LSAC RESEARCH REPORT SERIES

Constraining Item Exposure in Computerized Adaptive Testing with Shadow Tests

Wim J. van der Linden
Bernard P. Veldkamp
University of Twente, Enschede, The Netherlands

Law School Admission Council
Computerized Testing Report 02-03
December 2005

A Publication of the Law School Admission Council

The Law School Admission Council (LSAC) is a nonprofit corporation whose members are more than 200 law schools in the United States and Canada. It was founded in 1947 to coordinate, facilitate, and enhance the law school admission process. The organization also provides programs and services related to legal education. All law schools approved by the American Bar Association (ABA) are LSAC members. Canadian law schools recognized by a provincial or territorial law society or government agency are also included in the voting membership of the Council.

This study is published and distributed by the Law School Admission Council (LSAC). The opinions and conclusions contained in these reports are those of the authors and do not necessarily reflect the position or policy of the Law School Admission Council.

Table of Contents

Executive Summary

A potential danger in computerized adaptive testing (CAT) is that the item selection algorithm may overuse the items with the most attractive statistical characteristics, and the security of these items may be thereby compromised. Therefore, item exposure constraints that assure that none of the items in an item pool are overexposed are an important component of a CAT item selection algorithm. In this paper an alternative to the currently widely applied Sympson-Hetter item exposure control method is presented, which is based on decisions about the eligibility of the items in the pool before the test taker takes the test. If an item is eligible it remains in the pool; if it is ineligible it is removed from the pool for the test taker. These decisions are based on the outcomes of a probabilistic experiment with probabilities of eligibility that constrain the item-exposure rates to be below the target value.

The main advantage of the method is that, unlike the Sympson-Hetter method, no time-consuming simulation studies are necessary to find admissible values for control parameters of items. Instead, the current method is self-adaptive and can be implemented "on the fly" during operational testing. The method counts certain events during the testing of the test takers and uses these counts to automatically adapt the probabilities of item eligibility to their optimal level, which is then maintained during the rest of the testing process.

An extensive simulation study with adaptive tests from a previous pool of LSAT items showed that the probabilities of item eligibility were already stable after 1,000 test takers were tested, and the method produced exposure rates that were below the target for all items.

Abstract

Item-exposure control in computerized adaptive testing is implemented by imposing item-ineligibility constraints on the assembly process of the shadow tests. The method resembles Sympson and Hetter's (1985) method of item-exposure control in that the decisions to impose the constraints are probabilistic. However, the method does not require time-consuming simulation studies to set values for control parameters prior to the operational use of the test. Instead, the probabilities of item ineligibility can be set on the fly using an adaptive procedure based on the actual item-exposure rates. An empirical study using an item pool from the Law School Admission Test (LSAT) showed that application of the method yielded perfect control of the item exposure rates and had negligible impact on the bias and mean-squared error functions of the ability estimator.

Introduction

Computerized adaptive testing (CAT) can be viewed as an item-selection process in which an objective function is optimized subject to several constraints. A popular objective function in CAT is Fisher's information measure at the estimated ability level of the examinee (for alternative functions, see van der Linden & Pashley, 2000). Constraints have to be imposed on the item-selection process to give the test the necessary composition with respect to such attributes as item content, format, gender orientation, word counts, and statistical item parameters. It is the purpose of this paper to show that the set of constraints can also be extended with constraints that control the exposure rates of the items in the test.

The optimization problem involved in CAT can be represented by a test assembly model with 0-1 decision variables for the items. Let $I = 1, ..., I$ denote the items in the pool. Decision variable x_i is defined to take the value 1 if this item is selected in the test and the value 0 if item I remains in the pool. Suppose that the objective is to maximize Fisher's information measure, which takes the value $I_i(\theta)$ for item I at ability level θ. In addition, suppose that the constraints have to control the composition of the tests with respect to a set of categorical attributes (e.g., item content and format) and quantitative attributes (e.g., statistical item parameters and word counts). The categorical attributes partition the item pool into subsets of items with the same value on the attribute, whereas the quantitative attributes have numerical values. We will use V_g, $g = 1, ..., G$, as a generic symbol to denote subsets of items that share a categorical attribute and q_i as a generic symbol to denote the value of item I on a quantitative attribute.

A general representation of the optimization problem in CAT is:

$$\text{maximize} \sum_{i=1}^{I} I_i(\theta) x_i \tag{1}$$

subject to

$$\sum_{i \in V_g} x_i \leq u_g, g = 1, \ldots, G,$$

$$\sum_{i \in V_g} x_i \geq l_g, g = 1, \ldots, G,$$

$$\sum_{i=1}^{I} q_i x_i \leq u_q,$$

$$\sum_{i=1}^{I} q_i x_i \geq l_q,$$

$$\sum_{i=1}^{I} x_i = n,$$

$$x_i \in \{0, 1\}, i = 1, \ldots, I. \quad (2)$$

The objective function maximizes the test information at the examinee's true ability value. The first two sets of constraints require the number of items from the set V_g to be between upper bound u_g and lower bound l_g. Likewise, the next two constraints require the sum of values of attribute q_i to be between upper and lower bound u_q and l_q. The final two constraints define the length of the test and the range of the decision variables. For examples of more specific types of constraints that may be needed in test assembly models for CAT, see van der Linden (1998, 2000).

