Measurement and Research Department Reports 2007-1

# A Multiple Objective Test Assembly Approach for Exposure Control Problems in Computerized Adaptive Testing

**Bernard P. Veldkamp**
**Angela J. Verschoor**
**Theo J.H.M. Eggen**

Measurement and Research Department Reports 2007-1

# A Multiple Objective Test Assembly Approach for Exposure Control Problems in Computerized Adaptive Testing

Bernard P. Veldkamp University of Twente

Angela J. Verschoor Cito

Theo J.H.M. Eggen Cito

Cito
Arnhem, 2007

## Abstract

Over exposure and under exposure of items in the bank are serious problems in operational computerized adaptive testing (CAT) systems. These exposure problems might result in item compromise, or point at a waste of investments. The exposure control problem can be viewed as a test assembly problem with multiple objectives. Information in the test is maximized, item compromise is minimized, and pool usage in optimized. In this paper, a multiple objective method is developed to deal with both types of exposure problems. In this method, exposure control parameters based on observed exposure rates are implemented as weights for the information in the item selection procedure. The method does not need time consuming simulation studies, and it can be implemented conditional on ability level. The method is compared with Sympson Hetter method for exposure control, with the Progressive method and with alpha-stratified testing. The results show that the method is successful in dealing with both kinds of exposure problems, at the costs of an increased RMSE, while the bias remains comparable for all methods.

**Keywords:** computerized adaptive testing, exposure control, item bank usage, item selection, item response theory.

## Introduction

In computerized adaptive testing (CAT), items are selected on-the-fly. Adaptive procedures are used to select items with optimal measurement characteristics at the estimated ability level of examinees. CAT possesses the same advantages as other computer-based testing procedures, like increased flexibility and connection of administrative systems. Besides, for a CAT it also holds that test length can be decreased by almost 40 percent without decrease of measurement precision, and examinees are no longer frustrated by items that are either too difficult or too easy (see e.g. Wainer, Dorans, Flaugher, Green, Mislevy, Steinberg, & Thissen, 1990, van der Linden, & Glas, 2000).

CAT systems are theoretically based on the properties of item response theory (IRT). In IRT, person parameters and item parameters are separated. The item parameters are supposed to be invariant for different values of the person parameters. Therefore, items can be calibrated and the item parameters can be stored in item banks. From these item banks, items that provide most information at the estimated person parameter are selected. In many large scale testing programs, paper-and-pencil test have been replaced by CATs. For example for the Graduate Record Examination (GRE) and the Armed Services Vocational Aptitude Battery (ASVAB), CAT-versions are available now.

CITO (National Institute of Educational Measurement) in the Netherlands administers several CATs, like MATHCAT (CITO, 1999), TURCAT (CITO, in press), DSLcat (CITO, 2002) and KindergartenCAT. MATHCAT is developed for diagnosing Mathematics deficiencies for college students (Verschoor, & Straetmans, 2000), TURCAT tests proficiency of Turkish as a second language, DSLcat tests Dutch as a Second Language, and KindergartenCAT contains tests for measuring ordering, language, and orientation in time and space abilities of young children (Eggen, 2004). These CATs, like almost all operational CAT systems encounter an unevenly distributed use of items in the bank.

In general, most item selection procedures favor some items above others, due to superior measurement properties or favorable item characteristics. As a result, some items are overexposed. This might result in item compromise, which undermines the validity of score-based inferences (Wise & Kingsbury, 2000). On the other hand, some items might be underexposed, which is a waste of investments. Therefore, choosing a strategy for controlling the exposure of items to examinees has become an integral part of test development (Davis & Dodd, 2003).

## Theoretical background

One of the first methods developed to deal with exposure control problems, is the 5-4-3-2-1 technique (Hetter, & Sympson, 1997, McBride, & Martin, 1983) applied in the CAT-ASVAB. This randomized procedure was developed to reduce probability of item sequences in the first five iterations of CAT. Kingsbury and Zara (1989) and Thomasson (1998) developed different randomization methods aimed to reduce overall item exposure. Rotating item pool methods (Way, 1998, Way, Steffen, and Anderson, 1998, Ariel, Veldkamp, and van der Linden, 2004) and CAST (Luecht & Nungester, 1998) were developed to spread the items over different tests by a priori reducing the availability of items for selection. However, in CAT industry item-exposure control method based on the Sympson and Hetter method (1985) are most commonly applied.

