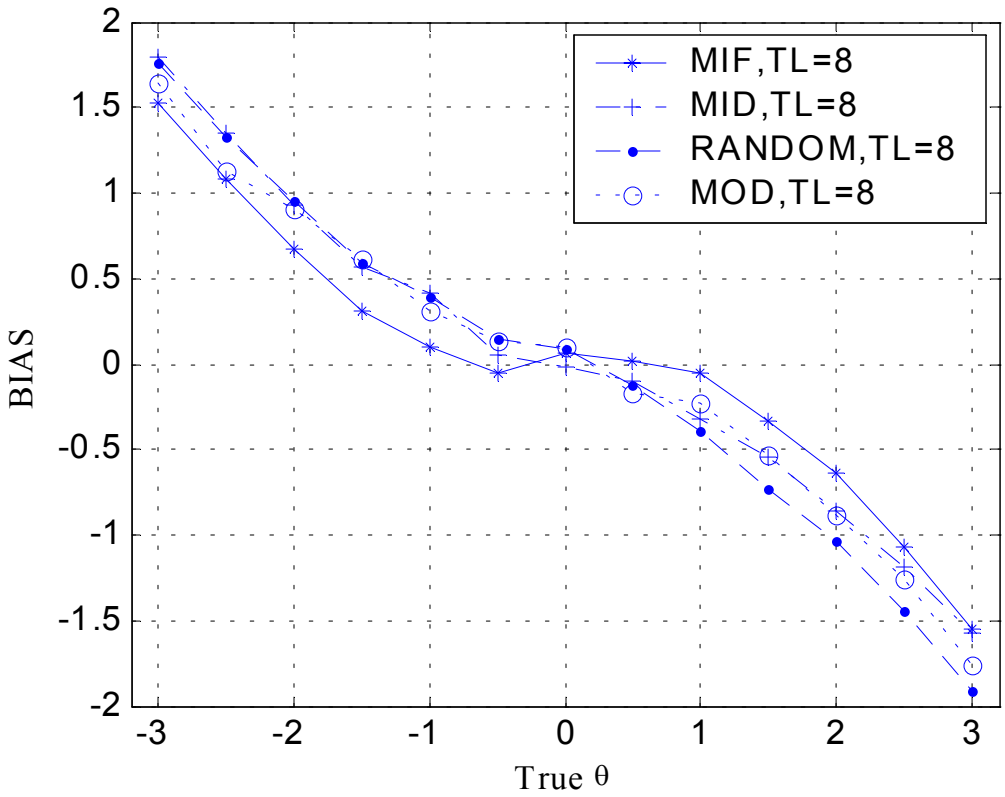

Figure 1a. BIAS as a Function of True θ for MIF, MID, RANDOM and MOD Methods When Test Length (TL) = 8.

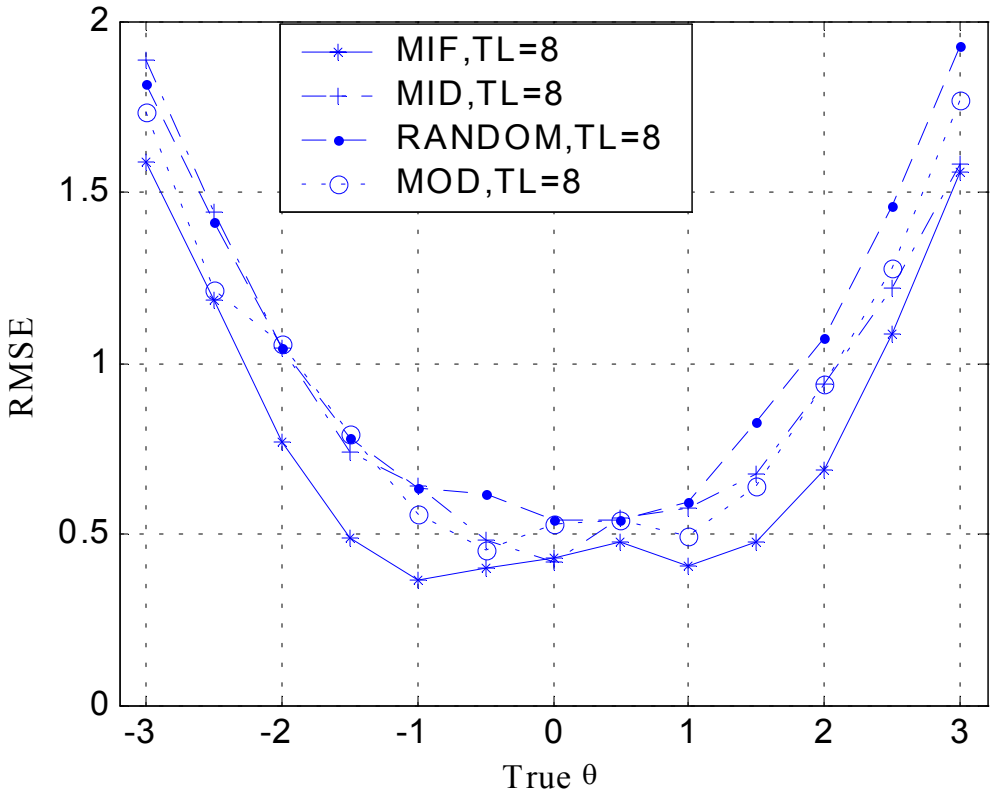

Figure 1b. RMSE as a Function of True θ for MIF, MID, RANDOM and MOD Methods When Test Length (TL) = 8.

2. Test Length=15

Figures 2a and 2b shows the BIAS and RMSE results for these four CAT methods when test length was set at 15. The BIAS or RMSE for the low and high abilities were still relatively large for the case of TL=15, although some improvement occurred over TL=8. The ranking of performance among these four methods was as the same as found in the case of TL=8.

