Constructing Rotating Item Pools for Constrained Adaptive Testing

Adelaide Ariel, Bernard P. Veldkamp, and Wim J. van der Linden[1)]

University of Twente

Enschede, The Netherlands

1) The paper was completed while this author was a Fellow at the Centre of Advanced Study in the Behavioral Sciences, Stanford, CA. He is indebted to the Spencer Foundation for a grant awarded to the Centre to support his Fellowship.

Abstract

Preventing items in adaptive testing from being over- or underexposed is one of the main problems in computerized adaptive testing. Though the problem of overexposed items can be solved using a probabilistic item-exposure control method, such methods are unable to deal with the problem of underexposed items. Using a system of rotating item pools, on the other hand, is a method that potentially solves both problems. In this method, a master pool is divided into (possibly overlapping) smaller item pools, which are required to have similar distributions of content and statistical attributes. These pools are rotated among the testing sites to realize desirable exposure rates for the items. In this paper, a test assembly model for the problem of dividing a master pool into a set of smaller pools is presented. The model was motivated by Gulliksen's (1950) matched random subtests method. Different methods to solve the model are proposed. An item pool from the Law School Admission Test (LSAT) was used to evaluate the performances of computerized adaptive tests from systems of rotating item pools constructed using these methods.

Key words: computerized adaptive testing, item pool design, matched random subtests method, mathematical programming, rotating item pools, test assembly.

Introduction

In paper-and-pencil (P&P) testing, the same set of items is administered to a population of examinees. A disadvantage of this testing format is its inability to deal with a broad range of abilities in the population. This disadvantage is remedied by computerized adaptive testing (CAT) where each examinee takes a test with items selected to match their ability estimates during the test. In doing so, CAT emulates an important aspect of oral examination, namely the practice of an examiner who chooses a more difficult question if the examinee responds correctly and an easier question if (s)he responds incorrectly. A CAT algorithm implements this practice by updating the examinee's ability estimate, $\hat{\theta}$, and choosing the next item to be optimal at this estimate.

Maximum-information and Bayesian item selection are commonly used criteria to select items (Hambleton, Swaminathan, & Rogers, 1991; van der Linden & Pashley, 2000). When an item is selected to maximize information at the current ability estimate, the algorithm prefers items for which both the difference between the current ability estimate $\hat{\theta}$ and the item difficulty parameter b_i is small and the item discrimination parameter a_i is high (Veerkamp & Berger, 1999). These items are of high quality but, unfortunately, for most item pools, their number is low. For a population of examinees, these items are selected more often and the high exposure rate of these items makes them vulnerable to security breaches. Typically, the other items are hardly selected at all and the resources invested in writing and calibrating them have futile effect (Veldkamp, 2001).

Various methods have been proposed to solve this problem, including methods of item-exposure rate control, item pool design, and rotating item pools. Sympson and Hetter (1985)

introduced a probabilistic approach to control the exposure rate of every item in the pool (for a review of this method, see van der Linden, in press). McBride, Wetzel and Hetter (1997) advocated an algorithm that reduced the exposure rates of items most popular at the beginning of the test. Veldkamp and van der Linden (2000) suggested to calculate a blueprint for the item pool such that distribution of the exposure rates for the items in the pool tends to be even. Stocking and Swanson (1998) introduced a method of rotating item pools with the same goal of even exposure rates. The item pools in this method are assembled from a master pool by splitting it into several smaller pools. The item pools are randomly rotated during operational testing. The goal of more uniformly distributed item-exposure rates is realized by assigning items with higher exposure rates to a smaller number of pools and items with lower rates to a larger number of pools. Unfortunately, studies of this method are rare. It is the purpose of this paper to propose some methods for assembling systems of rotating item pools from a given master pool and evaluate their properties.