Shadow Test Approach

If all items in the test were to be selected simultaneously and θ could be set to a known value, the model in Equations 1 and 2 could be solved immediately for an optimal set of values for the decision variables x_i, $I = 1, ..., I$, using a general-purpose software package for 0-1 linear programming (LP), for example, the software package CPLEX 6.6 (ILOG, Inc., 2000). These values tell us which set of items constitutes the optimal test. However, in adaptive testing items are selected one at a time and each next item has to be optimal at an estimate of θ that changes during the test.

An effective way to solve the assembly model for adaptive tests is through a shadow test approach (van der Linden, 2000; van der Linden & Reese, 1998). In this approach, prior to the selection of each item, the model in Equations 1 and 2 is solved for a full-size linear test (= shadow test) at the current ability estimate with the decision variables of the items already administered fixed to the value 1. The CAT algorithm then picks the optimal item at the ability estimate from the free items in the shadow test for administration. Because each shadow test meets the constraints, the adaptive test automatically meets them. Likewise, because each shadow test is optimal at its ability estimate and the optimal item at the estimate is administered, the adaptive test is optimal.

Online assembly of shadow tests has become possible through recent developments in 0-1 LP. The techniques in this area have become so fast that it takes less than a second for a current PC to assemble a test from an item pool of the size typically used in CAT (Veldkamp, 2001). However, the shadow test approach is a general scheme for optimizing the selection of items in CAT; it can also be used in combination with other techniques for optimal test assembly, such as the algorithms proposed by Leucht (1998) and Swanson and Stocking (1993).

Number of Constraints and Feasibility

For a typical test assembly problem, the number of constraints in the model needed to represent all test specifications can easily run into the hundreds. A set of values for the decision variables that meets all constraints is known as a *feasible solution*. Algorithms in 0-1 LP algorithms search for a set of values that has the best value for the objective function among the feasible solutions. This solution is known as an *optimal feasible solution*.

In principle, LP algorithms are able to solve problems with unlimited numbers of constraints. However, a possible effect of adding a constraint to the model is a decrease in the value of the objective function for the optimal solution. The tighter the constraints, the larger the effect can be. It is even possible to overconstrain the model, in which case no feasible solution is left, and the LP problem is said to be infeasible.

Feasibility of an LP model is a property that depends only on the constraints. In an adaptive test, the only quantity in the model for the shadow test that changes during the test is the objective function. Therefore, if the model for the shadow tests in Equations 1 and 2 is feasible for any examinee, it is feasible for all examinees. Also, feasibility is maintained during the adaptive test because the items for which the decision variables are fixed at $x_i = 1$ were part of a previous solution. In practical applications, feasibility of test assembly problems is never a problem. If it were, the event should be taken to signal a flaw in the composition of the item pool relative to the specifications of the test (van der Linden, Veldkamp, & Reese, 2000; Veldkamp & van der Linden, 2000).

Item-Exposure Control

Item-exposure control in CAT is necessary to prevent possible item compromise due to overexposure of items. Methods for item-exposure control based on Sympson and Hetter's (1985) proposal have become the de facto standard of the CAT industry. These methods are based on the definitions of two events: (1) item I is selected by the CAT algorithm (S_i) and (2) item I is administered (A_i). Because

$$A_i \subset S_i, \tag{3}$$

it holds that

$$P(A_i) = P(A_i \mid S_i)P(S_i). \tag{4}$$

The probabilities $P(S_i)$ are determined by the composition of the item pool and the nature of the CAT algorithm. However, the conditional probabilities $P(A_i \mid S_i)$ are control parameters that can be set to guarantee that

$$P(A_i) \leq r_{\max}, i = 1, \ldots, I, \tag{5}$$

with $r_{\max}$ being a target value determined by the testing agency. In the Sympson-Hetter method, after an item is selected, the control parameters, $P(A_i \mid S_i)$ are implemented through a probability experiment conducted to determine if the item is actually administered. If an item is not administered, it is removed from the pool, and the experiment is repeated for the next best item.

Admissible values for the control parameters in the Sympson-Hetter method cannot be calculated analytically. Instead, they have to be found through a series of iterative adjustments, with each iteration step being based on a sufficiently large number of simulated administrations of the adaptive test. Though it is possible to modify the Sympson-Hetter method to speed up the iterative adjustment process considerably (van der Linden, 2002), the process is generally tedious and time-consuming, particularly if the control parameters have to be set conditional on a set of realistic ability values for the population of examinees.

For empirical studies on the Sympson-Hetter method or other implementations of the idea to determine item administration probabilistically, see Davey and Parshall (1995, April), Eggen (2001), McBride and Martin (1983), Parshall, Hogarty, and Kromrey (1999, June), Revuelta and Ponsoda (1998), van der Linden (2002), and van der Linden and Veldkamp (2002).