### Sympson-Hetter methods

Although some variations exist, the general idea underlying these methods can be described as follows. To define these methods two events have to be distinguished, the event that item $i$ is selected by the CAT algorithm ($S_i$), and the event that item $i$ is administered ($A_i$). The probability that event $A_i$ occurs is the probability that $A_i$ occurs given that $S_i$ has occurred times the probability that $S_i$ occurs:

$$P(A_i) = P(A_i|\, S_i) * P(S_i). \quad (1)$$

To control the item exposure, one could focus on either of both probabilities. In the Sympson-Hetter methods, exposure control is conducted after an item is selected. The conditional probabilities $P(A_i|\, S_i)$ are used as control parameters. These control parameters guide the probability experiment in which it is determined whether the selected item is administered or removed temporarily for the person tested from the pool.

When $r_{max}$ is the target value for the maximum exposure rate, the conditional probabilities can be set to make sure that $P(A_i) \leq r_{max}$. The procedure to find appropriate values for the control parameters is quite time consuming. In a series of iterative adjustment, the appropriate values can be found.

These Sympson-Hetter methods suffer from several drawbacks. When the population is categorized based on ability, the exposure rates within sub groups might still be high. Time-consuming simulation studies have to be conducted for calculating the exposure control parameters. Moreover, the procedure for calculating the control parameters does not converge properly. Finally, it is also known that the Sympson-Hetter method is hardly effective in dealing with underexposure problems. Underexposure refers to the problem that items in the pool are administered so seldom, that the expense for constructing them can not be justified.

Several improvements of the original procedure have been developed. Stocking & Lewis (1998) proposed to conduct exposure control conditional on ability level, to overcome the problem of high exposure rates for specific ability levels. They defined the events in (1) conditional on ability level. The new relationship can be described as

$$P(A_i|\theta_j) = P(A_i|\, S_i, \theta_j) * P(S_i|\theta_j). \qquad j=1,..,J, \tag{2}$$

where $J$ defines the number of ability levels to take into account. The time needed to calculate the exposure control parameters increases $J$ times, because control parameters have to be calculated for all $J$ ability parameters. When this new procedure is applied, exposure rates within subgroups of the ability scale will also be below the specified level. This modification solves one of the problems of the method, but convergence problems and loss of total test information still exists.

Van der Linden (2003) proposed to modify the Sympson-Hetter method to speed up the iterative adjustment process to find the exposure control parameters. In the Sympson-Hetter method, the exposure parameters are adjusted with the following rule:

$$P^{t+1}(A_i \mid S_i) := \begin{cases} 1 & \text{if } P^t(S_i) \le r_{\max} \\ r_{\max} / P^t(S_i) & \text{if } P^t(S_i) > r_{\max} \end{cases} \tag{3}$$

where $t$ is the iteration number, and $r_{max}$ is the desired target for the exposure parameters. The adjustment process can be speeded up by changing this rule into

$$P^{t+1}(A_i \mid S_i) := \begin{cases} P^t(A_i \mid S_i) & \text{if } P^t(A_i) \le r_{\max} \\ r_{\max} / P^t(S_i) - \gamma & \text{if } P^t(A_i) > r_{\max} \end{cases} \tag{4}$$

where γ is a parameter to increase the size of the adjustment. Although less time is needed for finding exposure control parameters, the process is still generally tedious and time-consuming, particularly if the control parameters have to be set conditionally on a set of realistic ability values for the population of examinees.

Barrada, Veldkamp & Olea (2006) modified the Sympson-Hetter approach by varying the exposure control parameters throughout the test administration. To avoid that all items with high discriminating power are selected when estimation of trait levels is still uncertain, low values for $r_{max}$ are imposed at the beginning of the test. The values of $r_{max}$ increase during CAT administration. So, highly discriminating items are reserved for the later stages of the test.

Recently, van der Linden and Veldkamp (2004, in press) proposed to formulate the exposure control problem as a problem of constrained test assembly. Like the Sympson-Hetter method a probabilistic algorithm is used. However, this method does not need time consuming simulation studies to find control parameters for the probabilistic experiment. Based on the observed exposure rates, the algorithm determines whether item eligibility constraints are added to the model for selecting the items in CAT. The method proved to perform well in dealing with (over)exposure of popular items in the bank.

Both the (modified) Sympson-Hetter methods and the Eligibility methods mainly focus on overexposure of popular items in the pool. Although decrease of exposure rates of the most popular items results in some increase of exposure rates of less popular items, only exposure rates of items with almost as favorable attributes as the most popular items increase. Unpopular items are still hardly selected.