In this paper, several methods to construct the system of rotating item pools are presented. All methods are based on a two-stage assignment process, in which items are first assigned to interim sets of (closely) parallel items and then from these sets to item pools. This process is optimized using the idea underlying the matched random subtests method introduced by Gulliksen (1950) to split a test into parallel subtests to the largest possible split-half reliability, which is always a lower bound to classical test reliability. Interestingly, the same idea can be generalized to the problem of assembling a system of rotating items pools. To illustrate how to put our methods into practice, several examples using an item pool from the Law School Admission Test (LSAT) are given. A constrained CAT algorithm was applied to evaluate the performance of the systems of rotating item pools constructed by these methods.

Current Methods to Construct Rotating Item Pools

Way (1998) observes that probabilistic item-exposure control may be inadequate to guarantee a secure CAT program. His suggestion is that a system of rotating item pools may be a more promising approach to prevent item compromise. Way, Steffen, and Anderson (1998) discuss several examples of strategies of managing rotating item pools and using such systems to enhance the security of the computerized testing.

For systems of rotating item pools to be effective, the availability of automated item-selection procedures to assemble such pools from a master pool is crucial. These procedures should have the following properties: First, each pool should have similar distributions of content and statistical attributes to support uniform measurement quality to examinees. Second, the composition of the pools should support uniform usage of the items. Third, the pools should have enough items to allow item selection for the adaptive tests to be constrained with respect to all the specifications to be imposed on the test.

Stocking and Swanson (1998) demonstrate a method to construct rotating item pools. In their method, the items in the master pool are assigned to pools by their weighted deviations model (WDM). The first step in this method is to calculate an average (nonoverlapping) pool from the master pool. The actual pools are then assembled to minimize the differences between these pools and the average pools with respect to (1) the item attributes in the constraints to be imposed on the CAT and (2) item information at selected θ levels. In doing so, the deviations are weighed to get a single objective function. The resulting pools are expected to have similar distributions of content and statistical attributes and, hence, to support each test administration equally well.

Two versions of the Stocking-Swanson method exists, one with nonoverlapping and another with overlapping item pools. The former was presented above. However, to get more uniformly distributed exposure rates, a system of overlapping item pools is more efficient. This system allows us to reduce the exposure rate of more popular items by assigning them to a smaller set of pools and to increase the usage of less popular items by assigning them to a larger set. Overlapping pools are assembled by first calculating the required numbers of pools the items should figure in and then assigning the items to the pools until these numbers are realized. For empirical results with these methods, see Stocking and Swanson (1998).

New Methods for Constructing Rotating Item Pools

The new methods proposed for constructing rotating item pools are all based on techniques of constrained combinatorial optimization. The objective functions in the optimization problems focus on the values of the item parameters in the pools. The goal is to give the pools identical distributions of parameters. At the same time, constraints are introduced to match the pools in terms of content attributes and to control the overlap between the pools.

All methods were motivated by Gulliksen's (1950) matched random subtests method. Gulliksen's method was proposed to split a test into two halves that are statistically as closely parallel as possible. The split-half reliability calculated from these halves is a lower bound to the classical test reliability. Gulliksen's method has two stages. In the first stage, the items are assigned to pairs of items that have minimal differences between their parameter values. In the second stage, the items are assigned to test halves. A formalization of Gulliksen's method as a problem of constrained combinatorial optimization is given in van der Linden and Boekkooi-Timminga (1988).

We apply the same method to the problem of splitting a master pool into a set of smaller pools for use as rotating item pools in CAT. To illustrate the method, its stages for the case of nonoverlapping pools are briefly described: First, interim sets of items are assembled, each of a size equal to the number of (nonoverlapping) pools to be constructed. Second, items in the same interim set are assigned to different pools. If the composition of these pools is required to meet certain constraints to support CAT, the assignment is subject to these constraints.

Figure 1 illustrates the general process of dividing a master pool into four nonoverlapping pools. In this figure, to get four nonoverlapping pools, the *n* items in the master pool are assigned to *n*/4 interim sets, each consisting of four different items.