Constraining Item Exposure Rates

It is the purpose of this paper to present an alternative method of item-exposure control that is not based on the idea of controlling for item exposure after an item is selected but *before* an examinee takes the test. That is, the decisions to be made are no longer if an item that is selected should be administered, but which of the items in the pool are eligible for the examinee. If an item is eligible, it remains in the pool; if it is ineligible, it is removed from the pool for the examinee. The idea underlying the method in this paper is to base the decisions on the outcomes of a probabilistic experiment with probabilities of eligibility that constrain the item-exposure rates to be below the target value.

The main advantage of the method is that no time-consuming simulation studies are necessary to find admissible values for control parameters of items. Instead, it is possible to implement the method on the fly during operational testing. The method automatically adapts the probabilities of item eligibility to their optimal level and maintains these during the rest of the testing process.

Ineligibility Constraints

A natural way to implement control of item eligibility in adaptive testing with shadow tests is through the inclusion of constraints in the assembly model in Equations 1 and 2. If the decision is made that item I is ineligible for the current examinee, the following constraint is added to the model for the shadow tests:

$$x_i = 0. \tag{6}$$

If item I remains eligible, no constraint is added. We will refer to the constraints in Equation 6 as *ineligibility constraints*.

As noted earlier, a potential effect of adding constraints to a test assembly model is infeasibility of the model. In principle, infeasibility may thus occur if a large number of ineligibility constraints is added to the model. However, for a real-life CAT program, the probability of running into infeasibility due to item ineligibility constraints is generally small. The reason is that the method in this paper imposes ineligibility constraints only for items that have a tendency to be overexposed. It is a common experience in adaptive testing that this number is low. In fact, the majority of the items in real-life CAT programs are hardly exposed at all.

If the addition of ineligibility constraints results in an infeasible model for the shadow tests, a simple measure to restore feasibility is to remove all ineligibility constraints from the model and solve the relaxed model. The amount of time needed to detect infeasibility of the model is negligible. Besides, as will be shown below, the adaptive nature of the probabilistic experiment in this paper results in an automatic correction for an occasional extra exposure of an item due to infeasibility later in the testing process.

Probabilistic Constraints on Item Eligibility

The process involved in selecting a test for an examinee in adaptive testing with shadow tests is as follows: First, for each item in the pool, a probability experiment is conducted to determine which items are eligible and which are ineligible; second, ineligibility constraints are added to the model for the shadow tests, and the model is solved for optimal shadow tests at the sequence of updated ability estimates during the test. Third, if the addition of the ineligibility constraints leads to an infeasible model, the constraints are removed from the model, and the relaxed models for the shadow tests are solved. Fourth, from each shadow test, the (free) item with maximum information at the ability estimate is administered.

In summary, the process of item administration is thus defined by the following events:

E_i: item I is determined to be eligible for the examinee by the probability experiment;

F: the model for the assembly of the shadow test for the examinee remains feasible after the eligibility constraints have been added;

S_i: item I is selected in a shadow test for the examinee;

A_i: item I is administered to the examinee.

The event of removing the ineligibility constraints from the model is equivalent to $\overline{F}$. Observe that S_i is now used to denote selection of the item in a shadow test and not selection for administration.

Analogous to Equations 3 and 4, it now holds that

$$A_i \subset S_i \subset \left\{E_i \cup \overline{F}\right\}, \tag{7}$$

and

$$P(A_i) = P(A_i \mid S_i)P(S_i \mid E_i \cup \overline{F})P(E_i \cup \overline{F}). \tag{8}$$

However, because

$$P(A_i \mid S_i)P(S_i \mid E_i \cup \overline{F}) = P(A_i \mid E_i \cup \overline{F}),$$

we may ignore event S_i and Equation 8 reduces to

$$P(A_i) = P(A_i \mid E_i \cup \overline{F})P(E_i \cup \overline{F}). \tag{9}$$

The problem is to select values for the probabilities of eligibility, $P(E_i)$, such that the exposure rate for item I is below r. Combining Equation 5 with Equation 9, it follows that the target $r_{\max}$ for the item exposure rate for item I is met if:

$$P(E_i \cup \overline{F}) \leq \frac{r_{\max}}{P(A_i \mid E_i \cup \overline{F})}. \tag{10}$$

However, this inequality does not impose any direct constraint on $P(E_i)$. We therefore make the following independence assumption

$$P(E_i \cap F) = P(E_i)P(F). \tag{11}$$

In addition, from Equation 7

$$P(A_i \cap (E_i \cup \overline{F}) = P(A_i).$$

Using

$$P(E_i \cup \overline{F}) = 1 - P(\overline{E}_i \cap F), \tag{12}$$

and

$$P(\overline{E}_i) = 1 - P(E_i),$$

it follows that Equation 10 holds if:

$$P(E_i) \leq 1 - \frac{1}{P(F)} + \frac{r_{\max}P(E_i \cup \overline{F})}{P(A_i)P(F)}, P(A_i) > 0, P(F) > 0. \tag{13}$$

Because we want the exposure rates of the items with a tendency to overexposure to be just below the target in Equation 5, the probabilities of item eligibility should be set at the upper bound in Equation 13.

