

**Implementing Simpson-Hetter Item-Exposure Control in a Shadow-Test Approach  
to Constrained Adaptive Testing**

Bernard P. Veldkamp

Wim J. van der Linden

University of Twente

The Netherlands

Word count: 5852

Submission date: May 19th

Bernard P. Veldkamp

University of Twente, Faculty of Behavioral Sciences (OMD), P.O. Box 217, 7500 AE Enschede, The Netherlands. E-mail: [b.p.veldkamp@utwente.nl](mailto:b.p.veldkamp@utwente.nl).

Wim J van der Linden

Current address: CTB/McGraw-Hill, 20 Ryan Ranch Road, Monterey, CA 93940, USA.

E-mail: [wim\\_vanderlinden@ctb.com](mailto:wim_vanderlinden@ctb.com)

## Implementing Sympson Hetter in the Shadow Test Approach -

Implementing Sympson Hetter in the Shadow Test Approach -

**Implementing Sympson-Hetter Item-Exposure Control in a Shadow-Test Approach  
to Constrained Adaptive Testing**

Submission date: May 19th,

**Abstract**

In most operational computerized adaptive testing programs, the Sympson-Hetter method is used to control the exposure of the items. Several modifications and improvements of the original method have been proposed. The Stocking and Lewis (1998) version of the method uses a multinomial experiment to select items. For severely constrained CAT, the list on which this experiment is conducted not only has to be of appropriate length but also needs to balance the composition of the test with respect to its specifications. In this paper, it is shown how the SH method for exposure control can be implemented in the shadow test approach. The method was applied to an adaptive test with 433 constraints on various attributes. Both a single and a multiple shadow-test approach was used to compare different list lengths for the SH method.

Key words: Adaptive Testing; Item-Exposure Control; Shadow Test; Sympson-Hetter Method.

## **Implementing Sympson-Hetter Item-Exposure Control in a Shadow-Test Approach to Constrained Adaptive Testing**

An ideal of adaptive testing is to run test administrations with a distribution of exposure rates for the items in the pool that approximates uniformity. Because item selection in adaptive testing is based on optimization of the information in the test, it is common to see a minority of informative items administered to many examinees while the majority of items is hardly used at all. When items are overexposed, it becomes easy for the test takers to get familiar with these items and share them with future test takers. On the other hand, for items that are underexposed, it is very hard to justify their costs of development and pre-testing. An important contribution to a solution to exposure problems is the use of probabilistic control exposure in the testing program. Following Way (1998), exposure control methods can be classified as randomized item selection and conditional item selection. A third type --stratification procedures--can be added to complete the classification. Some methods combine elements from different types.

The first to introduce a method of probabilistic control of item exposure were McBride and Martin (1983). This randomization method was actually intended to reduce exposure problems due to memorization of the first few items and has, to the authors' knowledge, never been used operationally in CAT. In the 5-4-3-2-1 method, lists of items with the highest information at a sequence of ability levels are prepared, and the items are randomly selected for administration from the list closest to the current ability estimate. McBride and Martin proposed to begin with a list of fixed length (5 items, say). The length of the list decreases after each item is administered. By the fifth item, the most informative item is selected for administration. A great advantage of this method is its simplicity. However, since the items are randomly sampled from the list, it is impossible to impose quantitative constraints, like constraints related to word count or time limits (see also, van der Linden, 2005), on the selection of the items for the test.

A different probabilistic method was suggested in Sympson and Hetter (1985; see also Hetter & Sympson, 1997). This conditional method intervenes during the test as soon as an item is selected for administration. It then conducts a probability experiment to determine if the item should be actually

## Implementing Sympson Hetter in the Shadow Test Approach -

administered. If it should not, the item is removed from the pool during the rest of the tests and a new item is selected. In the McBride and Martin method, items were sampled with replacement. In the Sympson-Hetter (SH) method they are removed from the pool, when they are not selected. Stocking and Lewis (1998) suggested using the method for exposure control conditionally on the ability parameters  $\theta$ . An advantage of the conditional version of the SH method is that it prevents the possibility of the marginal exposure rates of all items at acceptably low values but much larger rates for some groups of test takers of nearly the same ability level (Stocking & Lewis, 2000). More specifically, the conditional SH method is as follows. Suppose we want to control the exposure rates at a set of ability intervals with typical values  $\theta_k$ ,  $k = 1, \dots, K$ , for instance, their midpoints. The method is based on the following relation between probabilities defined at these points  $\theta_k$ :

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

where

$$\begin{aligned} P(A_i | \theta_k) : & \quad \text{Probability of administering item } i \\ P(A_i | S_i, \theta_k) : & \quad \text{control parameter for item } i \\ P(S_i | \theta_k) : & \quad \text{Probability of selecting item } i \end{aligned} \quad (2)$$

The relation in (1) holds because an item can only be administered if it has been selected; therefore,  $P(A_i | \theta_k) = P(A_i, S_i | \theta_k)$ . Control parameters  $P(A_i | S_i, \theta_k)$  are the conditional probabilities of administering an item given its selection. The objective of the SH method is to force the conditional exposure rates,  $P(A_i | \theta_k)$  to take values below a prespecified target,  $r^{\max}$ , for all items  $i = 1, \dots, I$ , in the pool. Generally, the control parameters require different values to realize this target for different items; for popular items they are low but for unpopular items they may even be equal to one.