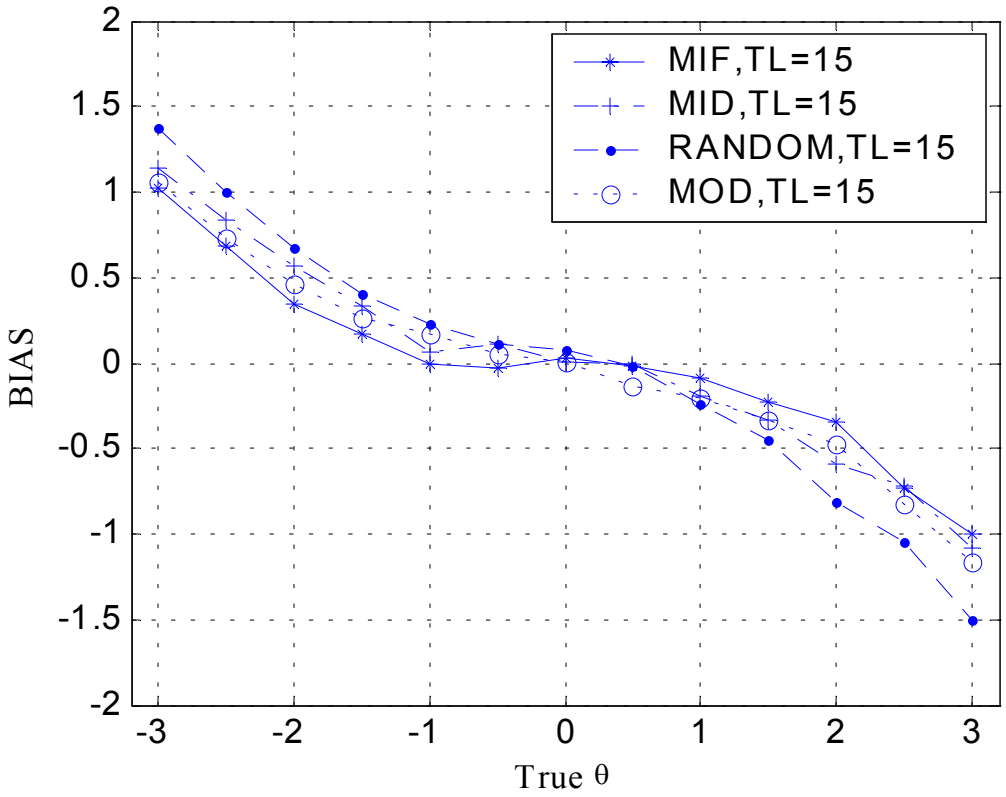

Figure 2a. BIAS as a Function of True θ for MIF, MID, RANDOM and MOD Methods When Test Length (TL) = 15.

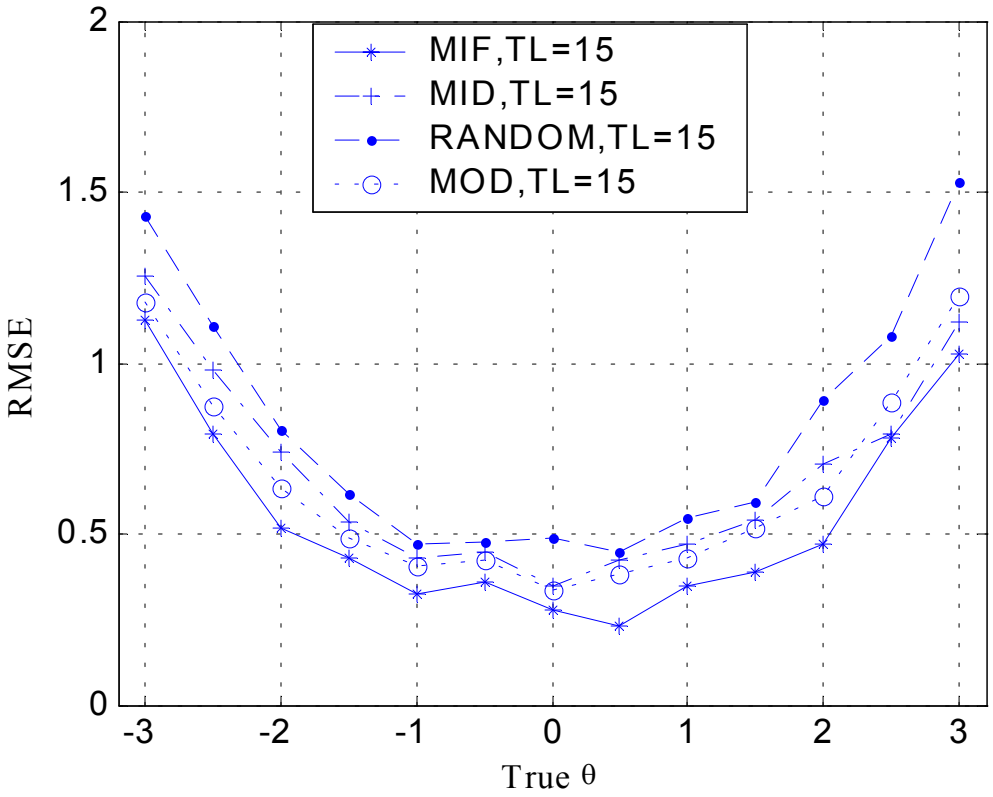

Figure 2b. RMSE as a Function of True θ for MIF, MID, RANDOM and MOD Methods When Test Length (TL) = 15.

3. Test Length=30

Figures 3a and 3b show the BIAS and RMSE results for these four CAT methods when test length was 30. Under TL=30, the best method, MIF, made a sizeable improvement in recovering those high and low abilities that were relatively poorly estimated when test length equaled 15 or 8. For the condition of TL=30, the MOD or MID, in general, produced comparable results with the MIF. The MID, performed slightly poorer than the MOD at some ranges of abilities, but at some ranges of low abilities this method produced less BIAS or RMSE than the MOD did. The poorest method, RANDOM, did not make significant improvement as the rest of the three methods did as the test length increased, especially for the low and high abilities.