Observe that these probabilities are all marginal with respect to the distribution of θ. To obtain constraints that impose item exposure control conditional on θ (Stocking & Lewis, 1998), all probabilities should be replaced by conditional probabilities given θ. This operation is straightforward, and its results are not presented here. We will return to the issue of conditional item-exposure control later in this paper.

Adapting the Probabilities of Item Eligibility

The idea is to set this upper bound adaptively during the test. That is, after an examinee is tested, the probabilities of the events in the right-hand side of Equation 13 are updated using the events recorded for the examinee. More specifically, assume that j examinees have been tested and $j + 1$ is the next examinee. The probabilities are then updated by

$$P^{(j+1)}(E_i) = \min\left\{1 - \frac{1}{P^{(j)}(F)} + \frac{r_{\max} P^{(j)}(E_i \cup \bar{F})}{P^{(j)}(A_i) P^{(j)}(F)}, 1\right\}, \tag{14}$$

provided $P^{(j)}(A_i) > 0$ and $P^{(j)}(F) > 0$.

The following argument helps to understand Equation 14. As already noted, for a well-designed real-life CAT program, the probability of infeasibility is small. That is, we expect the testing procedure to approximate

$$P^{(j)}(F) = 1 \tag{15}$$

and

$$P^{(j)}(E_i \cup \bar{F}) = P^{(j)}(E_i). \tag{16}$$

Under these conditions, it follows from Equation 14 that

$$\begin{aligned} &P^{(j+1)}(E_i) < P^{(j)}(E_i) \text{ if } P^{(j)}(A_i) > r_{\max}, \\ &P^{(j+1)}(E_i) = P^{(j)}(E_i) \text{ if } P^{(j)}(A_i) = r_{\max}, \\ &P^{(j+1)}(E_i) > P^{(j)}(E_i) \text{ if } P^{(j)}(A_i) < r_{\max}. \end{aligned} \tag{17}$$

These relations show the behavior we want the probabilities of item eligibility to have: If the probability of exposure for an item is too high at some stage during the testing process, its probability of eligibility should go down. If the probability of exposure for an item is already below the target, the probability of eligibility should be relaxed. If the conditions in Equations 15 and 16 do not hold, the relation between $P^{(j+1)}(E_i)$ and $P^{(j)}(E_i)$ is slightly more complicated, because the probability of feasibility intervenes and the actual probability of item I being available for selection for examinee j is not $P^{(j)}(E_i)$ but $P^{(j)}(E_i \cup \bar{F})$.

The adaptive nature of the update in Equation 14 also explains our earlier claim that an occasional extra exposure of some of the items in the pool due to the addition of ineligibility constraints to the model for the shadow tests is automatically corrected later in the testing process. Using Equation 12, the expression in the right side of Equation 14 can be written as

$$1 - \frac{1}{P^{(j)}(F)}\left(1 - \frac{r_{\max}}{P^{(j)}(A_i)}\right) - \frac{r_{\max} P^{(j)}(\bar{E}_i)}{P^{(j)}(A_i)}. \tag{18}$$

If infeasibility occurs for examinee j, $P^{(j)}(F)$ decreases. The effect of this decrease is an increase in the probabilities of eligibility, $P^{(j+1)}(E_i)$, in Equation 14, provided $P^{(j)}(A_i) < r_{\max}$.

Independence Assumption

This assumption of independence in Equation 11 is the only empirical assumption on which the method is based. The assumption is generally realistic. Unless the item pool is badly designed, it typically has multiple sets of items with the combinations of attributes required by the content constraints in the test assembly model. The probability of a feasible solution, therefore, does not depend on the ineligibility of a small number of the items in the pool.

The assumption can easily be tested for an operational CAT program by a scatter plot for the items in the pool with estimates of the probabilities $P(E_i \cap F)$ against products of the estimates of $P(E_i)$ and $P(F)$. Deviations from the identity line would point at violations of the assumption of independence. If some of these probabilities are extremely low, we could plot estimates of $\ln P(E_i \cap F)$ against $\ln P(E_i)$ and check for nonlinearity.

Implementing the Method

Suppose that for the previous j examinees we have recorded the following counts:

φ_j: number of examinees for which the shadow test has been feasible (event F);

ρ_{ij}: number of examinees for which item I has been eligible or the ineligibility constraints have been removed after infeasibility (event $E_i \cup \bar{F}$);

α_{ij}: number of examinees to whom item I has been administered (event A_i).