**Methods for controlling underexposure**

For solving the problem of underexposure, different methods have been developed. Chang & Ying (1999) introduced α-stratified testing. In their approach, item pools are stratified with respect to values of their discrimination parameters α. The first items are chosen from the stratum with lowest α values. A second group of items are chosen from the subsequent stratum, and the last items in the test from the stratum with highest α values. This approach is based on the observation that estimates of the ability parameters are very unstable during the first few iterations of a CAT. Because of this, less discriminating items should be used in the earlier stages, while the most discriminating items should be used when estimates have been stabilized. The claim is that this approach

would lead to a more balanced item exposure distribution and improve item pool utilization. Unfortunately, this method does not impose any bounds on exposure rates. Some observed exposure rates might be much higher than expected (Parshall, Kromrey, & Hogarty, 2000). Besides, the method is highly dependent on item bank properties. Usually, discrimination parameters are not uniformly distributed or the discrimination and the difficulty parameters might correlate positively.

A different method for solving the problem of underexposure is based on the observation that exposure problems result from the item selection criterion that is applied. When items are selected that maximize Fisher's Information criterion, items with high discrimination values tend to be selected more often than the others. One way to reduce both over- and underexposure is to add a random component to the item selection criterion. Revuelta and Ponsoda (1998) elaborated this idea in their Progressive method. When this method is applied, a random value $R_i$ in the interval [0,H], where H is the maximum value of the information function, is assigned to each item in the bank. Items are selected based on a weighted combination of the random component and Fisher's information criterion:

$$\arg(i) = \max_i (1 - \frac{s}{n}) R_i + \frac{s}{n} I_i(\hat{\theta}) \qquad (5)$$

where the weighting factor is determined by the serial position $s$ of the item in the test, and the total test length $n$. For selecting the first item, the value of the criterion is dominated by the value of the random component, while for selecting the last item, the random component does not influence the criterion anymore. This method proved to be effective against underexposure, however, it is not conditional on ability level, and it can not be guaranteed that targets for exposure rates will be met. Another drawback is that items that are completely off target might be presented to a candidate.

Dealing with exposure control problems in CAT is rather complicated. Although several promising methods have been developed, all of them seem to suffer from various drawbacks. Because of this, exposure control problems still exist. In most large scale testing systems, a rather pragmatic approach is used and a combination of over- and underexposure control methods is implemented. For example, in most CATs developed by CITO, a combination of the Sympson-Hetter method and a generalization of the

Progressive method is implemented (Eggen, 2001). By implementing a combination of methods, an attempt is made both to maximize measurement accuracy, and to balance item pool usage.

## Multi-objectivity and exposure control.

When an exposure control method is implemented, the test assembly problem can be formulated as an instance of multiple objective decision making (Veldkamp, 1999). The first objective is to assemble tests accordingly to the test specifications. In general, the amount of information in the test is maximized, while a number of constraints on test content, item format, word count or gender orientation of the items have to be met. The second objective in the process is related to exposure of the items. The objective is to obtain an evenly distributed use of items in the bank. The observation that the exposure control problem is a problem of multiple objectives in test assembly is the corner stone of the method presented in this paper. The main idea is that exposure control methods should represent this multiple-objectivity.

Both objectives can be formulated in mathematical programming terms. The first objective can be formulated as:

$$
\begin{aligned}
&\max \sum_{i=1}^{I} I_i(\theta) x_i \\
&\text{subject to} \\
&\sum_{i \in S_j} x_i \leq b_j \qquad \text{(categorical)} \\
&\sum_{i=1}^{I} a_{ij} x_i \leq b_j \qquad \text{(quantitative)} \\
&\sum_{i \in S_e} x_i \leq 1 \qquad \text{(inter - item dependencies)} \\
&\sum_{i=1}^{I} x_i \leq n \qquad \text{(test length)} \\
&x_i \in \{0,1\}
\end{aligned}
\tag{6}
$$

where $x_i$ denotes whether an item is selected ($x_i = 1$) or not ($x_i = 0$). The information in the test is maximized. The first general constraint represents constraints like content or item type. The second constraint represents specifications related to quantitative attributes like

word count or response times. The third constraint is formulated to deal with dependencies between items like enemies, but also item sets. In this way, the first objective can be obtained. To formulate the second goal is slightly more complicated. In van der Linden and Veldkamp (in press) it is shown that the following equality holds:

$$\sum_i \varphi_i = n, \tag{7}$$

where $\varphi_i$ is the observed exposure rate, and $n$ represents the test length. Because of this, it suffices to minimize the maximum exposure rate to obtain an evenly distributed use of the items in the bank. Therefore, the second objective can be formulated as