Operational values for the control parameters are found using computer simulations of the adaptive test. In these simulations, the control parameters are adjusted iteratively until the exposure rates  $P(A_i | \theta_k)$  converge to admissible values. The rule proposed by Sympson and Hetter is

$$P^{(t+1)}(A_i | S_i, \theta_k) := \begin{cases} 1 & \text{if } P^{(t)}(S_i | \theta_k) \leq r^{\max}, \\ r^{\max} / P^{(t)}(S_i | \theta_k) & \text{if } P^{(t)}(S_i | \theta_k) > r^{\max}, \end{cases} \quad (3)$$

where  $t$  denotes the iteration number. However, convergence may be slow and is not always guaranteed; for an analysis of the difficulties that can be encountered, see van der Linden (2003).

In principle, the number of probability experiments that has to be conducted before an item is selected for administration is bounded only by the size of the item pool. To prevent unfortunate results, Stocking and Lewis (1998) suggest an implementation based on an alternative experiment. In this multinomial experiment, a list of the most informative items at the selected ability levels closest to the ability estimate is established, the control parameters for these items are renormed to sum to one, and an item is sampled for administration. In fact, this version of the SH method much resembles the McBride-Martin method; the only difference is simple random sampling from the lists of the most informative items versus multinomial sampling with the probabilities dictated by the control parameters.

Generally, lists of items for probabilistic item-exposure control have to meet two conditions:

- The first condition regards the size of the lists. This condition is about the number of items to conduct the probability experiment on. If the lists are short, the method selects the items only from a small portion of the item pool. Consequently, it loses its effectiveness as a method of exposure control but the accuracy of the ability estimates remains close to the maximum possible for the pool because not many items are rejected. If the lists are long, another phenomenon occurs. The probabilities in the multinomial experiment decrease toward the end of the list, and for longer lists

## Implementing Sympson Hetter in the Shadow Test Approach -

the probabilities of selecting any of the items located there may become negligible. It is therefore hard to formulate general rules for setting the size of the lists. In order to navigate adequately between the extremes, operational experience with the actual distribution of the item parameters in the pool and the nature of the content constraints on the test is required.

- The second condition is on the composition of the lists. This condition is about the kind of items that are selected for the list. The lists should not only consist of the most informative items but also have a balanced composition with respect to the content specifications of the test. This requirement involves complicated combinatorial constraints. For example, if an item on a list has a combination of an attribute that is overrepresented among the items already administered with another attribute that is still missing in the test, the item should never be allowed to enter a list because its selection leads to a test that cannot satisfy its specifications. Especially toward the end of the test, when each of the items already administered restricts the choice, finding a list with feasible items may become a challenge.

The Stocking and Lewis implementation of the SH method was originally developed within the framework of the weighted deviation method (WDM) for constrained test assembly (Stocking & Swanson, 1993). In this framework, test specifications are being seen as desired properties instead of hard constraints. As a consequence, some of the specifications might be violated in the assembly process. van der Linden (2005b) showed that the shadow test approach (STA) for constrained test assembly can be used to guarantee that all specifications are met without any loss of efficiency (for a detailed description of the method, see the next section).

The question of how to implement the SH method for exposure control within the framework of the STA still has to be addressed. Therefore the goal of this research was to develop a method for SH exposure control within the STA, which both balances the composition of the list for the multinomial experiment and maintains an appropriate length of it toward the end of the test.

### Implementing SH Control in the STA

The STA to adaptive testing (van der Linden, 2000) was introduced to obtain optimal measurement precision in adaptive testing and to guarantee that large amounts of test specifications would be met. In the shadow test approach, items are selected not directly from the pool but from shadow tests. A shadow test is a full test assembled in real time prior to the selection of the item that (i) meets all content constraints to be imposed on the adaptive test, (ii) is optimal at the ability estimate, (iii) contains all items already administered to the test taker.

The following pseudo-algorithm describes the regular STA:

- Step 1: Initialize the ability estimator;
- Step 2: Assemble a shadow test;
- Step 3: Administer the free items in the shadow test that is best at the current ability estimate;
- Step 4: Update the ability estimate;
- Step 5: Adjust the set of constraints to allow for the attributes of the items already administered;
- Step 6: Return all unused item to the pool;
- Step 7: Repeat Steps 2-6 until  $n$  items have been administered.