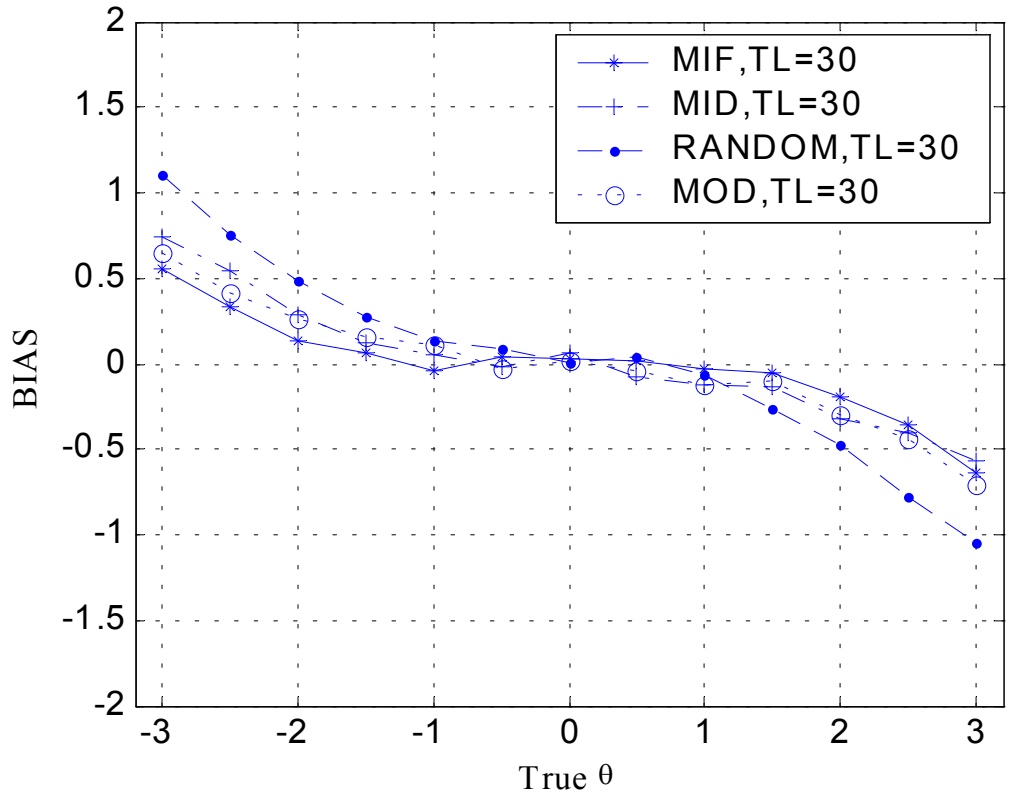

Figure 3a. BIAS as a Function of True θ for MIF, MID, RANDOM and MOD Methods When Test Length (TL) = 30.

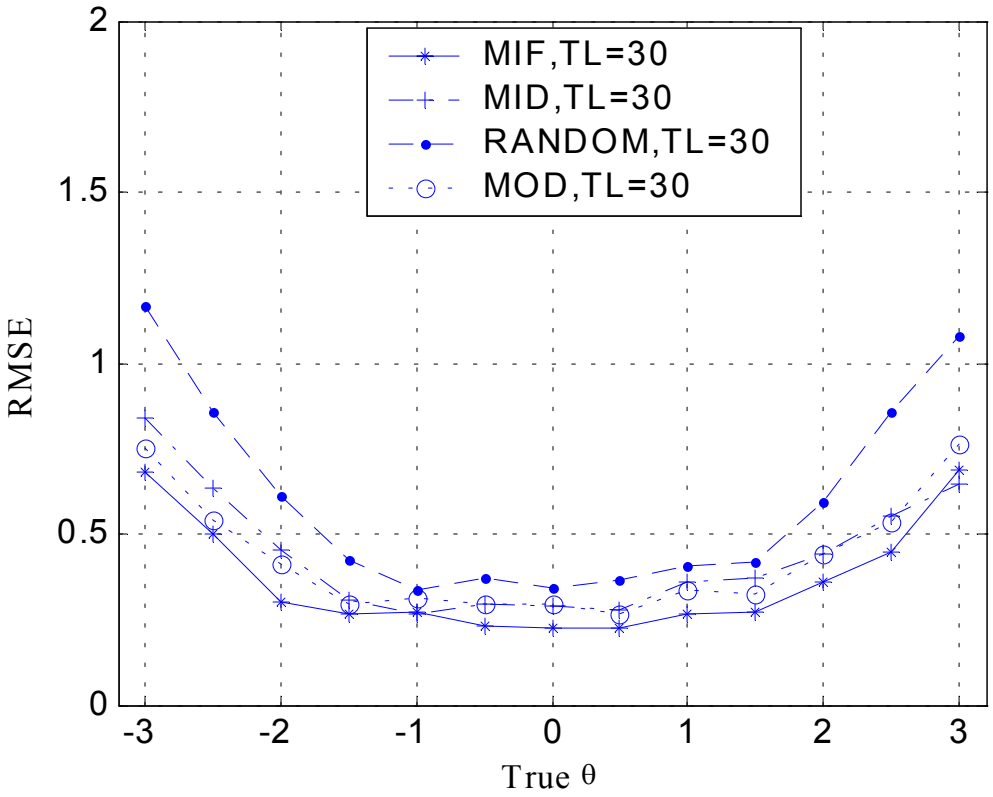

Figure 3b. RMSE as a Function of True θ for MIF, MID, RANDOM and MOD Methods When Test Length (TL) = 30.

4. Test Length=45

For the condition of TL=45, the COMBINED method together with the other four methods was used. As expected, this method became the second best algorithm in recovering CAT's abilities, as seen in Figures 4a and 4b, which shows BIAS and RMSE as a function of true θ for these five methods. This COMBINED method was created by combining of the best method, MIF, the poorest method, RANDOM, and the other two alternative methods, MID and MOD. Its performance in estimating abilities only slightly differs from the best method.

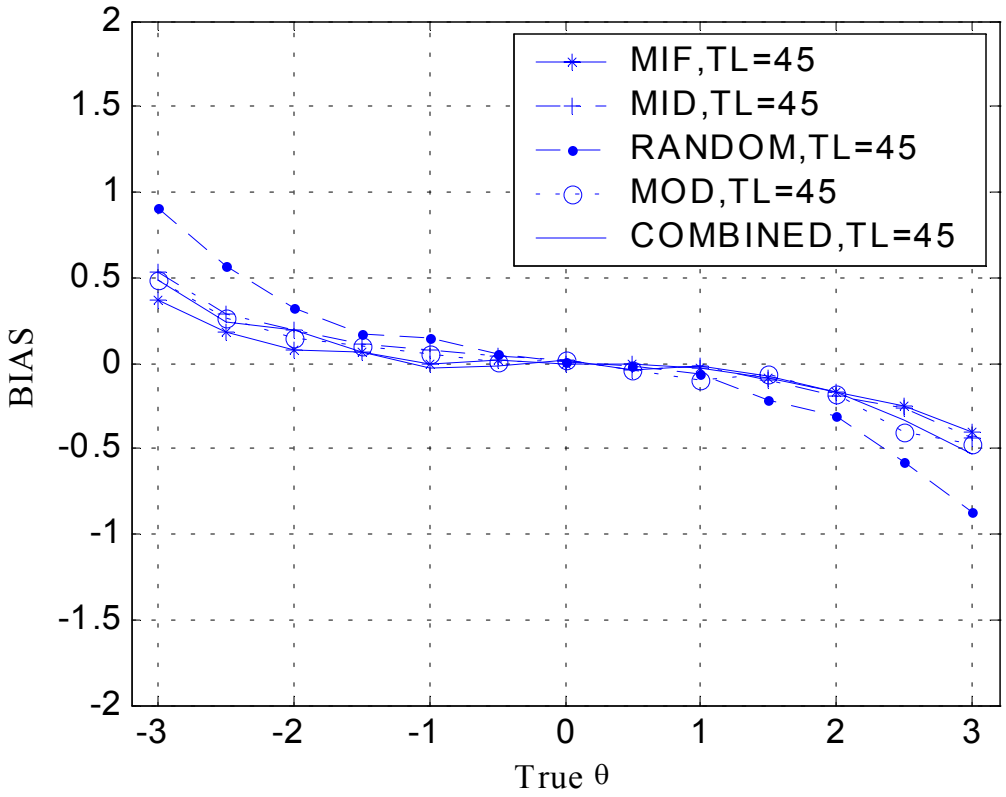

Figure 4a. BIAS as a Function of True θ for MIF, MID, RANDOM, MOD and COMBINED Methods When Test Length (TL) = 45.