$$\min \max_i \frac{n\varphi_i + x_i}{n+1}. \tag{8}$$

These two objectives might conflict. To maximize the amount of information in the test, highly discriminating items are often selected. On the other hand, to obtain an evenly distributed use of the bank, these popular items can not be administered to all candidates. It comes down to the test assemblers preferences, how to deal with these conflicting objectives. One method for dealing with multiple objective test assembly problems is to use weighting functions (Veldkamp, 1999). When this method is applied to the exposure control problem, the information is weighted with some function of the observed item exposure rates. The resulting objective of the test assembly problem can be formulated as:

$$\max \sum_i w(\varphi_i) I_i(\theta) x_i \tag{9}$$

where $w(\varphi_i)$ is a weighting function that represents the test assemblers preferences.
Several weighting functions can be applied. For example, the function can be based on the observation that the use of popular items can be reduced by temporarily removing them from the pool of available items, until their observed exposure rate is smaller than $r_{max}$. This weighting function is shown in Figure 1a.

A second example is based on the observation that the use of unpopular items ($\varphi_i << r_{max}$) can be increased by increasing their weights. To boost the use of unpopular

items, the weighting function might decrease for increasing exposure rates. This observation results in a weighting function shown in Figure 1b.

The third example is related to test fairness. Because expelling some items from administration for some students, as in the first and second weighting function, might not be considered fair, assigning a small weight for popular items ($\varphi_i > r_{max}$) reduces the probability that they are selected, but does not make them ineligible. Two weighting functions that combine observations two and three are shown in Figures 1c and 1d.

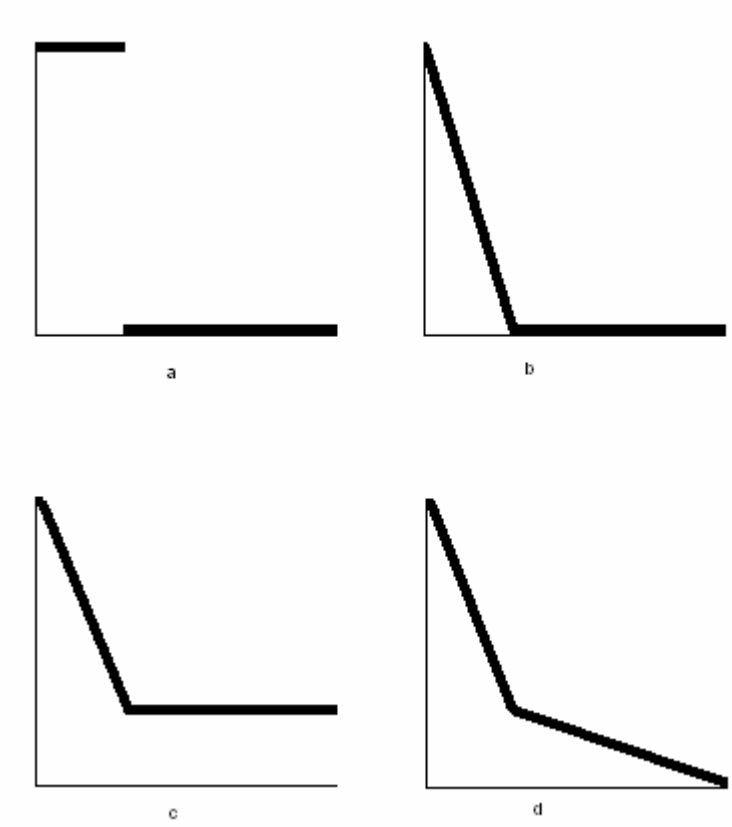


Figure 1. Weighting functions.

Moreover, the causes of over exposure can be taken into account when the weighting function is defined. The main cause of exposure problems lays in the amount of information provided by the item. Since the amount of information presented by an item is related to the squared discrimination of an item, a weighting function that takes the amount of information into account can be formulated as:

$$w_i(\phi_i) = a_i^{-2} \quad (\phi > \mathrm{r}_{\max}) \tag{10}$$

In all these examples, a difference is made between items that are overexposed ($\varphi_i > r_{max}$) and those who are not ($\varphi_i \leq r_{max}$). For both intervals different weighting functions can be defined, based on a number of observations. However, the question remains which weighting function performs best for which interval.
A systematic approach to answer this question would be to distinguish between both intervals and to see which function for which interval results in the best exposure control method.