An important step in the STA is the second step with the assembly of the shadow test. The assembly should optimize the information function of the test subject to a potentially elaborated set of constraints on its content. An appropriate technique for this assembly is 0-1 linear programming (van der Linden, 2005). We then model the objective function and constraints using variables  $x_i$  for the selection of item  $i$  that are equal to one if the item is selected and equal to zero otherwise. The model can easily be solved in real time using standard commercial software, for instance, the CPLEX package (ILOG, 2003).

## Implementing Sympson Hetter in the Shadow Test Approach -

Let  $\hat{\theta}^{g-1}$  denote the update of ability estimate after  $g-1$  items in the test and  $R_g$  the set of indices of the items in the pool that are available for the selection of the  $g$ th item. An example of a model for the assembly of a shadow test is:

$$\text{maximize } \sum_{i=1}^I I_i(\hat{\theta}^{(g-1)})x_i, \quad (\text{maximum information}) \quad (4)$$

subject to

$$\sum_{i=1}^I x_i = n \quad (\text{test length}) \quad (5)$$

$$\sum_{i \in V_c} x_i \leq n_c, \quad \text{for all } c, \quad (\text{categorical attributes}) \quad (6)$$

$$\sum_{i=1}^I q_i x_i \leq b_q, \quad \text{for all } q, \quad (\text{quantitative attributes}) \quad (7)$$

$$\sum_{i \in R_g} x_i = g - 1, \quad (\text{previous items}) \quad (8)$$

$$\sum_{i \in V_e} x_i \leq 1, \quad \text{for all } e, \quad (\text{enemies}) \quad (9)$$

$$x_i \in \{0,1\}, \quad \text{for all } i. \quad (\text{range of variables}) \quad (10)$$

The objective function that is maximized in (4) is the value of the test information function at  $\hat{\theta}^{(g-1)}$ . The test length is fixed in (5). The constraint in (8) fixes the values of the decision variables of all items that have already been administered to one. In doing so, this constraint automatically allows for the attributes of these items with respect to the test specifications when selecting the other items in the test. The constraints in (10) define the range of values for the decision variables. The remaining constraints illustrate how the model deals with content specifications for the test. In (6) and (7),  $V_c$  and  $q_i$  denote the set of items with a common categorical attribute (e.g., a content category) and the value of item  $i$  for a quantitative

attribute (e.g., a word count), respectively. The constraints require the number of items with the categorical attribute and the sum of the values for the quantitative attribute to be between bounds. The constraint in (9) can be used to deal with sets of items,  $V_e$ , that are not allowed to be selected for the same test.

This model only illustrates a few possible types of content constraints. In a real-world application, we will need more constraints of the type in (6) and (7), and possibly also constraints to deal with the presence of item sets in the pool and the attributes of their stimuli. For more details about modeling shadow tests, see van der Linden (2005, chap. 9). As for the solution of the model, it is still possible to set up an algorithm that selects shadow tests in a negligible time (see the empirical example below).

### **Item List Composition**

To conduct the multinomial experiment in the SH method, its list of items should both be long enough and balance the composition of the test with respect to its specifications. The set of free items in the shadow test forms a natural list of items for the multinomial experiment in the SH method. As each shadow test meets all constraints on the test, it follows that as long as the items are selected from the free items in the shadow tests, the adaptive test automatically meets all constraints. However, the set of free items in the shadow test is dynamic, and the number of free items decreases to zero towards the end of the test. Consequently, composing a list for the SH method will be possible at the beginning of the administration but may become severely restricted toward the end. The current research was motivated by this implementation problem.

The problem can be handled in several ways. One is to add items to the list from other sources. The only requirement is that these items balance the composition of the list with respect to the test specifications. This necessity motivated the multiple shadow test approach (MSTA).

The MSTA generalizes the STA in that prior to the selection of each item not one shadow tests but a small set of shadow tests is assembled. In addition to the three defining properties of a shadow test above, the shadow tests in this set have the following features: (i) all shadow tests are parallel in the sense that they

## Implementing Sympson Hetter in the Shadow Test Approach -

meet the same set of content constraints; (ii) each individual shadow test contains the set of items that have already been administered to the test taker; and (iii) there is no overlap among the free items between the shadow tests. Because later sets of shadow tests automatically correct for the consequences of earlier choices, a powerful feature of the MSTA is that *any subset of the free items in the shadow tests can be chosen*.