Insert Figure 1 about here

The two stages are explained in more detail as follows:

Stage 1: Assigning Items to Interim Sets

In the first stage, the items in the master pool are assigned to interim sets. For notational simplicity, we formulate the optimization problem for interim sets consisting of two items. Generalization to larger interim sets is straightforward.

A metric δ_{ij} is used to represent the differences between items *i* and *j* in interim sets. If the goal is to minimize the differences between the values of the items for the *a* and *b* parameters in the sets, a possible metric for these differences is

$$\delta_{ij} = \left|a_i - a_j\right| + w\left|b_i - b_j\right| \qquad (1)$$

where *w* is a parameter that can be used to correct for differences between the scales of the two parameters. Other types of metric and larger numbers of item parameters are possible.

The problem of assigning items to interim sets can be formulated as a 0-1 mathematical programming problem with decision variables, x_{ij}, $i \neq j = 1,...,I$, which are equal to 1 if item *i* and *j* are chosen in the same set and are equal to 0 otherwise, where *I* represents the number of items in the master pool. The objective function is

$$min \sum_{i=1}^{I-1} \sum_{j=i+1}^{I} \delta_{ij} x_{ij} \tag{2}$$

subject to

$$\sum_{\substack{i \\ i<j}} x_{ij} + \sum_{\substack{i \\ i>j}} x_{ji} = 1, \forall j \tag{3}$$

The constraint in (3) is to guarantee that every item is assigned to an interim set only once (van der Linden and Boekkooi-Timminga, 1988).

Stage 2: Assigning Items to Pools

In the second stage, the items in the interim sets are assigned to the item pools. Different assignment models for the assignment of items to nonoverlapping and overlapping pools are formulated.

Nonoverlapping pools. The general idea in generating the item pools is to make them as similar as possible. In the mathematical programming model below, the objective function

minimizes the differences between the total information in the pools at several ability values, while constraints are introduced to assign every item exactly once.

The model can be formulated as

$$min\ z \tag{4}$$

subject to

$$\sum_i I_i(\theta_k)y_{is} - \sum_j I_j(\theta_k)y_{jp} \le z,\ \forall\theta_k, s, p, s \ne p, i \ne j \tag{5}$$

$$\sum_i I_i(\theta_k)y_{is} - \sum_j I_j(\theta_k)y_{jp} \ge -z,\ \forall\theta_k, s, p, s \ne p, i \ne j \tag{6}$$

$$\sum_{i \in Q_r} y_{is} = 1, \forall s \tag{7}$$

$$\sum_s y_{is} = 1, \forall i \tag{8}$$

$$y_{is} \in \{0,1\} \tag{9}$$

where i and j are indices for the items, Q_r is the rth interim set, s and p indicate item pools, θ_k is ability level k, $I_i(\theta_k)$ is the information about θ_k in item i, and y_{is} is a decision variable that is equal to one if item i assigned to pool s, and equal to zero otherwise.

In (5)-(6) the difference between the total information in the item pools at the ability values θ_k are constrained to be in the interval (-*z*, *z*). The size of this interval is minimized in (4). The constraints in (7) require items in the same interim set to be assigned to different pools. The constraints in (8) guarantee that all items are assigned once, whereas the constraints in (9) define the decision variables to be 0-1. This model can be modified to allow for the case of overlapping pools. Moreover, the model can be extended to allow for additional constraints on the test required to deal with such issues as test content and word count. We will address these topics below.

Overlapping pools. Overlapping item pools are obtained by changing the constraint on number of times an item is assigned to a pool in (8). This constraint is then replaced by

$$\sum_s y_{is} \leq n^{(r)}, \forall i \tag{10}$$

$$\sum_s y_{is} \geq n_r, \forall i \tag{11}$$

with $n^{(r)}$ and n_r denoting the maximum and minimum number of item replications admitted, respectively. Unpopular items should have larger values $n^{(r)}$ and n_r, and popular items should have lower values.