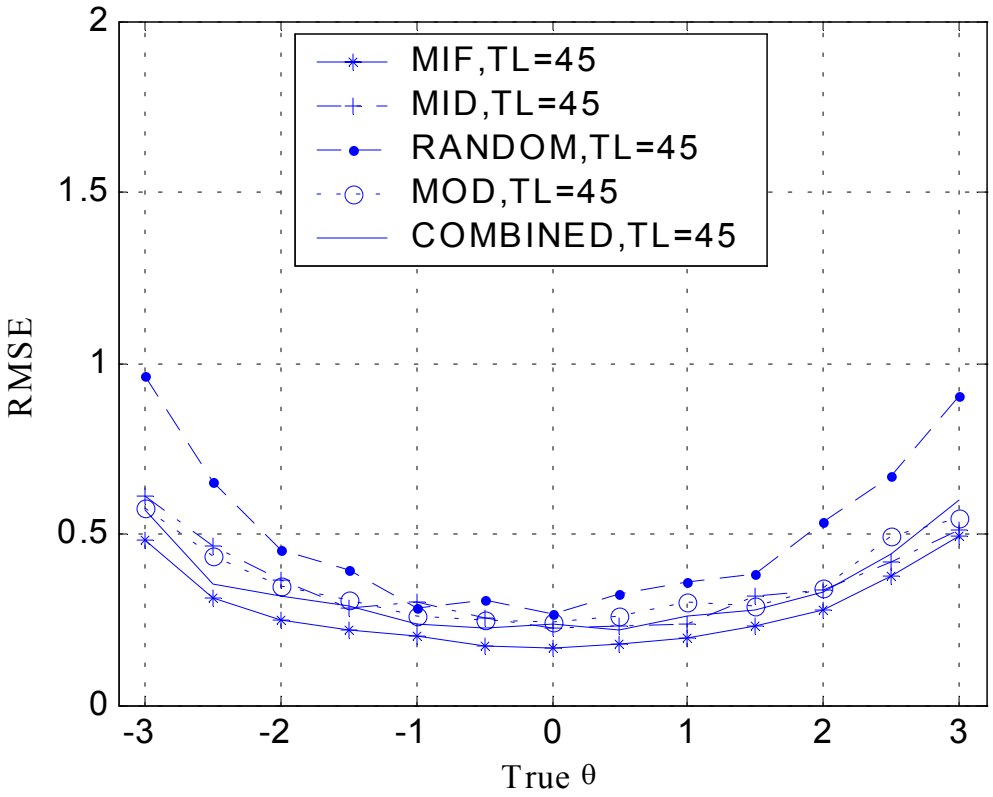

Figure 4b. RMSE as a Function of True θ for MIF, MID, RANDOM, MOD and COMBINED Methods When Test Length (TL) = 45.

B. Item Exposure Rate

1. Comparisons of Exposure Rates Among Five CAT Methods

The results of item exposure rates when TL=45 for the Random, MIF, MID, MOD and COMBINED methods are presented in Table 4. The plot of each item's exposure rate for these five CAT methods are presented in Figure 5.

The rate of unused items is high in MIF; 54.87% of the items were never selected for any of the 5000 examinees. In contrast, the poorest CAT's ability estimator, Random used all items in the pool. The MID and MOD had about 1.20% of the items unused. The second best ability estimator, COMBINED, like the RANDOM method, used all items.

If the maximum desirable item-exposure rate was set at .30, which is a typical value (ranging from .2 to .4 in the context of the Sympson-Hetter's item-exposure control method), the percent of items whose item-exposure rates are larger this criterion for each method is presented in the last second row from the bottom in Table 4. The overexposure rate beyond this criterion is 9.40% for MIF, 3.24% for MID, 2.56% for MOD and 1.7% for COMBINED. This finding implies that the COMBINED worked well in producing low percent of overexposured items even though the item-exposure control was not involved into its algorithm.

Excluding the RANDOM, the COMBINED method was best at producing the most homogeneous item exposure rates, MOD was second, MID was third, and MIF was poorest. Compared with the variance of item exposure rates produced by MIF, the COMBINED, MOD

and MID reduced the variance of item exposure rates by 80.54%, 66.68% and 57.17%, respectively. Those figures are presented in the last row of Table 4. The ultimate goal of increasing the homogeneity of CAT's item-exposure rates, which has been pursued in this research, was achieved by using the COMBINED algorithm. This finding combined with the results regarding the accuracy of CAT ability estimates suggests that when compared with MIF CAT, the COMBINED CAT is the CAT algorithm of choice because this method not only resulted in comparable results as the MIF CAT, but also made full use of all items in the pools. On the other hand, more than half of items in the pools would never be administered to examinees under MIF CAT testing method under the realistic conditions studied here.

Table 4
Item Exposure Rate Distribution for the RANDOM, MIF, MID, MOD and COMBINED Methods Under the Shadow-test-constraint Control

Item Exposure Rate (x100)	Methods				
	RANDOM	MIF	MID	MOD	COMBINED
0: Items Never Administered	0.00	54.87	1.20	1.20	0.00
1-5	14.86	16.23	50.09	46.85	41.02
5-10	67.53	5.83	27.18	29.74	35.04
10-15	16.58	4.62	8.89	10.60	13.68
15-20	1.03	3.08	4.62	4.27	4.96
20-25	0.00	3.76	2.39	3.25	2.22
25-30	0.00	2.22	2.39	1.54	1.37
30-35	0.00	1.88	1.03	1.03	0.85
35-40	0.00	1.03	0.68	0.17	0.34
40-45	0.00	1.71	0.17	0.17	0.34
45-50	0.00	1.54	0.17	0.34	0.17
50-60	0.00	2.05	0.34	0.51	0.00
60-70	0.00	0.68	0.51	0.17	0.00
70-80	0.00	0.34	0.34	0.17	0.00
80-90	0.00	0.17	0.00	0.00	0.00
90-99	0.00	0.00	0.00	0.00	0.00
100: Items Always Administered	0.00	0.00	0.00	0.00	0.00
Mean	7.69	7.69	7.69	7.69	7.69
SD	2.75	14.76	9.66	8.52	6.51
Minimum	0.60	0.00	0.00	0.00	0.02
Maximum	17.32	81.38	78.96	70.78	47.08
Larger than 30	0.00	9.40	3.24	2.56	1.70
Reduced Variance of Item Exposure Rate	96.53	Anchor	57.17	66.68	80.54