It may seem laborious to assemble a set of shadow tests but actually the change in the test-assembly model time is rather simple and straightforward. Let  $t$  denote an individual shadow tests in a set of  $T$  tests that is to be assembled prior to the selection of the  $g$ th item. The decision variables we now need are doubly indexed:

$$x_{it} = \begin{cases} 1, & \text{if item } i \text{ is selected for shadow test } t, \\ 0, & \text{otherwise,} \end{cases} \quad (11)$$

$i = 1, \dots, I$ , and  $t = 1, \dots, T$ . The model in (4)--(10) now becomes

$$\text{maximize } y \quad (\text{objective}) \quad (12)$$

subject to

$$\sum_{i=1}^I I_i(\theta^{(g-1)})x_{it} > y \quad \text{for all } t, \quad (\text{test information}) \quad (13)$$

$$\sum_{i=1}^I x_{it} = n \quad (\text{test length}) \quad (14)$$

$$\sum_{i \in V_c} x_{it} \leq n_c, \quad \text{for all } c, \quad (\text{categorical attributes}) \quad (15)$$

$$\sum_{i=1}^I q_i x_{it} \leq b_q, \quad \text{for all } q, \quad (\text{quantitative attributes}) \quad (16)$$

$$\sum_{t=1}^T x_{it} \leq 1, \quad \text{for all } i \in R_g, \quad (\text{item overlap}) \quad (17)$$

$$\sum_{i \in R_g} x_{it} = g - 1 \quad \text{for all } t, \quad \text{(previous items)} \quad (18)$$

$$\sum_{i \in V_e} x_{it} \leq 1, \quad \text{for all } t \text{ and } e, \quad \text{(enemies)} \quad (19)$$

$$y > 0, \quad (20)$$

$$x_{it} \in \{0,1\}, \quad \text{for all } i \text{ and } t. \quad \text{(range of variables)} \quad (21)$$

Except for the change of variables and the fact that all previous constraints now hold for each shadow test  $t$ , the only differences between the models for the STA and the MSTA are the constraints in (12)-(13), (17) and (20). In (13), the information in each of the shadow tests at  $\hat{\theta}^{(g-1)}$  is required to be larger than a common bound  $y$ , which is maximized in (12). The bound is defined as a positive real-valued variable in (20). This new objective is an application of the maximin principle, which guarantees that “no shadow test will be left behind”. The principle is important in that it enables us to assemble the lists for the multinomial experiment in the SH method from the joint set of free items in *all* shadow tests without having to sacrifice any information. The constraints in (18) prevent overlap between the free items in the shadow tests.

### Feasibility

For the STA, it is easy to show that if the item pool admits one test that satisfies all content constraints, the algorithm is always able to produce a shadow test. This feature does not necessarily hold if SH exposure control is introduced. The items on the lists that are passed over in the multinomial experiment are removed from the pool. As a consequence, when the item pool is not on target, depending on the length of the test, too many items with certain attributes may be removed from the pool and shadow tests that satisfy all constraints are no longer feasible. In fact, this consequence is not tied exclusively to the STA;

under the same conditions, any adaptive testing algorithm would fail to select a test that meets its specifications.

The aim of the MSTA is to produce longer lists of items for the multinomial experiment. However, longer lists tend to entail larger numbers of items that are passed over in the experiment. Furthermore, in the MSTA we assemble more than one shadow test at a time. If an infeasible problem is met, a simple solution, however, is to return all items that have been used to the pool and restart the assembly.

For a well-designed pool, we do not expect this to happen frequently. In the empirical study below, the measure was necessary for fewer than 0.1 % of the total number of items selected. The impact of such figures on the item-exposure rates is negligible.

### **Item Sets**

It is not unusual to have adaptive tests with items organized in sets around a common stimulus. Using separate variables for the selection of the stimuli and the items, models as in (12)-(21) can easily be extended to deal with the assembly of shadow tests with item sets. For the SH method, it is most convenient to treat item sets in shadow tests as “superitems”, that is, as single entries on the list of items for the multinomial experiment. Once such an entry is selected, item selection is constrained to be from the set until all lower bounds in the constraints on the set are met. From that point on, the competition between items within and outside the set is open again. The set is closed as soon as an item outside it has been chosen.

When item sets are present in the pool, it is only necessary to control the exposure at the level of the their stimuli. Since the exposure rate of any item in a set can never be larger than the rate of its stimulus, this choice effectively controls the exposure rates at the item level as well.

### **Example Study**

The STA was used to implement SH exposure control in an adaptive version of the Law School Admission Test (LSAT). We used a target  $r^{\max}=.20$  for all conditional exposure rates. Six different cases were studied. Each case was a combination of the following two factors:

1. One or two shadow tests;
2. Lists of the 5, 10, or 20 most informative items among the free items in the shadow test(s).

The case of a single shadow test with a list of five items can be seen as a baseline in this study.

### **Item Pool and Test**

The item pool was a previous 753-item pool from the LSAT. All items were calibrated using the 3-parameter logistic response model (Hambleton & Swaminathan, 1985). The LSAT consists of three different sections. Two of the sections have an item-set structure. The number of stimuli available in the pool for these two sections was 48.

A 50-item version of the LSAT was simulated. The length was half the length of the paper-and-pencil version of the LSAT. To retain all specifications for the LSAT, the bounds in the constraints in the model for the shadow tests were scaled down by a factor .50. The set of specifications implied 433 different constraints on various attributes in the test at test, stimulus, item set and individual item level,

### **CAT Algorithm**

Test administrations were simulated for equal numbers of examinees at  $\theta = -2, -1.5, \dots, 2$ . The reason why we used equal numbers and did not sample from some population distribution lies in the uniform precision of the estimates of the exposure rates of the items as well as the bias and MSE of the abilities we wanted to obtain. Because each of these estimates is a conditional quantity given  $\theta$ , as a consequence of our choice, the results generalize to populations of test takers with any distribution over  $\theta$ .