Whether or not an item is popular is usually known only after conducting a CAT simulation. Based on items performance in the CAT algorithm, we then can decide what values $n^{(r)}$ and n_r should have. However, we can also decide on these values using the values of the items for the discrimination parameter (see empirical examples below).

Additional content constraints. The model previously described does not allow for possible additional content constraints on the CAT yet. Suppose, for example, that the CAT has to be constrained with respect to item type and word counts. In order to ensure that the pools have comparable distributions of these item attributes, the model in Stage 2 has to be extended with the following constraints:

$$\sum_{i \in V_m} y_{is} \geq n_m, \forall s \tag{12}$$

$$\sum_{i \in V_m} y_{is} \leq n^{(m)}, \forall s \tag{13}$$

$$\sum_{i} w_i y_{is} \geq n_w, \forall s \tag{14}$$

$$\sum_{i} w_i y_{is} \leq n^{(w)}, \forall s \tag{15}$$

where V_m is the set of items of type m in the master pool, w_i is the word count of item i, n_m and n_w are the lower bounds on the number of items of type m and the word count in the test, respectively, and $n^{(m)}$ and $n^{(w)}$ are the upper bounds on these attributes.

Algorithms for Solving the Models

Various algorithms for solving the previous two types of models are presented. We first discuss the methods for assigning items to interim sets.

Assigning Items to Interim Sets

For a system of rotating nonoverlapping item pools, the number of parallel items in every interim set is equal to the number of pools to be created. Various methods for assigning items to interim sets are given.

Sequential assignment. A simple method would be to use a greedy heuristic. This heuristic has been applied to test assembly problems in combination with the Weighted Deviation Model (Stocking & Swanson, 1993) and in combination with the Normalized Weighted Absolute Deviation Heuristic (Luecht, 1998). When this heuristic is applied to solve the model, the interim sets are constructed sequentially, that is, items with the smallest value for (1) are selected until all items have been used. It is obvious that this approach is easy to implement, but does have some drawbacks. For each subsequent item, the value of (1) will be larger and the quality of the interim sets will therefore deteriorate. Therefore a method based on simultaneous assembly of all interim sets is proposed.

Simultaneous assignment. If the number of pools to be assembled increases, the combinatorial optimization problem involved in their assembly becomes larger and we may soon reach a point at which a heuristic is needed to solve the model. In the empirical example below, we used a heuristic method known as simulated annealing. Simulated annealing is an iterative Monte Carlo method. At each step a candidate for a new solution is generated by randomly modifying the previous solution, after which it is decided whether this candidate is or is not accepted. An attractive feature of simulated annealing is that it accepts a worse solution with a nonzero probability. Typically, the probability becomes smaller after some iterations. This feature allows the algorithm to backtrack if it gets stuck in a bad path. The algorithm is terminated as soon as no further improvement can be made. Examples of an earlier applications

of this heuristic to test assembly problems are given in van der Linden, Veldkamp and Carlson (submitted), and Veldkamp (1999). A more theoretical explanation of simulated annealing can be found, for example, in van Laarhoven and Aarts (1987).

A simple but effective implementation of simulated annealing for the current problem is to generate new interim sets from randomly selected old sets. We initialized the algorithm solution by assigning *all* items to *all* interim sets. This initialization is possible because the final solution does not depend on the initial solution. (If the number of items in the master pool is not a multiplicity of the number of pools, a dummy interim set has to be created to which the remaining items are assigned.) Candidate interim sets were then constructed by randomly swapping items between sets. The metric in (2) was used to evaluate the candidate sets against the old sets. If the new value for the objective function was smaller, the candidate sets were accepted with probability one. If it was larger, the decision was made with a probability. At the next iteration step, new candidates were generated from the current interim sets.

Assigning Items to Pools

Two methods for assignment of items from interim sets to item pools are discussed.