Thus, for examinee j + l, item I is eligible with estimated probability

$$\widehat{P^{(j+1)}(E_i)} = \min\left\{1 - \frac{j}{\varphi_j} + \frac{j\rho_{ij} r_{\max}}{\alpha_{ij}\varphi_j}, 1\right\}, \tag{19}$$

with $\alpha_{ij} > 0$ and $\varphi_j > 0$. The decision to add an ineligibility constraint for item I to the model for the shadow test is based on a probability experiment with this probability.

Because of the adaptive nature of the method, the method can be applied directly to an operational test. The only conditions that should be avoided for the estimates in Equation 19 are $\alpha_{ij} = 0$ and $\varphi_j = 0$. A simple way to avoid these conditions is to initialize $\widehat{P^{(j+1)}(E_i)}$ at 1 and keep it at this value as long as the counts α_{ij} and φ_j are equal to zero. This measure was taken in the empirical examples below.

Stabilization of Probability Estimates

If the method is applied directly to an operational test, the relative counts in Equation 19 stabilize if the number of examinees tested increases. As long as these counts are unstable, some of the estimates of the probabilities of eligibility may fluctuate considerably. This behavior is particularly likely for items that are favored by the CAT algorithm (e.g., items with high values for the discrimination parameter and values for the difficulty parameter near the center of the ability distribution). These items will be the first to become ineligible. Also, as soon as they are eligible again, they tend to be administered. If the relative counts become stable, the opposite phenomenon will be observed. The probability estimates in Equation 19 then hardly change, and items will remain eligible or ineligible for long periods.

If a more consistent behavior of the estimates of the probabilities of item eligibility during the testing process is desirable, they should be updated using the technique of fading introduced in the literature on Bayesian networks (Jensen, 2001, sect. 3.3.2). In this technique counts of events are updated by combining new data with old data weighted by a fading factor chosen from the interval (0,1). Let ω_i be the choice for the factor for item I. A common factor ω for all items will suffice in most applications. If j examinees have been tested, the weighted updates of the count of the events A_i is given by the following relation:

$$\alpha^*_{i(j+1)} = \begin{cases} \omega_i \alpha^*_{ij} + 1 & \text{if } i \text{ was administered to } j, \\ \omega_i \alpha^*_{ij} & \text{if } i \text{ was not administered to } j, \end{cases} \tag{20}$$

where the asterisk is used to denote a weighted version of the counts. The updates of the weighted counts φ^*_j and ρ^*_{ij} are analogous, whereas the weighted count of the number of examinees is updated by

$$j^*_{i(j+1)} = \omega_i j^*_{ij} + 1. \tag{21}$$

If these weighted counts are substituted into Equation 19, the impact of old events fades away during the testing process. In fact, the choice of weight ω_i implies an effective sample size that tends to $1/(1-\omega_i)$ (Jensen, 2001). As a consequence, the estimates of the probabilities of item eligibility tend to be based permanently on the last $1/(1-\omega_i)$ examinees.

If operational testing should begin with stable values for the probabilities in Equation 19, this goal can be realized by simulating the adaptive test on the computer with weighted updates of the counts for $1/(1-\omega_i)$ examinees prior to the start of the testing process. In the empirical example below, we used weights $\omega_i = .999$, which amounted to an effective sample size of 1,000 examinees.

Item Sets

In adaptive testing, item pools sometimes have a structure with sets of items organized around common stimuli (e.g., text passages or descriptions of cases). It is possible to apply the techniques of item-exposure control in this paper at the level of individual items within sets. However, in tests with item sets, the concern typically is about overexposure of stimuli rather than items. It is therefore recommended to apply the techniques in this paper at the level of the stimuli. Because the exposure rates of items in sets can never be larger than the rates of their stimuli, the criterion in Equation 5 is automatically satisfied for the items if it is for the stimuli.

Empirical Examples

A simulation study was conducted to check the efficacy of the method of exposure control presented in this paper and estimate its effects on the statistical properties of the ability estimator. The following conditions were compared:

1. CAT without item-exposure control;
2. CAT with item-exposure control, with the probabilities of item eligibility updated without fading; and
3. CAT with item-exposure control, with the probabilities of item eligibility updated with fading.

The condition without item-exposure control was included to provide a base line for the evaluation of the performances of the exposure control methods in the other conditions. For the two conditions with exposure control, the target for the maximum exposure rate was set at $r = .25$, a choice believed to be typical of actual targets in real-life CAT programs. For the condition with fading, factor ω_i in Equations 20 and 21 was set equal to .999.

Item Pool and Test

The item pool was a previous 753-item pool from the Law School Admission Test (LSAT). All items fit the 3-parameter logistic item response model (Hambleton & Swaminathan, 1985). The pool had both discrete items and items organized as sets around a common stimulus. The exposure of the items in the sets was at the level of the stimuli. The total number of discrete items and stimuli in the pool was 353.

A 50-item adaptive version of the LSAT was simulated. The length of this test was half the length of the paper-and-pencil version of the LSAT; all specifications for the LSAT were therefore scaled back to 50%. The LSAT consists of three different sections, two of the sections having an item-set structure. The total number of constraints in the LP model for the shadow tests needed to represent all test specifications with respect to such attributes as item and stimulus content, (sub)types, gender and minority orientation, answer key, and word count was 433.