The estimator of  $\theta$  was the expected a posteriori (EAP) estimator with an uniform prior distribution over  $[-4,4]$ . The ability estimator was initialized at  $\hat{\theta}^{(0)} = 0$ . The items were selected using the maximum-information criterion.

The control parameters of the SH method were set using the iterated simulations with the adjustment rule in (3). At each step, we simulated 300 test administrations at each of the  $\theta$  values. In the main study, we simulated 700 administrations at each  $\theta$  value. As already stated, the goal of this study was to estimate the final conditional exposure rates for the items in the pool as well as the bias and the mean-squared error (MSE) functions of the ability estimator. These quantities were estimated at  $\theta = -2, -1.5, \dots, 2$ . Other measures, like item overlap (Chang and Zhang, 2002) could be applied, but we decided to follow Stocking & Lewis (1998) in their use of observed exposure rates as a measure for exposure control.

## Results

The shadow-test approach to SH item-exposure control was implemented in software developed in Delphi 6.0. CPLEX 9.0 was used as a solver for the assembly of the shadow tests. On a Intel Pentium IV, 3,20 GHz computer, the average time for selecting an item varied from 0.43 seconds for one shadow test to 1.01 seconds for two shadow tests with hardly any variation.

The numbers of iteration steps required to find satisfactory values for the SH control parameters was different for the six cases studied. In Table 1, for each case, the maximum conditional exposure rates over the nine values  $\theta = -2, -1.5, \dots, 2$  found at each of the steps are shown. In Table 2, the proportion of items with a violation of the target value of  $r^{\max} = .20$  over the same  $\theta$  values at each step is shown. Although both tables reveal a strong tendency for the maximum rate as well as the proportion of violations to decrease, differences between the rates for the different cases are visible. The following conclusions can be drawn:

1. The cases with two shadow tests tended to have the best results at an earlier iteration step than the cases with one shadow test.

2. The longer the list used in the multinomial experiment in the SH method, the better the maximum performance.
3. Both observations hold for the maximum exposure rate as well as the proportion of violations of the target of  $r^{\max}$ .

=====  
Tables 1 and 2 about here  
=====

The values of the control parameters at the iteration steps with the best results in Table 1 were used in the main study. The estimates of the conditional exposure rates for the items found in this study are shown in Figure 1. As in the previous process of iterative adjustment of the control parameters, the conditional rates of a few items still exceeded the target of .20; the largest conditional rate found in this study was equal to .29. Also, the largest conditional rates were generally found for lists of the five most informative items but hardly any differences existed between the lists of 10 and 20 most informative items. This can be explained by the nature of the multinomial experiment, for which, as already indicated, the probability of selection selected for the later items in the list when its length increases. When SH control for the STA was compared with that for the MSTA, slightly better results were obtained for the latter when the lists contained five or ten items. For lists of 20 items, hardly any differences were found between the two approaches.

In Figure 1, the observed exposure rates are reported conditional on ability level. The marginal observed exposure rates were smaller than 0.09 for all six conditions in this study, which is much lower than the target of .20. Conditional on ability level, between 125 and 150 items were being administered. The MSTA used slightly more items than the approach with only one shadow test. Overall, the total number of items used for the one shadow test approach varied from 425-434 items. For the two shadow test approach, the number of items used varied from 435-445. So, the MSTA produced a slightly more uniform usage of the items.

## Implementing Sympson Hetter in the Shadow Test Approach -

The highest exposure rates were observed for items selected toward the end of the tests when it becomes more difficult to balance the test content with respect to the specifications. In addition, the number of free items decreases even for a set of two shadow tests. However, in this simulation study, due to the quality of the item pool, the differences between the conditions with one and two shadow tests were not large.

=====

Figure 1 about here

=====

The estimates of the bias and MSE functions are shown in Figures 2 and 3. To assess the impact of the exposure control, the figure also shows the estimates for the same adaptive test without any exposure control. As expected, both the bias and the MSE were larger for the condition with exposure control method: The SH method ignores some of the best items when administering the adaptive test, and the loss of these items involves a price. However, the differences between the bias and MSE functions were nowhere larger than .04 --- an amount that we consider as negligible for most practical purposes. Based on Figure 2 and 3, we conclude that the exposure control method hardly influenced the resulting bias and MSE. In other words, for the item bank at hand, the exposure control was not very strict. This might be one of the reasons why we did not find many differences between one or two shadow tests.

=====

Figures 2 and 3 about here

=====

Although we studied different cases, the results were still for one item pool and one--albeit elaborate--set of test specifications. Variation in the size and composition of the item pool or in the test specification may

have an impact on the results. Care should therefore be taken if the conclusions are generalized to other applications.