Random assignment method. Because in the first stage of the method the interim sets are constructed to be as similar as possible, it may be possible to assign items in the same interim set randomly to item pools. If items in all interim sets are assigned, we may still have a dummy set that had to be created because the number of items in the master pool was not a multiplicity of the number of pools. The items in this set can also be assigned randomly. However, in the empirical example below, the information functions for each pool was calculated and the items were assigned to the pools with the smallest values for these functions.

Mathematical programming. A more precise method is to solve the model in (4) until (9) and the additional constraints in (12) and (14) using a mathematical programming algorithm. Use of such algorithms is recommended especially when sequential assignment results in quick deterioration of the value of the objective function for the interim sets. In the empirical study below we used one of the standard algorithms in the mathematical programming software packages AIMMS (2001).

Choosing the Number of Overlapping Pools

If overlapping pools are to be constructed, the bounds on the number of times an item can be assigned, $n^{(r)}$ and n_r, have to be specified for each item. The choice of these numbers can either be based on the values of the item parameters or on the actual exposure rates of the items. The first method is based on the discrimination parameter of the items. Since items with higher a_i values have a larger chance of being selected, such items should be given low values for $n^{(r)}$ and n_r. On the other hand, items with low a_i have a low probability of being selected and should be given higher values for $n^{(r)}$ and n_r. The second method is based on the actual exposure rates of the items. Items with high exposure rates are given low values for both $n^{(r)}$ and n_r, while items with low exposure rates are given high values.

Empirical Example

A previous pool of 2,131 items from the LSAT fitting the 3-parameter logistic model (Hambleton, Swaminathan & Rogers, 1991) was used as the master pool. Figure 2 shows the distribution of the values for the a_i and b_i parameters for the items in the pool. We ignored the c_i parameter because its variability was small and this parameter typically does hardly have any

impact on the composition of the item pools. The items in the master pool were of nine different types, leading to 20 content constraints. Because word counts of the items were available we used constraints to match the word counts of the pools as well. All constraints appeared relatively easy to satisfy. In this study, we first classified the items in the master pool based on their type and then solved the (extended) model in (1)-(3) per item type. The benefit of this approach was that the model in (4)-(15) reduced proportionally and (12)-(13) could be ignored. Observed that when the number of item of a certain type was small, the number of nonoverlapping pools produced would be limited since each pool had to satisfy (12)-(13).

Insert Figure 2 about here

Following Stocking's (1994) recommendation that the size of a CAT item pool size be at least 12 times the length of the adaptive test (which was 24 in our study), at most four nonoverlapping pools could be produced. The same restriction did not apply for the case of overlapping pools, however, because it allowed for the duplication of some items to meet (12)-(13). As a result, we were able to assemble one system of nonoverlapping pools with four pools and two different systems of overlapping pools with six and eight pools.

All item pools in this study were assembled using an objective function based on the metric in (1), with $w = 1$. All results were evaluated through a CAT simulation study. The CAT algorithm in these studies was based on a constrained CAT with shadow tests approach (van der Linden, 2000). In this approach, prior to the selection of an item first a full test meeting all constraints is assembled. From this test, the most informative item is administered to the examinee. After the ability estimate is updated, the shadow test is reassembled keeping all items

already administered. The process continues until the required number of items administered is reached. The shadow tests were calculated using the AIMMS (2001) optimization software package. Test length was fixed at 24 items and the values of the examinees for the ability parameter were randomly drawn from the standard normal distribution. Abilities were estimated by the method of maximum likelihood estimation (MLE). As long as the simulated responses for an examinee were all correct or all incorrect, the ability estimate for the examinee was set equal to 3 and -3, respectively. In each condition, 1,000 examinees were simulated and ability estimation was always initialized at $\hat{\theta} = 0$.

To evaluate the performances of the methods, the exposure rates of the items and the bias and mean squared error (MSE) functions for the ability estimators were calculated. No method of probabilistic item-exposure control was used; the effects of such methods would have confounded the evaluation of impact of the pool assembly methods on the exposure rates of the items. In a practical application, however, the use of additional exposure control may be necessary.