CAT Algorithm

Test administrations were simulated for examinees randomly sampled from a uniform distribution over $\theta = -2.0, -1.5, \ldots, 2.0$. For the two conditions with the method presented in this paper, we used l, 000 replications to study stabilization of the probability updates, and then another 9,000 replications to get the results presented below. Sampling from a uniform distribution of θ was chosen to enable the estimation of the bias and mean-squared error (MSE) functions of the estimator of θ with equal precision along its scale (1,000 replications for each θ value). The items were selected using the maximum-information criterion. The estimator of θ was the expected a posteriori (EAP) estimator with an uninformative prior over [-4,4]. In all conditions, the ability estimate was initialized at $\hat{\theta} = 0$.

Results

In both conditions with exposure control, the models for the shadow test completed with the ineligibility constraints always had a feasible solution. Thus, the independence assumption in Equation 11 was automatically met for all items.

Figure 1 shows the exposure rates of the discrete items and stimuli in the pool for the three conditions after 10,000 examinees. For the condition without exposure, 38 of the items and stimuli had exposure rates that seriously violated the target rate of .25. However, both conditions with exposure control showed perfect exposure rates. Except for a few random violations (smaller than .01), all items had exposure rates below the target value of .25.

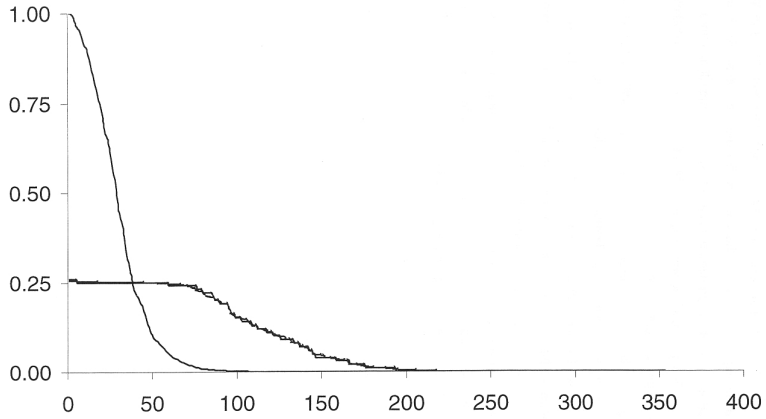

FIGURE 1. *Distribution of exposure rates for the items and stimuli in the pool for CAT without exposure control (solid line), with exposure control (dashed line), and with exposure control and probability updates based on fading (dotted line)*

Observe that the distributions of exposure rates for the two conditions with exposure control are indistinguishable. Application of the technique of fading thus had no impact whatsoever on the distribution of the exposure rates. Also, from this figure it is obvious that the control of the exposure rates of the 38 items and stimuli with a tendency to overexposure led to an increase in the exposure rates of several of the other items and stimuli. This result is to be expected, because the average exposure rate in the pool is always equal to nI^{-1}, where n is the length of the test and I is the size of the pool (van der Linden, 2002, proposition 1).

To find out if the exposure rates for the method in this paper stabilized quickly, we checked the rates of the items and stimuli after 1,000 examinees. Figures 2 and 3 show the distributions of the differences between the actual rates after 1,000 and 10,000 examinees for the condition without and with probability updates with fading, respectively. For both conditions, nearly all differences were smaller than 1%, indicating that the method was already stable after 1,000 examinees.

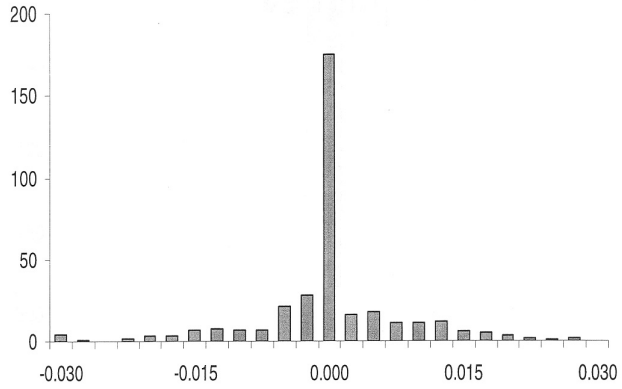

FIGURE 2. *Differences between exposure rates after 1,000 and 10,000 examinees for the CAT with exposure control*

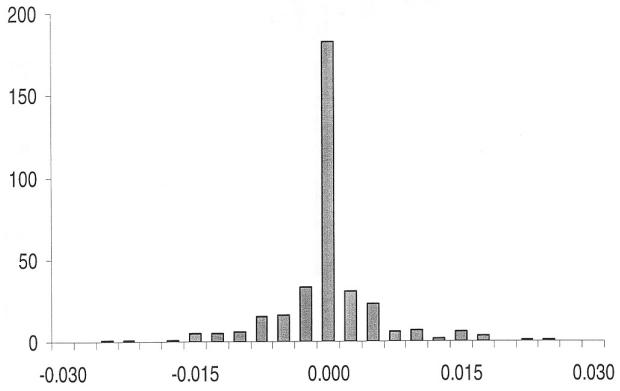

FIGURE 3. *Differences between exposure rates after 1,000 and 10,000 examinees for the CAT with exposure control and probability updates based on fading*

The bias and MSE functions for the ability estimator estimated from 10,000 simulated examinees are given in Figures 4 and 5, respectively. The differences between these two functions for the three conditions are negligible for all practical purposes. The method of item exposure control in this paper thus had negligible impact on the statistical quality of the ability estimator.