### **Concluding Observations**

In this paper, it has been demonstrated how the SH method for exposure control can be implemented in constrained adaptive testing using the STA. The list of items needed for the multinomial experiment can be constructed from the shadow tests. In this way, the two requirements for successful implementation of the SH method identified earlier, namely sufficient numbers of items to choose from and content balancing throughout the test, can be met. An important result from the empirical study not shown in the figures above was that, for each simulated test taker, all content constraints on the adaptive test were realized. The power of the STA is precisely this feature: Even though the SH method rejects a substantial number of items before it finds the next item for administration, the use of the (M)STA guarantees an adaptive test that still meets all of the specifications. In this way, the SH method is also applicable even when large numbers of specifications have to be met.

In the simulation study, the lists in the multinomial experiment consisted of items from either one or two shadow tests. For the well-designed operational item bank used in this study, even the implementation with one shadow test provided good results. Besides, it hardly mattered whether the list consisted of 10 or 20 items. The main reason is that beyond a length of 10 items, the probabilities of selecting an item experiment became very small. Due to its dependence on the nature of the test specifications and the item pool, it is hard to generalize this critical length to other adaptive tests.

Since the decrease of the lists of items in the SH method occurs only at a rate of one item at a time towards the end of an adaptive test, other implementations of the STA can be considered. For example, the test could start with a single shadow tests but move to multiple shadow tests when the remaining set of free items becomes too small. In fact, we could even increase the number of shadow tests gradually to avoid any

## Implementing Simpson Hetter in the Shadow Test Approach -

decrease of the length of the list at all! A strong point of the STA is that, whatever implementation is chosen, the method always guarantees that all test specifications are met.

The SH method reduced the overexposure of the popular items in the pool considerably. Also, as a result of the exposure control, 435 from the 753 items in the pool became active during the test. From a cost-benefit perspective, this number is still low. But it is much higher than the 10% or so of the items generally used in adaptive testing without exposure control. To further increase item usage, different methods, such as  $\alpha$ -stratified adaptive testing (Chang & Ying, 1999), item-pool design with exposure control as an explicit objective (Veldkamp & van der Linden, 2000; van der Linden, Ariel & Veldkamp, 2006), or rotating item pools (Ariel, Veldkamp, & van der Linden, 2004; Stocking & Swanson, 1998) can be used. The Progressive method (Revuelta and Ponsoda, 1998) and the Progressive S-H method (Eggen, 2001) can also be applied to reduce the underexposure of some of the items. Each of these methods attacks the same problem from a different angle.

## 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--359.
- Chang, H., & Ying, Z. (1999). a-stratified multistage computerized adaptive testing. *Applied Psychological Measurement, 23*, 211--222.
- Chang, H.H., & Zhang, J. (2002). Hypergeometric family and item overlap rates in computerized adaptive testing. *Psychometrika, 67*, 387-398.
- Eggen, T.J.H.M. (2001). *Overexposure and underexposure of items in computerized adaptive testing*, Arnhem, The Netherlands: CITO.
- Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*, Boston, MA: Kluwer Nijhoff.
- Hetter, R. D., & Sympson, J. B. (1997). Item exposure control in CAT-ASVAB. In W. A. Sands, B. K. Waters, & J. R. McBride (Eds.), *Computerized Adaptive Testing: From Inquiry to Operation* (pp. 141-144). Washington, D.C.: American Psychological Association.
- ILOG, Inc. (2003). *CPLX 9.0*. Incline Village, NV: ILOG, Inc.
- 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.
- Revuelta, J., & Ponsoda, V. (1998). A comparison of item exposure control methods in computerized adaptive testing. *Journal of Educational Measurement, 35*, 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.

## Implementing Sympon Hetter in the Shadow Test Approach -

- Stocking, M. L., & Swanson, L. (1993). A method for severely constrained item selection in adaptive testing. *Applied Psychological Measurement, 17*, 277--292.
- Stocking, M. L., & Swanson, L. (1998). Optimal design of item banks for computerized adaptive testing. *Applied Psychological Measurement, 22*, 271--279.
- Sympson, J. B., & Hetter, R. D. (1985, October). Controlling item-exposure rates in computerized adaptive testing. *Proceedings of the 27<sup>th</sup> annual meeting of the Military Testing Association* (pp. 973-977). San Diego, CA: Navy Personnel Research and Development Center.
- 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). Dordrecht, the Netherlands: Kluwer Academic Publishers.
- van der Linden, W. J. (2003). Some alternatives to Sympon-Hetter item-exposure control in computerized adaptive testing. *Journal of Educational and Behavioral Statistics, 28*, 249-265.
- van der Linden, W. J. (2005). *Linear models for optimal test design*. New York: Springer.
- van der Linden, W. J. (2005b). A comparison of item-selection methods for adaptive tests with content constraints. *Journal of Educational Measurement, 42*, 283-302.
- van der Linden, W. J., Ariel, A., & Veldkamp, B. P. (2006). Assembling a CAT item pool as a set of linear test forms. *Journal of Educational and Behavioral Statistics, 31*, 81-100.
- 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). Dordrecht, the Netherlands: Kluwer Academic Publishers.
- Way, W.D. (1998). Protecting the integrity of computerized testing pools. *Educational Measurement: Issues and Practice, 17*, 12-27.