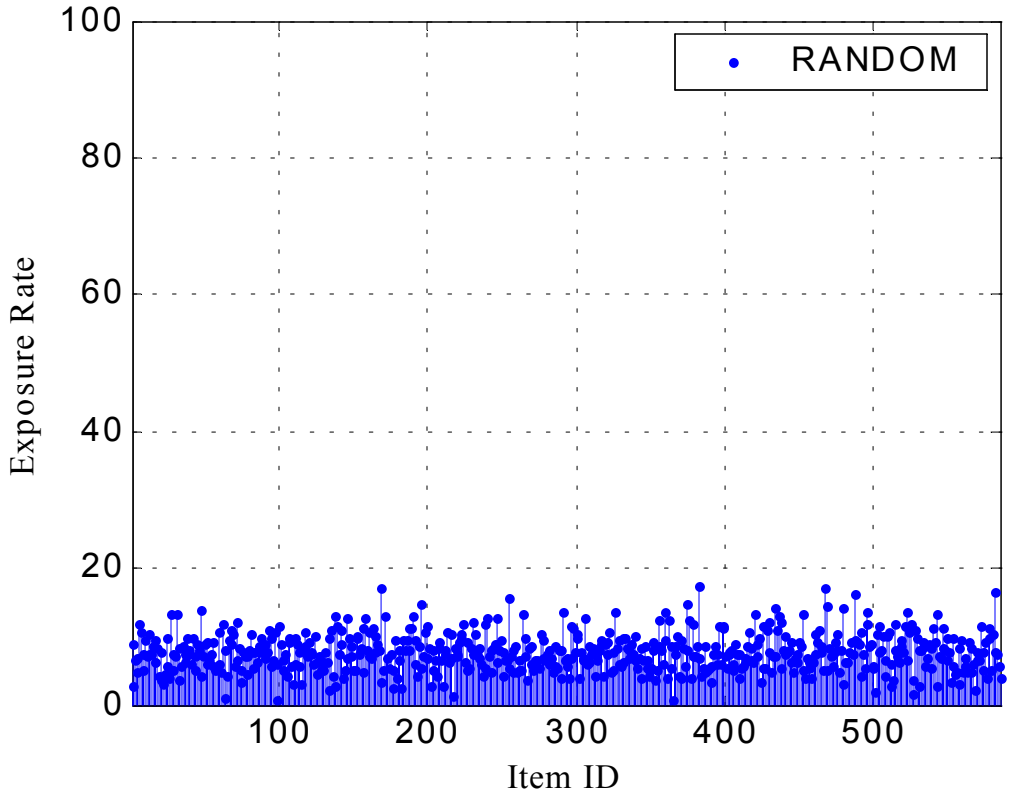

Figure 5a. Plot of Each Item's Exposure Rate for the RANDOM Method

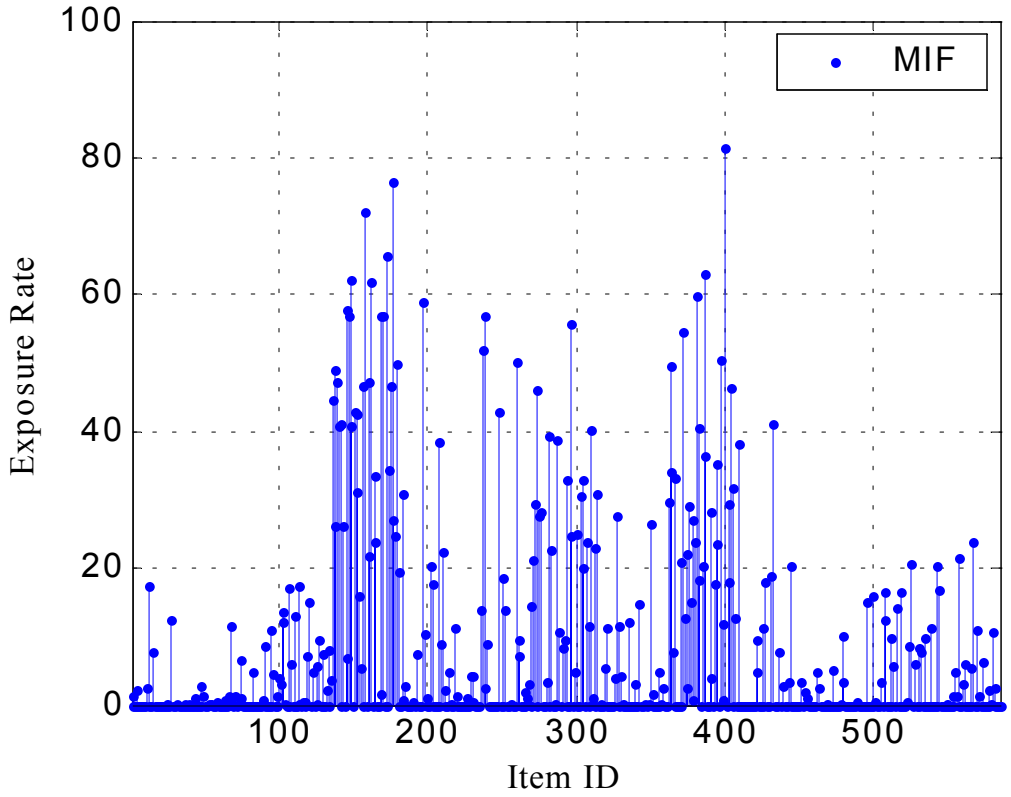

Figure 5b. Plot of Each Item's Exposure Rate for the MIF Method

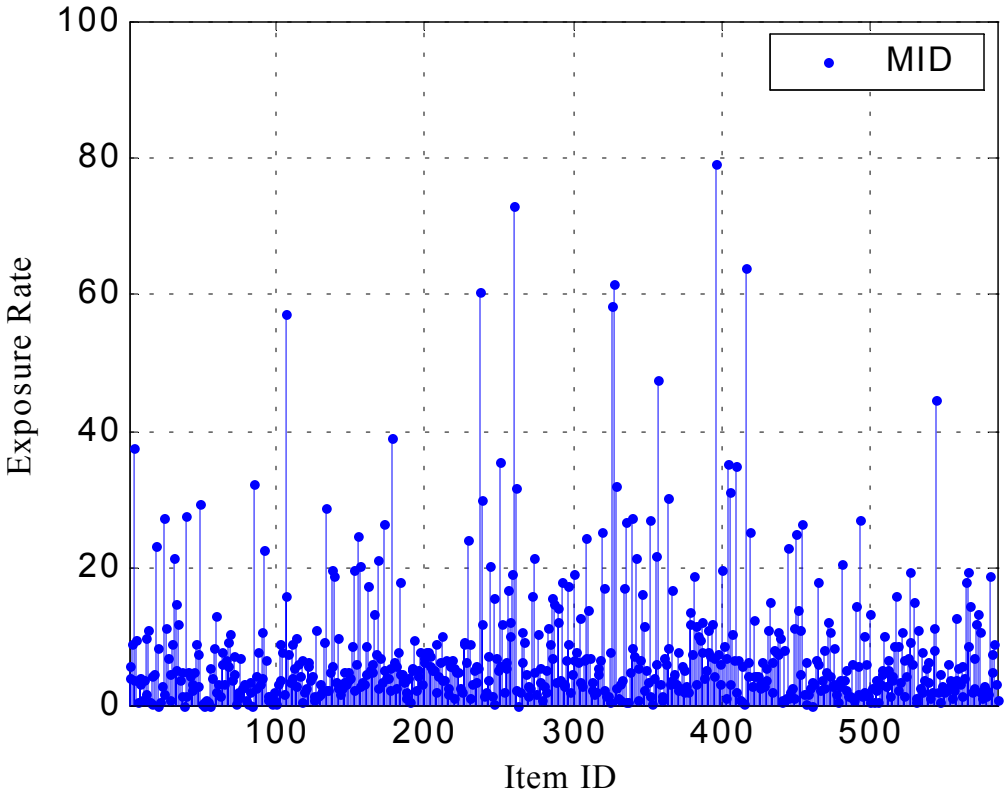

Figure 5c. Plot of Each Item's Exposure Rate for the MID Method

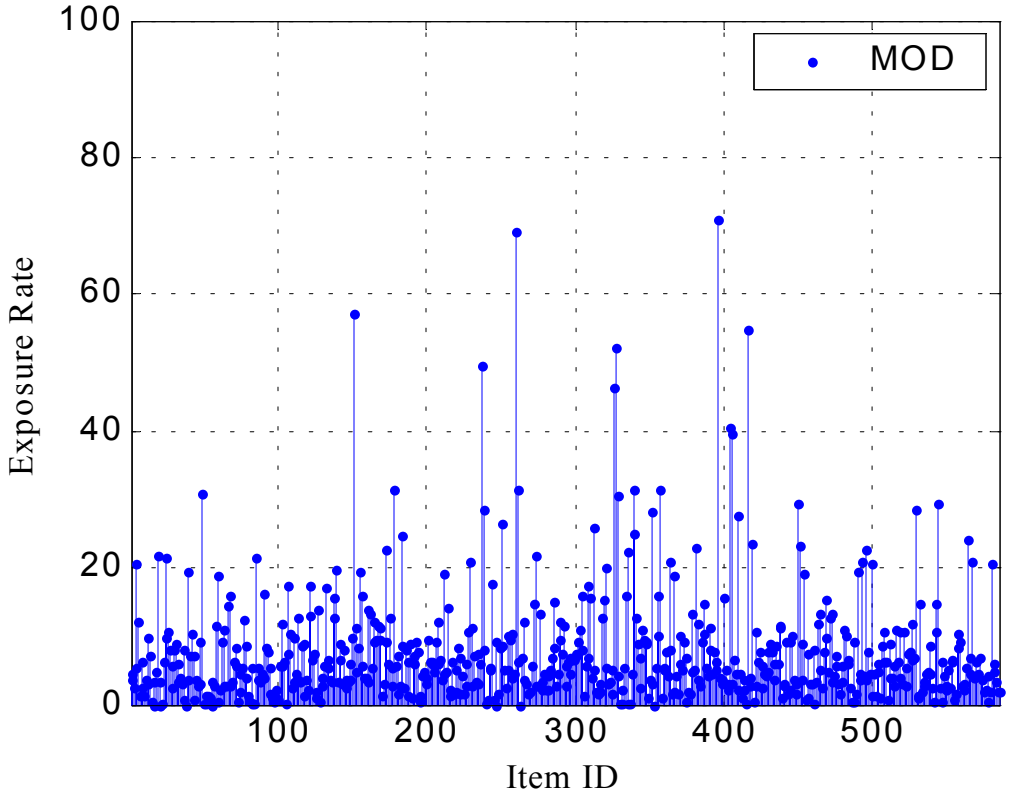

Figure 5d. Plot of Each Item's Exposure Rate for the MOD Method

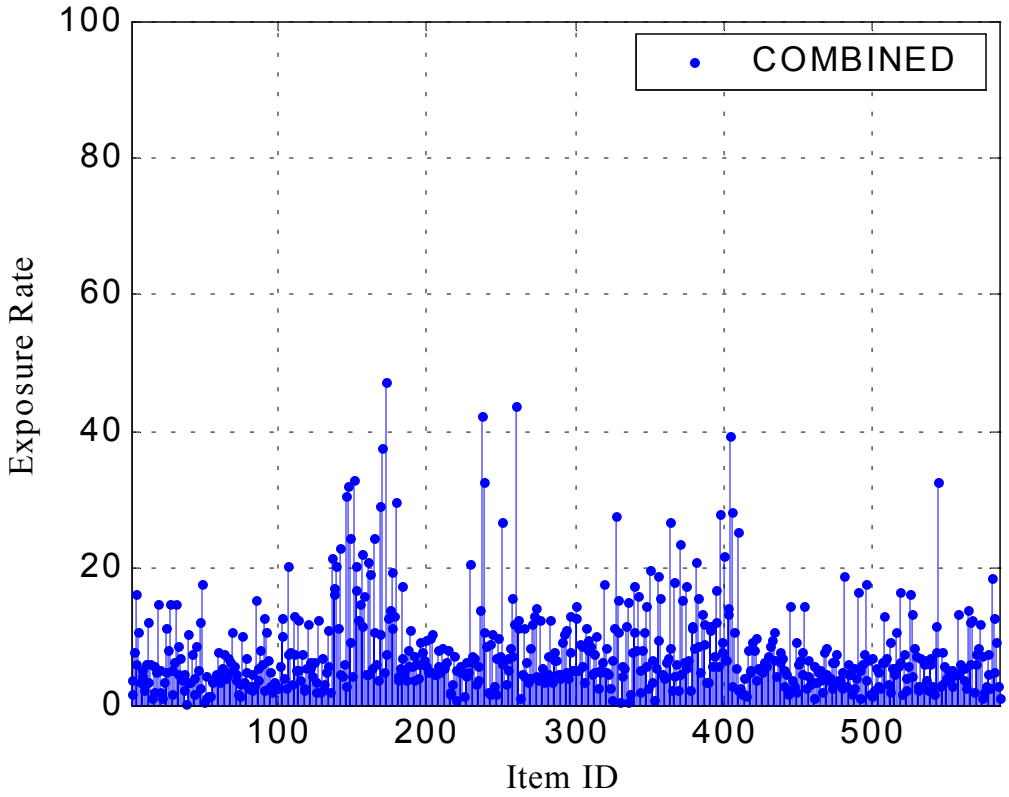

Figure 5e. Plot of Each Item's Exposure Rate for the COMBINED Method

2. Refining the COMBINED CAT Algorithm

In this study, an item-exposure control method (Sympson & Hetter, 1985) was not chosen and incorporated into this COMBINED CAT algorithm and consequently it turned out only about 1.7 % of items were overexposured when the maximum-allowed item-exposure rate was set at .30. This fact might indicate that without the involvement of an item-exposure control method, test security under this COMBINED algorithm still might not be seriously threatened because of its performance in maintaining a low overexposure rate.

On the other hand, if we intend to refine this algorithm to make each item's exposure rate as low as possible, the Sympson-Hetter method, which requires repeated simulation studies to set values of item-control parameters prior to the operational use of the test, could be included, but as a very low overexposured rate occurred here, this very time-consuming method might no longer be needed. The method of setting a maximum item-exposure rate during the on-line CAT environment might be practicable. A more comprehensive method for this approach can be found in van der Linden and Veldkamp's study (2002).

A less complex method, but supported empirically, is the restricted maximum exposure rate (RMER), proposed by Revuelta & Ponsoda (1998). This RMER method, like the Sympson-Hetter's method, sets the maximum-allowed rate, k, for a CAT algorithm. During the process of CAT, the actual item-exposure rate for each item is calculated as A_i/N, where A_i is the number of times the item has been administered and N is the number of examinees who have already taken CAT. This value was then compared with the pre-specified criterion, k. If this value is larger than k, the item was "temporary" removed form item bank and became an unavailable candidate to be selected. As the CAT continues, the number of examinee increases and the quotient of