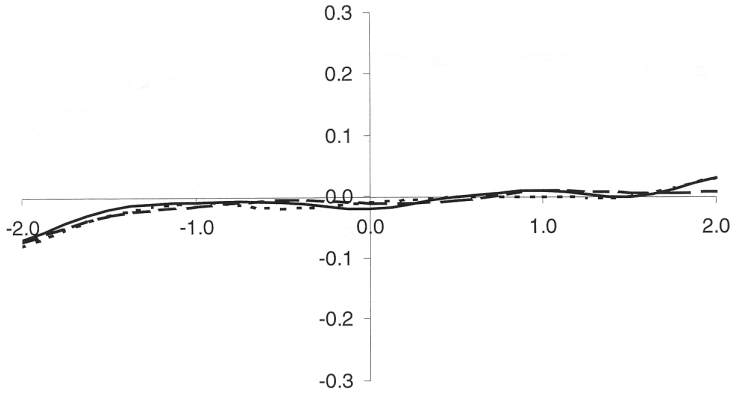

FIGURE 4. *Estimated bias functions of the ability estimator for CAT without exposure control (solid line), with exposure control (dashed line), and with exposure control and probability updates based on fading (dotted line)*

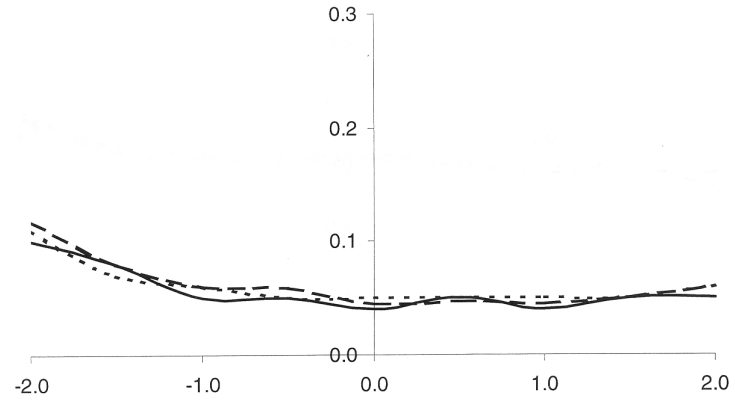

FIGURE 5. *Estimated MSE functions of the ability estimator for CAT without exposure control (solid line), with exposure control (dashed line), and with exposure control and probability updates based on fading (dotted line)*

From these results it seems safe to conclude that the method works well. The distribution of the actual exposure rates was exactly as desired and already stable after 1,000 examinees. The fact that the impact of the methods on the statistical properties of the ability estimator was negligible is easily explained by the model for the assembly of the shadow tests in Equations 1 and 2. The MSE of the ability estimator depends directly on the value of the objective function for the optimal shadow test. This value, in turn, depends on the number and severeness of the constraints. For real-life CAT programs, the number of ineligibility constraints that have to be added to the model is expected to be small relative to the number of content constraints. For example, for the LSAT the number of content constraints was 433, whereas we only had to constrain the exposure rates of 38 items to a value below .25. The average exposure rate of these items in the condition without exposure control was .70, which implies that on average for each examinee some 17 ineligibility constraints had to be added to the model. This implied an increase in the number of constraints by only 3.9%. If the item pool is well designed, ineligibility constraints are not severe and their impact on the value of the objective function for the solution is negligible.

Discussion

A practical advantage of the method of exposure control introduced in this paper is that, unlike the Sympson-Hetter method, it does not need a long iterative process of setting values for control parameters. The adaptive nature of the method guarantees that the exposure rates are automatically set, and the method can be applied directly to operational tests.

The adaptive nature of the method also opens up a whole new range of strategies of item pool management. For example, if one or more items are detected to be flawed or to have been victims of security breaches, these items can simply be removed from the pool or replaced by other items; the exposure rates for the items in the new pool adapt automatically to below the target value. For the Sympson-Hetter method, the pool would have to be taken out of operation to have new values set for its control parameters (Chang & Harris, 2002).

We could also use such replacements intentionally, periodically replacing its parts of the item pool for which the absolute numbers of exposures have reached a predetermined level. This strategy is an alternative to item-exposure control strategies based on random rotation of multiple item pools (Mills & Steffen, 2000). As a matter of fact, it can be seen as a rotation method that permanently rotates different item pools among the examinees.