## Implementing Sympson Hetter in the Shadow Test Approach -

**TABLE 1.**

Maximum exposure rates over  $\theta = -2.0, -1.5, \dots, 2.0$  at each iteration step for the six cases studied. (Note: first best result for each case is printed in bold)

No. of Shadow Tests	1			2		
Length of List	5	10	20	5	10	20
Iteration step						
1	1.00	1.00	1.00	1.00	1.00	1.00
2	.898	.884	.871	.905	.894	.897
3	.673	.680	.693	.696	.667	.676
4	.502	.522	.541	.550	.470	.541
5	.403	.392	.368	.404	.293	.416
6	.287	.296	.269	.281	.258	.297
7	.282	.243	.249	<b>.268</b>	<b>.249</b>	.251
8	.292	.243	.256	.284	.253	.259
9	.287	<b>.238</b>	.240	.296	.260	<b>.238</b>
10	<b>.267</b>	.259	<b>.236</b>	.280	.269	.244

## Implementing Sympson Hetter in the Shadow Test Approach -

**TABLE 2**

Average percentage of violations of target exposure rate over  $\theta = -2.0, -1.5, \dots, 2.0$  at each iteration step for the six cases studied. (Note: first best result for each case is printed in bold)

No. of Shadow Tests	1			2		
Length of List	5	10	20	5	10	20
Iteration step						
1	3.13	3.08	3.22	3.13	3.13	3.00
2	3.04	3.23	3.00	2.89	2.48	2.97
3	2.61	2.79	2.82	2.89	2.81	2.38
4	2.67	2.42	2.32	<b>2.32</b>	2.36	2.69
5	2.85	2.38	2.55	2.69	2.95	<b>1.74</b>
6	2.72	2.38	2.46	2.70	2.33	2.94
7	2.72	2.30	<b>2.07</b>	2.48	<b>2.17</b>	1.82
8	2.98	2.36	2.52	2.85	2.29	2.48
9	<b>2.49</b>	<b>2.20</b>	2.10	2.42	2.42	2.11
10	2.73	2.20	2.38	2.63	2.29	1.99

**Figure Captions**

*Figure 1.* Conditional exposure rates at  $\theta = -2.0, -1.5, \dots, 2.0$ . The items are ordered by their rate.

*Figure 2.* Estimated bias function for the ability estimator with and without item-exposure control.

*Figure 3.* Estimated MSE function for the ability estimator with and without item-exposure control.

Figure 1.