A_i/N will decrease. The item will be again available because its exposure rate is again below k. When this RMER procedure is incorporated into the COMBINED CAT algorithm, those overexposured items (e.g., 1.7% in this study) should never take place with small enough k.

Reviewing CAT literature and the results generated from this research, we found that no matter what method we used, examinees with middle-level abilities were well estimated, even when test lengths were short. In addition, MIF made a dramatic improvement for high and low ability estimates. These findings might help us to refine the COMBINED CAT algorithm as designed in this study by the following modification. After the second-quarter test of CAT, if examinee's ability estimates fall within the middle range of θ (e.g. $-0.5 \le \theta \le 0.5$), the MOD method is used to the end of CAT; if they are high (e.g., $\theta > 2$) or low (e.g., $\theta < -2$), the MIF is used to end of CAT, and for the rest the original COMBINED method to the end of CAT. When this modified COMBINED CAT is implemented, direct item-exposure control methods such as the RMER procedure might not be needed to prevent any items from being overexposed.

3. Correlations between 3PL Item Parameters and CAT Algorithms

As stated in the introduction, the MIF method might tend to select highly discriminating items from the item pool in the constrained CAT. This can be confirmed as we computed the correlations between the 3PL item parameters and the item exposure rates for the five CAT algorithms, RANDOM, MIF, MID, MOD, and COMBINED. As seen in Table 5, we found that the correlation of item-exposure rate and the a_i parameters was .60 for the MIF CAT, .35 for the COMBINED CAT, and almost zero for the RANDOM, MIDO and MOD. This finding implies that the more highly discriminating the items are; the more likely they are to be administered in MIF CAT; in contrast, the two alternative methods, MID and MOD, did not follow this pattern.

The correlation between item-exposure rate and the b_i parameter was almost zero for the MIF method. The correlation for the item difficulty oriented method, MID, was .22, and the correlation was .10 for the MOD method. The slightly positive correlations for the b_i parameters for the MID and MOD methods may result from an interaction between the item-pool characteristics and the level of examinee's abilities. As seen in Table 1, the simulated item pool had more easy than difficult items (mean of difficulty = -.40), but we simulated a normally distributed ability construct in the group to take this item pool. When the MID was used in