Nonoverlapping Item Pools

The performances of all four possible combinations of the methods of sequential and simultaneous assignment of items to interim sets and random assignment and mathematical programming were compared. Each combination was evaluated using CAT simulations. The simulations were based on the full systems of rotating pools, each with four nonoverlapping pools. Because each pool had approximately 500 items, each evaluation was based on approximately 2,000 items. In addition, CAT administrations directly from the master pool were simulated. The results for CAT from the master pool served as a point of reference for our

evaluation of the results in the other four conditions. In the mathematical programming method, item information was controlled at $\hat{\theta} = -1, 0, \text{and } 1$.

Figure 3 shows the item exposure rates for CAT from all item pools assembled by the four combinations of methods as well as directly from the master pool. For each method, the rates of only 600 of

---

Insert Figure 3 about here

---

the items (out of the total of 2,000 items) are given; the rates of all other items were equal to zero. The highest exposure rate for an item in CAT from nonoverlapping pools was less than 0.3. However, for CAT from the master pool the highest rate was equal to 1.0. Figure 3 also reveals that the number of items with nonzero rates is larger for CAT from nonoverlapping pools. These results show that, compared with CAT from the master pool, CAT from nonoverlapping pools improved the exposure rates of the items equally well for all four methods used to construct these pools.

Figure 4 summarizes the errors in the ability estimates for all five conditions. Both the bias and the MSE functions are lower for CAT from the master pool than from nonoverlapping pools. The reason is the fact that during CAT from the master pool all items were available for selection for each of the examinees where with CAT from the nonoverlapping pools each item had to selected from a smaller set. For all practical purposes, the differences are small though. The differences between the functions for the four conditions with rotating item pools were also negligible.

Insert Figure 4 about here

Overlapping Item Pools

Unlike the previous study, only the second stage was used to construct overlapping pools. The upper and lower bounds $n^{(r)}$ and n_r to the overlap in (10)-(11) were determined using the two methods presented below. For each method, systems of six and eight overlapping item pools were created from the master pool.

The first method used to determine the upper and lower bounds on the item overlap between pools was based on the values of the items in the master pool for the discrimination parameter. These values were in the interval [0.2,1.7]. This interval was divided into equally wide intervals, six and eight intervals for the cases of six and eight overlapping pools, respectively. Items with the highest value for the a_i parameter (the first interval) were assigned only once ($n^{(r)}$ and n_r are equal to 1), items in the second interval twice ($n^{(r)}$ and n_r are equal to 2), etc.

The second method was based on the empirical exposure rates of the items. Using the exposure rates in the first study, the items were divided into two sets. The criterion was an exposure rate below or above 0.5. The items with larger exposure rates were assigned only once, the items with smaller rates were assigned to all item pools.

The estimated exposure rates as well as the bias and MSE functions for both methods are given in Figure 5-7, respectively. For items that were assigned to multiple pools, the exposure rates were cumulated across pools. In Figure 5, however, only items with nonzero exposure rates are displayed. The rates were slightly more favourable than in Figure 3 but showed the same

pattern. Also, both for the case of six and eight overlapping pools, the method for setting the bounds $n^{(r)}$ and n_r based on empirical exposure rates outperformed the method based on the a_i parameter. Also, the differences between the two methods were larger for the case of eight than six pools. However, as Figures 6 and 7 shown, these more favourable exposure rates were obtained at the costs of slightly higher bias and MSE.

Insert Figure 5 - 7 about here

The differences in results between the methods for setting bounds on item overlap follow directly from the criterion on which they were based. In the method based on empirical exposure rates, every pool contains limited number of good items because these items are assigned to only one pool. As a result, that CAT algorithm is forced to choose considerable numbers of worse items and the errors in the ability estimates increase. For the method based on a_i values, the items in the highest category are the only ones assigned only once. Because the item pools have larger numbers of good items, the estimation errors are smaller but the worse items remain hardly used at all.