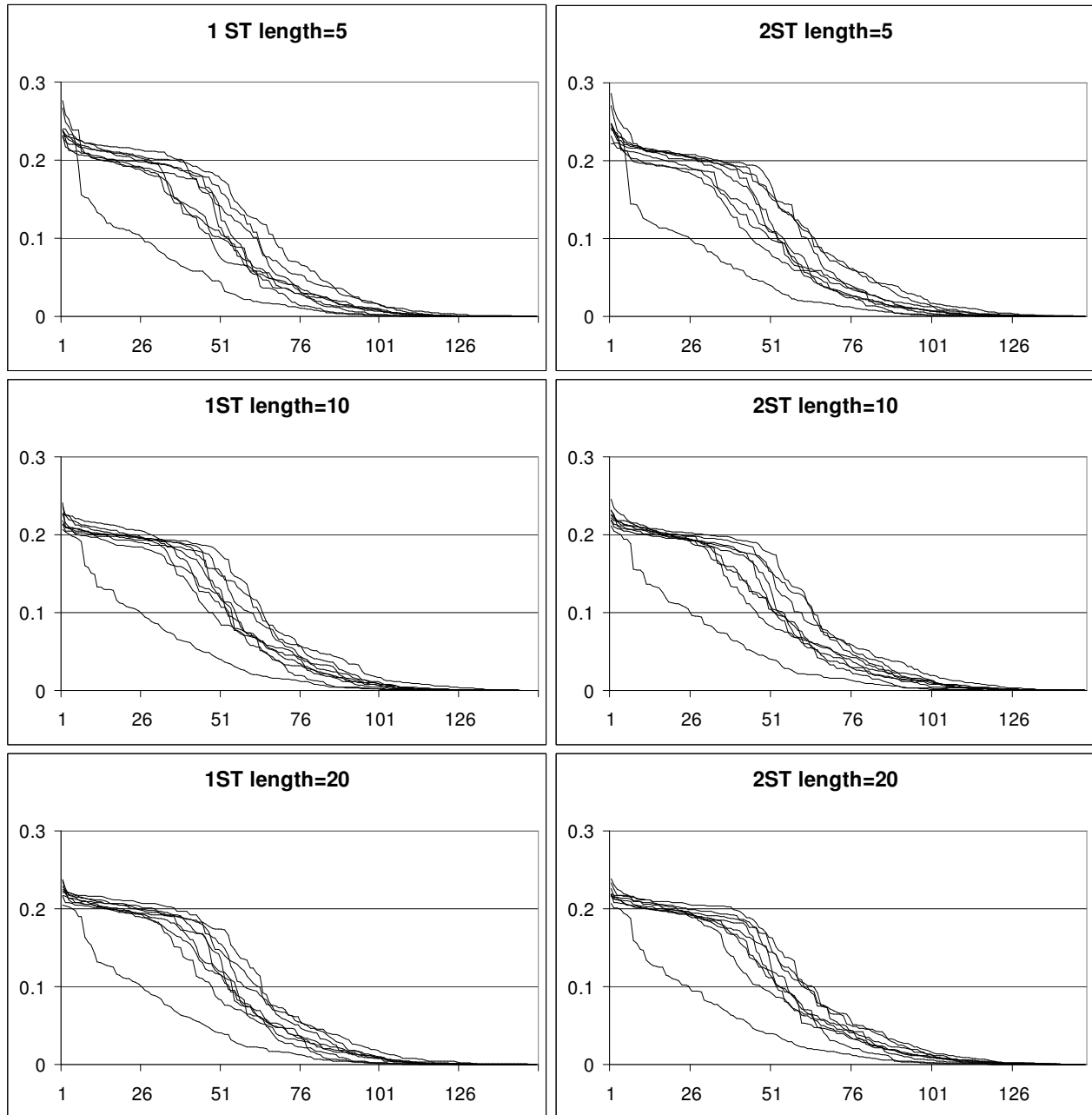


Figure 2.

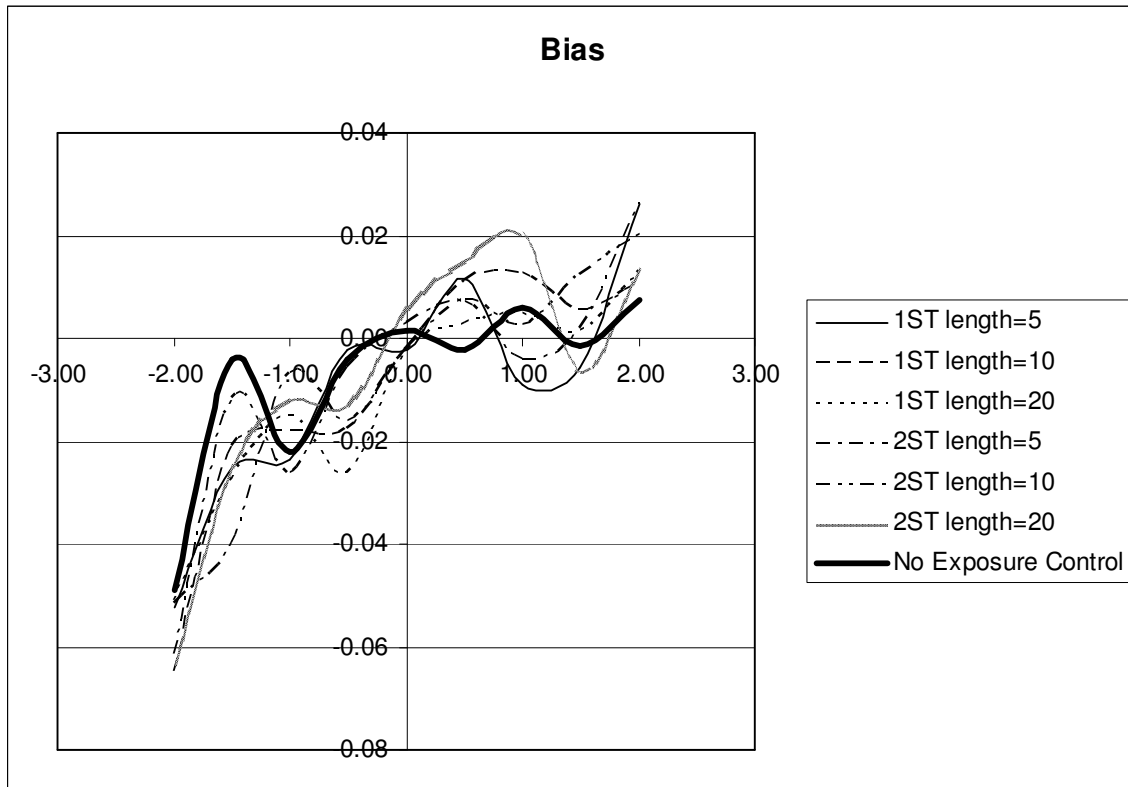


Figure 3.