CAT, the items with item difficulties that are nearest to the examinees' abilities had more chance to be selected. As a result, harder items are, in general, preferable for this simulated examinee group. Here, the scale of item difficulty and ability parameter is the same. By the above logic, we might expect to obtain almost zero correlation between item-exposure rate and the b_i parameters for the MID or MOD method if we used another item pool whose average item difficulty is close to the average ability for the group of examinees.

Table 5
The Intercorrelations in Item Exposure Rates among 3PL Item Parameters as well as the Five CAT Methods

	A	b	c	RANDOM	MIF	MID	MOD	COMBINED
	1.00							
b	0.03	1.00						
c	0.19	0.04	1.00					
RANDOM	0.02	0.01	0.04	1.00				
MIF	0.60	0.06	-0.07	-0.02	1.00			
MID	0.04	0.22	0.06	0.04	0.20	1.00		
MOD	0.08	0.10	0.08	0.06	0.22	0.84	1.00	
COMBINED	0.35	0.12	-0.03	0.09	0.70	0.54	0.54	1.00

V. Conclusions

Key goals of CAT are to estimate each test-taker's ability efficiently and precisely. This seems to be best achieved by one of the family of maximum item information methods (e.g., MIF). Unfortunately, this type of item-selection method tends to select highly discriminating items from the beginning of CAT through to the end of testing. Negative consequences of that were discussed in the introduction and consequently interest in other promising item-selection algorithms is growing.

The family of maximum item information methods can be classified as item-discrimination-oriented methods. This study also explored two non-item-discrimination-oriented methods (e.g., MID and MOD) which have not often been evaluated in CAT studies. These two item-selection methods can be classified as item-difficulty-oriented based on their nature of the way of selecting items.

Based on the literature and the results of this study, these two methods do not outperform MIF in terms of accuracy of ability estimates, but both methods, especially MOD, were capable of producing comparably accurate results with MIF. When the factor of item-exposure rate was also taken into account, MIF might waste a large proportion (e.g. more than 50%) of the items in a pools and lead to overexposure of some items. Thus, the MOD method made use of more items and exhibited a more homogeneous item-exposure rate. These desirable features should make the MOD method more appealing in real CAT testing applications in the near future, especially if more studies using this promising item selection method are explored.