## Numerical examples

A comparison study was carried out to judge the performance of the multiple objective exposure control method. Several settings of the method were compared with the Sympson-Hetter method, the alpha-stratified method, randomized item selection, and CAT without exposure control. In the first example, different weighting functions were compared. Different methods for exposure control were compared in Example 2.

**Example 1.**

To find the best settings for the multiple objective exposure control method, several functions were implemented. The items in the bank were calibrated with the OPLM, a special version of the 2PLM, where the discrimination parameters are restricted to be integer. The OPLM is the general IRT model underlying all CATs developed by CITO. The item bank consists of 300. The test length of all CATs was set equal to 40 items. Fisher's Information criterion was used to select the items. The ability was estimated with the Weighted maximum likelihood estimator (Warm, 1989), assuming that the item parameters are known. For all examples, 5000 examinees were randomly sampled from a normal distribution. The maximum exposure rate $r_{\max}$ was set equal to $r_{\max} = 0.20$ in the examples. These settings most closely resembled the CITO context.

To compare the results, the following criteria were applied. The performance of the CAT was evaluated by taking both the bias and the root mean squared error (RMSE) into account.

$$bias = \frac{\sum_{\mathrm{p}=1}^{\mathrm{P}} (\hat{\theta}_{\mathrm{p}} - \theta_p)}{\mathrm{n}} \tag{11}$$

$$RMSE = \left[ \frac{\sum_{p=1}^{P} (\hat{\theta}_p - \theta_p)^2}{n} \right]^{\frac{1}{2}} \tag{12}$$

where $p = 1,\ldots,P$ runs over all persons.

To control for underexposure of the items, three different functions were distinguished for $\varphi_i \leq r_{max}$. The first function does not control for underexposure of the items ($w_i(\phi_i) = 1$). The second function tries to control for underexposure by assigning decreasing weights when the observed exposure rate increases. The function is defined such that the weight equals one for items that have not been administered yet ($w_i(\phi_i = 0) = 1$), and it linearly decreases, where the weight for items with observed exposure equal to $r_{\max}$ is set equal to a constant ($w_i(\phi_i = r_{\max}) = c$, where c << 1). The third function aims at the causes of underexposure, and relates the weights to the inverse of the squared discrimination.

For overexposure ($\varphi_i > r_{max}$), four different functions where distinguished in this study. First, overexposure was not allowed ($w_i(\phi_i) = 0$). In the second function, a small weight is assigned ($w_i(\phi_i) = c$). In the third function, the weight linearly decreases, where the weight for items with observed exposure equal to $r_{\max}$ is set equal to a constant ($w_i(\phi_i = r_{\max}) = c$, where c << 1), and the weight is set equal to zero when the observed exposure rate equals one ($w_i(\phi_i = 1) = 0$). The fourth function aims at the causes of overexposure, and relates the weights to the inverse of the squared discrimination. In the examples, the constant was set equal to $c = 0.2$.

When the multiple objective exposure control method is applied, any weighting function is a combination of function for controlling underexposure and a function for controlling overexposure of the items. The weighting functions were compared for two different settings, $r_{max}$ = 0.2. Since 40 items were selected from an item bank of 300 items, the lower bound for $r_{max}$ equals 0.14. Resulting bias and RMSE for $r_{max}$ = 0.2 are shown in Table 1 and Table 2. The exposure rates of the items are shown in Figure 2.

Figure 2. Observed exposure for different settings of the multiple objective exposure control method $r_{max}$=0.20.

Table 1. Bias for different combinations of weighting functions for under- and overexposure ($r_{max} = 0.2$).

| | Overexposure | | | |
|---|---|---|---|---|
| Underexposure | $w_i(\phi_i) = 0$ | $w_i(\phi_i) = c$ | $w_i(\phi_i) = \text{linear}$ | $w_i(\phi_i) = a_i^{-2}$ |
| $w_i(\phi_i) = 1$ | 0.001 | 0.000 | 0.002 | 0.000 |
| $w_i(\phi_i) = \text{linear}$ | 0.002 | 0.001 | -0.001 | 0.002 |
| $w_i(\phi_i) = a_i^{-2}$ | 0.002 | 0.002 | 0.002 | 0.000 |

As can be seen in Table 1, the values for the resulting biases hardly differ, and no significant differences between the conditions were found.