The rotation scheme does not require any physical transport of items. Neither does it require the assembly of multiple item pools with prior estimation of required overlap rates (Stocking & Swanson, 1998). Also, the size of the rotating pools as well as their item overlap rates are automatically set at optimal levels.

As a final example of a new strategy of item pool management, observe that there is no need to set the maximum exposure rate $r_{\max}$ at a common value for all items and examinees. This target could be set at different values for different sets of items, for example, at lower values for items or stimuli that are easier to memorize or for item pools that are used at locations with a higher risk of item compromise. Also, the target could be set at different levels during the history of an individual item pool. These and other strategies for optimizing item pool management deserve further investigation.

As already noted, the method presented in this paper can also be applied conditional on a series of ability levels. Formally, the only thing required to get constraints on conditional item-exposure rates is to formulate all probabilities conditional on θ. Practically, in a conditional version of the method, the estimates of the conditional probabilities of item eligibility in (19) are still updated after the examinee has completed the test, but the eligibility experiments with these probabilities are conducted more than once for each examinee, namely at each θ value at which the item-exposure rates are to be controlled. Further research involves a study, analogous to that carried out for the Sympson-Hetter method (Stocking & Lewis, 2000), to assess the possible impact during the test of estimation error in the ability estimate on actual conditional item-exposure rates.

References

Chang, S. W., & Harris, D. J. (2002, April). *Redeveloping the exposure control parameters of CAT items when a pool is modified*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Davey, T., & Parshall, C. G. (1995, April). *New algorithms for item selection and exposure control with computerized adaptive tests*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Eggen, T. J. H. M. (2001). *Overexposure and underexposure of items in computerized adaptive testing* (Measurement and Research Department Reports 2001-1). Arnhem, The Netherlands: Citogroep.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston: Kluwer-Nijhoff.

ILOG, Inc. (2000). CPLEX 6.6 [Computer program and manual]. Incline Village, NV: Author.

Jensen, F. V. (2001). *Bayesian networks and graphs*. New York: Springer.

Leucht, R. M. (1998). Computer-assisted test assembly using optimization heuristics. *Applied Psychological Measurement, 22*, 224–236.

McBride, J. R., & Martin, J. T. (1983). Reliability and validity of adaptive ability tests in a military setting. In D. J. Weiss (ed.), *New horizons in testing: Latent trait theory and computerized adaptive testing* (pp. 223–226). New York: Academic Press.

Mills, G. N., & Steffen, M. (2000). The GRE adaptive test: Operational issues. In W. J. van der Linden & C. A. W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 75–99). Boston: Kluwer.

Parshall, C., Hogarty, K., & Kromrey, J. (1999, June). *Item exposure in adaptive tests: An empirical investigation of control strategies*. Paper presented at the annual meeting of the Psychometric Society, Lawrence, KS.

Revuelta, J., & Ponsoda, V. (1998). A comparison of item-exposure control methods in computerized adaptive testing. *Journal of Educational Measurement, 38*, 311–327.

Stocking, M. L., & Lewis, C. (1998). Controlling item exposure conditional on ability in computerized adaptive testing. *Journal of Educational and Behavioral Statistics, 23*, 57–75.

Stocking, M. L., & Lewis, C. (2000). Methods of controlling the exposure of items in CAT In W J. van der Linden & C. A. W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 163–182). Boston: Kluwer.

Stocking, M. L., & Swanson, D. (1998). Optimal design of item banks for computerized adaptive testing. *Applied Psychological Measurement, 22*, 271–279.

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

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. 973–977). San Diego, CA: Navy Personnel Research and Development Center.

van der Linden, W. J. (Ed.) (1998). Optimal test assembly of psychological and educational tests. *Applied Psychological Measurement, 22*, 195–211.

van der Linden, W. J. (2000). Constrained adaptive testing with shadow tests. In W. J. van der Linden & C. A. W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 27–52). Boston: Kluwer.

van der Linden, W. J. (2002). *A formal characterization and some alternatives to Sympson-Hetter item-exposure control in computerized adaptive testing*. Manuscript submitted for publication.

van der Linden, W. J., & Pashley, P. J. (2000). Item selection and ability estimation in adaptive testing. In W. J. van der Linden & C. A. W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 1–25). Boston: Kluwer.

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). *Implementing Sympson-Hetter item-exposure control in adaptive testing with shadow tests*. Manuscript submitted for publication.

van der Linden, W. J., Veldkamp, B. P., & Reese, L. M. (2000). An integer programming approach to item pool design. *Applied Psychological Measurement, 24*, 139–150.

Veldkamp, B. P. (2001). *Implementing constrained CAT with shadow tests for large item pools*. Manuscript submitted for publication.

Veldkamp, B. P., & van der Linden, W. J. (2000). Designing item pools for computerized adaptive testing. In W. J. van der Linden & C. A. W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 149–162). Boston: Kluwer Academic Publishers.

Author's Note

The authors are most indebted to Wim M. M. Tielen for his computational assistance.