The ultimate goal of the current study was not to seek the best CAT algorithms in recovering ability estimates, but to find a compromise algorithm that is capable of increasing the homogeneity of CAT's item-exposure rates without significantly reducing the precision of ability estimates. Reviewing literature, a CAT using a modified shadow-test algorithm might be suitable to serve this purpose.

There are two key components for the typical shadow tests constraint. One is the item information function to be maximized; the other is the constraints imposed by users. At the beginning of the shadow-test CAT (e.g., the first two quarter tests), we are maximizing

something (e.g., information) that may not be as meaningful as we anticipate. We may be emphasizing information, technically by maximizing (or minimizing) other functions, while the on-line shadow test is assembled at the beginning of the CAT. A modification that allocates more emphasis to user constraints designed to minimize exposure may be possible. Such a modification might dramatically expand the use of all characteristics of items in the pool and guarantee that all constraints are fully met.

What other target (or objective) functions are legitimate (or appropriate) to be maximized (or minimized) in the CAT test? In this study, a combination of different CAT algorithms to be implemented at different stages of CAT was proposed. The Random method was used in the first-quarter stage of CAT. As explained previously, the poorest CAT algorithm, Random, does little harm in recovering ability at this early-stage CAT, but this method makes all items have almost equal likelihoods to be administered to any examinees. For the second and third quarter stages of CAT, better CAT item-selection methods, MID and MOD were used, respectively. Because the mechanism of each of these two methods does not depend of item parameter values and will make use of most items in the pool without significantly reducing the precision of ability estimates, these desirable features make them fit the middle-stages of CAT. At the final-stage (i.e., the fourth quarter of the test), test-takers' ability estimates are sufficiently accurate and stable for MIF to seek items with more information to improve upon test-takers' existing ability estimates.

Based on these results, the combined item-selection method under the shadow-test approach in CAT appears to have fully utilized all the items in the pool, reduced the maximum item exposure rates, generated more homogeneous exposure rates, and produced a very low rate of item overexposure (e.g., the proportion of items administered to 30% or more of the examinees was only 1.7%). Moreover, its accuracy in recovering ability estimates was similar to that of the MIF method. The COMBINED CAT method thus provided the best overall results. Some suggestions to refine this COMBINED method were provided in the previous section. More studies are needed to assess the impact of those modifications of the Combined CAT. Studies of how to implement the combined method for applications where the test length is variable are also needed.

References

Baker, F. B. (1992). Item response theory: Parameter estimation techniques. New York: Marcel Dekker, Inc.

Berger, M. P.F., & Veekamp, W.J. J. (1997). Some new item selection criteria for adaptive testing. Journal of Educational and Behavioral Statistics, 22, 203-226.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM. Psychometrika, 46, 443-459.

Bock, R. D., & Lieberman, M. (1970). Fitting a response model for n dichotomously scored items. Psychometrika, 35, 179-197.

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

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

Chen, S., Ankenmann, R. D., & Chang, H. H. (2000). A comparison of item selection rules at early stages of computerized adaptive testing.

Cheng, P. & Liou, M. (2000). Estimation of trait level in computerized adaptive testing. Applied Psychological Measurement, 24, 257-265.

CTBS/McGraw-Hill (1997). Teacher's guide to TerraNova. Monterey, CA. McGraw-Hill Companies, Inc.

De Ayala, R. J., Schafer, W., & Sava-Bolesta, M. (1995). An investigation of the standard errors for expected a posteriori ability estimates. British Journal of Mathematical and Statistical Psychology, 47, 385-405.

Dodd, B. G. (1990). The effect of item selection procedure and stepsize on computerized adaptive measurement using the rating scale model. Applied Psychological Measurement, 14, 355-366.

Hau, K., & Chang, H. H. (2001). Item selection in computerized adaptive testing: Should more discriminating items be used first? Journal of Educational Measurement, 38, 249-266.

Li, Y. H., & Schafer, W. D. (2003). The effect of item selection methods on the variability of CAT's ability estimates when item parameters are contaminated with measurement errors. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.

LINDO Systems, Inc. (2001). LINDO API: The premier optimization engine. [Computer program]. Chicago Illinois: INDO Systems, Inc.

Lord, F. M. (1980). Applications of item response theory to practical testing problems. Hillsdale NJ: Erlbaum.

Meijer, R. R., & Nering, M. L. (1999). Computerized adaptive testing: Overview and Introduction. Applied Psychological Measurement, 23, 187-194.

Mislevy, R. J., & Bock, R. D. (1990). BILOG-3 (2nd ed.): Item analysis and test scoring with binary logistic models. Mooresville, IN: Scientific Software.

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

Rao, P. S. (2000). Sampling methodologies with applications. New York: Chapman & Hall/CRC.

Sympson, J. B. & Hetter, R. D. (1985, October). Controlling item-exposure rates in computerized adaptive testing. Proceedings of the 27th annual meeting of the Military Testing Association (pp. 937-977). San Diego, CA: Navy Personnel Research and Development Center.

The MathWorks, Inc. (2001). MATLAB (Version 6.1): The language of technical computing [Computer program], Natick MA: The MathWorks, Inc.

Theunissen, J. J. J. M. (1985). Binary programming and test design. Psychometrika, 50, 411-420.

Theunissen, J. J. J. M. (1986). Optimization algorithms in test design. Applied Psychological Measurement, 10, 381-389.

van der Linden, W. J. (2000). Constrained adaptive testing with shadow tests. W. J. van der Linden and C. A. W. Glas (eds.), Computerized Adaptive Testing: Theory and practice, 27-52. Dordrecht, The Netherlands: Kluwer Academic Publishers.

van der Linden, W. J., & Boekkooi-Timminga, E. (1989). A maximin model for test design with practical constraints. Psychometrika, 54, 237-247.

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

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

van der Linden, W. J., & Veldkamp, B. P. (2002). Constraining item exposure in computerized adaptive testing with shadow tests (Research Report No. 02-06). University of Twente, The Netherlands.

Veerkamp, W., & Berger, M. (1999). Optimal item discrimination and maximum information for logistic IRT models. Applied Psychological Measurement, 23, 31-40.

Wang, T., Hanson, B. A., & Lau, C. A. (1999). Reducing bias in CAT trait estimation: A comparison of Approaches. Applied Psychological Measurement, 23, 263-278.

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. Psychometrika, 4, 427-450.

Weiss, D. J., & McBride, J. R. (1984). Bias and information of Bayesian adaptive testing. Applied Psychological Measurement, 8, 273-285.