Table 2. RMSEs for different combinations of weighting functions for under- and overexposure ($r_{max} = 0.2$).

| | Overexposure | | | |
|---|---|---|---|---|
| Underexposure | $w_i(\phi_i) = 0$ | $w_i(\phi_i) = c$ | $w_i(\phi_i) = \text{linear}$ | $w_i(\phi_i) = a_i^{-2}$ |
| $w_i(\phi_i) = 1$ | 0.123 | 0.118 | 0.120 | 0.120 |
| $w_i(\phi_i) = \text{linear}$ | 0.120 | 0.107 | 0.114 | 0.119 |
| $w_i(\phi_i) = a_i^{-2}$ | 0.123 | 0.108 | 0.113 | 0.116 |

With respect to functions controlling for overexposure, the results were more or less what we had expected. The conditions where no overlap was allowed resulted in highest values for the RMSE. Lowest values were obtained when small weights were assigned to overexposed items. Both adaptive functions ended up somewhere between them. An unexpected effect was that controlling for underexposure resulted in smaller RMSEs.

The observed exposure rates are shown in Figure 2. This figure has to be read in the same way as both tables; the first column of the first row describes the results for the condition of no underexposure control $w_i(\phi_i) = 1$, and no overexposure allowed $w_i(\phi_i) = 0$, etc..

For underexposure, the method with decreasing weights (row 2) performed best. It performed better than the cases were no underexposure control was applied (row 1), and

also better then the cases were the causes of underexposure were taken into account (row 3).

For overexposure, the results were less obvious. When underexposure was not controlled for (row 1), all proposed control functions performed equally well with respect to control of overexposure. When underexposure was taken into account (rows 2 and 3), the best results were obtained when no overexposure was allowed (column 1) or when the causes of overexposure were taken into account (column 4). Allowing overexposed items to be used (column 2) resulted in high overexposure of some popular items. These results can be explained by checking the weighting functions. Because the weighting functions just weight the information provided by an item, very informative items might still be selected when the difference in weights between overexposed and less popular items is small. The method of decreasing weights (column 3), also resulted in overexposure of the most popular items. Overall, the best results were obtained in when overexposure was controlled with decreasing weights (row 2) and no overexposure was allowed (column 1).

**Example 2.**

To evaluate the performance of the multiple objective exposure control method, it was compared with the alpha-stratified method, the Sympson-Hetter method, and the progressive method in combination with Sympson-Hetter. To add some benchmarks, both randomized item selection and item selection based on Fisher Information without exposure control were added to the example. For every exposure control method, 5000 CATs were simulated. The maximum exposure rates were set equal to $r_{max} = 0.20$ in these simulations. The results are shown in Table 3.

Table 3. Performance different exposure control methods $r_{max} = 0.20$.

| Method | Bias | RMSE |
|---|---|---|
| no exposure control | -0.001 | 0.085 |
| Sympson-Hetter method | -0.003 | 0.116 |
| Alpha-stratified method | -0.002 | 0.107 |
| Progressive method (S-H) | -0.007 | 0.116 |
| Randomized item selection | -0.004 | 0.134 |

When the results in Table 3 are compared with the results in Table 1 and Table 2 it can be observed that the multiple objective exposure control method results in slightly smaller biases. The Alpha stratified method, resulted in the smallest RMSE, both the Sympson-Hetter method and the Progressive method combined with Sympson-Hetter resulted in RMSE's comparable with those of the multiple objective exposure control functions method.

The observed exposure rates are shown in Figure 3. It can be seen that our implementation of the alpha-stratified method was not very successful. For some items the observed exposure rate exceeded 0.40. A different stratification might have performed better, although we did not succeed in finding good settings. Almost no differences were found between the Sympson-Hetter method and the Progressive method. The progressive method performed slightly better with respect to underexposure. The multiple objective exposure control method performed slightly better with respect to control of under-exposure.

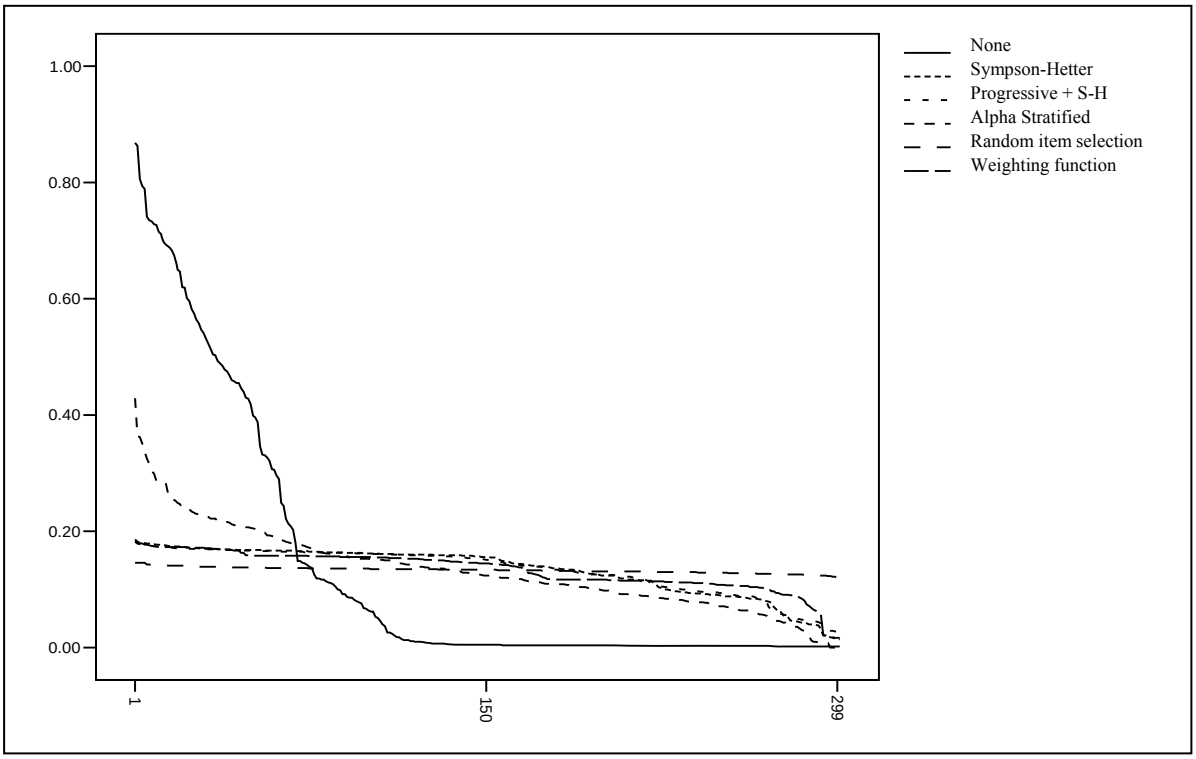


Figure 3. Observed exposure rates for different exposure control problems.

## Discussion

Exposure control is applied to computer adaptive testing programs for several reasons. The most important reason is to prevent item compromise. A second reason is to increase the usage of the item pool. Until now, several exposure control methods have been developed that deal with the problem of over exposure successfully. Under exposure of the items is still a problem in many adaptive testing programs.

The multiple objective exposure control method was developed to deal with both kinds of exposure control problems. One of the advantages of the new method is that no time consuming simulation studies have to be carried out. The new method can be implemented 'on the fly'. During the administration, the additional time for selecting an item with the multiple objective exposure control method was less than a millisecond. In the first example, it can be observed how the weighting functions influence the resulting tests. For example, the best results for the RMSE are obtained for an weighting function that allowed overexposure of some popular items. In other words, the tradeoff between RMSE and observed exposure rates can be controlled by defining appropriate weighting functions.

The multiple objective exposure control method was described as a deterministic method of exposure control. This implies that any administration of the test directly influence the weights for the next candidates. If such a dependency is undesirable, a probabilistic implementation might be considered. The weighting functions $w(\phi_i)$ determine the probability for an item $i$ to be selected. Before any CAT is administered, a probability experiment is carried out for every item to decide whether it is selected for the pool or not. For examinee $j$+1, item $i$ is eligible, that means available for selection, with estimated probability

$$P^{(j+1)}(E_i) = w(\phi_i) \tag{13}$$

where $E_i$ denotes the event that item $i$ is eligible. This probability experiment is comparable to the one described in van der Linden & Veldkamp (2004). However, in this approach the test specialist can define the function that relates the observed exposure rates

to the probability of being eligible. The result of this experiment is a subset of the item pool that can be used for test administration.

Finally, since the multiple objective exposure control method is an interactive method where the parameters effecting the exposure control method are updated during the test administration period, some remarks have to be made about practical implementation. In a web-based environment, with testing over the internet, updating the parameters on-the-fly seems rather straightforward. However, when thousands of examinees participate in a test at the same time updating the parameters every few minutes instead of continuously updating might be considered. This will reduce the probability of crashing the web-server. When the method is applied in classroom setting, which is most common for CITO CATs, the exposure rates resulting from different locations can be combined periodically.

## References


Ariel, A, Veldkamp, B.P., & van der Linden, W.J. (2004). Constructing rotating item pools for constrained adaptive testing. *Journal of Educational Measurement, 41,* 345-360.

Barrada, J.R., Veldkamp, B.P., & Olea, J. (2006). Multiple maximum exposure rates in computerized adaptive testing. *Manuscript submitted for publication.*

Chang, H-H, & Ying, Z. (1999). α-Stratified computerized adaptive testing. *Applied Psychological Measurement, 23,* 211-222.

CITO (1999). *WISCAT. Een computergestuurd toetspakket voor rekenen en wiskunde.* [Mathcat: A computerized test package for arithmetic and mathematics]. CITO: Arnhem.

CITO (2002). *NT2cat. Een computergestuurd toetspakket voor Nederlands als tweede taal.* [DSLcat. A computerized test package for Dutch as a Second Language]. CITO: Arnhem.

CITO (in press). *TURCAT. Een computergestuurd toetspakket voor Turks als tweede taal.* [TURCAT. A computerized test package for Turkish as a Second Language]. CITO: Arnhem.

Davis, L.L., & Dodd, B. (2003). Item exposure constraints for testlets in the verbal reasoning section of the MCAT. *Applied Psychological Measurement, 27,* 335-356.

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

Eggen, T.J.H.M. (2004). *CATs for kids: easy and efficient.* Paper presented at the 2004 meeting of Association of Test Publishers, Palm Springs, CA.

Hetter, R.D., & Sympson, J.B. (1997). Item exposure control in CAT-ASVAB. In W. Sands, B.K. Waters, & J.R. McBride (Eds.), *Computerized adaptive testing – from inquiry to operation* (pp. 141-144). Washington, DC: American Psychological Association.

Kingsbury, G.G. & Zara, A.R. (1989). Procedures for selecting items for computerized adaptive tests. *Applied Measurement in Education, 2,* 359-375.

Luecht, R.M., & Nungester, R.J. (1998). Some practical examples of computer-adaptive sequential testing. *Journal of Educational Measurement, 35,* 229-249.

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* (pp. 223-226). New York: Academic Press.

Parshall, C., Harmes, J.C., & Kromrey, J.D. (2000). Item exposure control in computer-adaptive testing: The use of freezing to augment stratification. *Florida Journal of Educational Research, 40,* 28-52.

Revuelta, J. & Ponsada, 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.

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.

Thomasson, G.L. (1998). *CAT item exposure control: New evalutation tools, alternate method and integration into a total CAT program.* Paper presented at the annual meeting of the National Council of Measurement in Education, San Diego.

van der Linden, W.J. (2000). Constrained adaptive testing with shadow tests. In W.J. van der Linden, and C.A.W. Glas (Eds.) *Computerized adaptive testing: Theory and practice* (pp. 1-25). Boston, MA: Kluwer Academic Publishers.

van der Linden, W. J. (2003). Some alternatives to Sympson-Hetter item-exposure control in computerized adaptive testing. *Journal of Educational and Behavioral Statistics, 28,* 249-265.

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

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

van der Linden, W. J., & Veldkamp, B. P. (2007). Conditional item-exposure control in adaptive testing using item-ineligibility probabilities. *Journal of Educational and Behavioral Statistics, 32.* In press.

Veldkamp, B.P. (1999). Multiple objective test assembly problems. *Journal of Educational Measurement, 36,* 253-266

Verschoor, A.J., & Straetmans, G.J.J.N. (2000). MathCAT: A flexible testing system in mathematics education for adults. In W.J. van der Linden, and C.A.W. Glas (Eds.) *Computerized adaptive testing: Theory and practice* (pp. 101-116). Boston, MA: Kluwer Academic Publishers.

Warm, T.A. (1989). Weighted maximum likelihood estimation of ability in item response theory. *Psychometrika, 54,* 427-450.

Wainer, H., Dorans, N.J., Flaugher, R., Green, B.F., Mislevy, R.J., Steinberg, L., Thissen, D. (1990). *Computerized adaptive testing: A primer.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Way, W.D. (1998). Protecting the integrity of computerized testing item pools. *Educational Measurement, Issues and Practice, 17,* 17-27.

Way, W.D., Steffen, M., & Anderson, G.S. (1998). *Developing, maintaining, and renewing the item inventory to support computer-based testing.* Paper presented at the colloquium on computer-based testing: Building the foundation for future assessments, Philadelphia, PA.

Wise, S.L., & Kingsbury, G.G. (2000). Practical issues in developing and maintaining a computerized adaptive testing program. *Psycologica, 21,* 135-156.