Discussion

The methods for constructing item pools presented in this paper are intended to optimize or to maximize the use of test items in CAT. It was shown that CAT from rotating item pools that do not overlap can improve the usage of items. As shown in Figure 3, the item exposure rates for CAT from nonoverlapping item pools were substantially better than for CAT directly from the master pool. The differences found between the exposure rates and statistical quality of

the ability estimates of the four combinations of methods used to assemble these pools were generally negligible, though. This finding seems to suggest that best strategy in real-life applications is to choose a method from these four that is easy to implement under local constraints.

Though the use of nonoverlapping item pools did not dramatically increase the exposure rates of the less frequently items, the use of overlapping pools was more successful in this respect. Especially the results obtained when the method for setting the bounds $n^{(r)}$ and n_r was based on empirical exposure rate were promising. Key to these results seems to be the fact that larger numbers of popular items were not allowed to be assigned to more than one pool. As a result, the CAT algorithm was forced to select larger numbers of less popular items, which thus obtained larger exposure rates.

The experimental results also showed that improving the exposure rates of the items tends to results in larger errors for the ability estimates. This trade-off should not come as a surprise: More uniform item exposure rates can only be obtained by imposing more severe constraints on the item selection. These constraints result in a lower value for the objective function optimized during item selection, that is, in less information about the examinees' abilities in the test. However, it is the opinion of the authors that the size of the increase of these errors in the present studies was still acceptably low and could easily have been compensated, for example, by a small increase in the length of the test. The best possible way to deal with these errors is topic of further research.

The empirical study shows that implementation of the methods proposed in this paper is straightforward. In theory, the methods can be applied to construct systems with large numbers of rotating item pools. The only possible obstacle is limited availability of some types of items in

the master pool. Though it seems attractive to deal with this obstacle by duplicating such items and assigning them to overlapping pools, and there is virtually no bound on the number of overlapping pools possible, we should be aware of the possible adverse effects of this measure on the exposure rates of some of the items as well as a possible increase in measurement error.

References

Gulliksen, H. (1950). *Theory of mental tests*. New York: Willey.

Hambleton, R.K., Swaminathan H., & Rogers, H.J. (1991). *Fundamentals of item response theory*. California, USA: Sage Publications.

McBride, J.R., Wetzel, C.D., & Hetter, R.D. (1997). Preliminary psychometric research for CAT-ASVAB: selecting an adaptive testing strategy. In W.A. Sands, B.K. Waters, & J.R. McBride (Eds.), *Computerized adaptive testing: from inquiry to operation* (pp. 83-95). Washington DC: American Psychological Association.

Paragon BV. (2001). AIMMS (Version 3.2) [Computer Software]. Haarlem, The Netherlands: Paragon BV.

Stocking, M.L. (1994). *Three practical issues for modern adaptive testing item pools*. ETS Research Report No. 93-2. Princeton, NJ: Educational Testing Service.

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

Swanson, L., & Stocking, M.L. (1993). A model and heuristic for solving very large item selection problems. *Applied Psychological Measurement, 17*, 151-166.

Sympson, J.B., & Hetter, R.D. (1993). 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., & Boekkooi-Timminga, E. (1988). A zero one programming approach to Gulliksen's matched random subtests method. *Applied Psychological Measurement, 12*, 201-209.

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). The Netherlands: Kluwer Academic Publishers.

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

van der Linden, W. J., Veldkamp, B. P., & Carlson, J. E. (submitted). *Optimizing balanced incomplete block designs for educational assessments*.

van Laarhoven, P.J.M., & Aarts, E.H.L. (1987). *Simulated annealing: theory and applications*. Dordrecht: Reidel.

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

Veldkamp, B.P. (2001). *Principles and methods of constrained test assembly*. Thesis, Enschede, The Netherlands: University of Twente.

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

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). The Netherlands: Kluwer Academic Publishers.

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, Philadelpia, PA.

Figure Captions

Figure 1. An illustration for constructing four nonoverlapping pools from a given master pool.

Figure 2. Scatter plot of item parameters a and b.

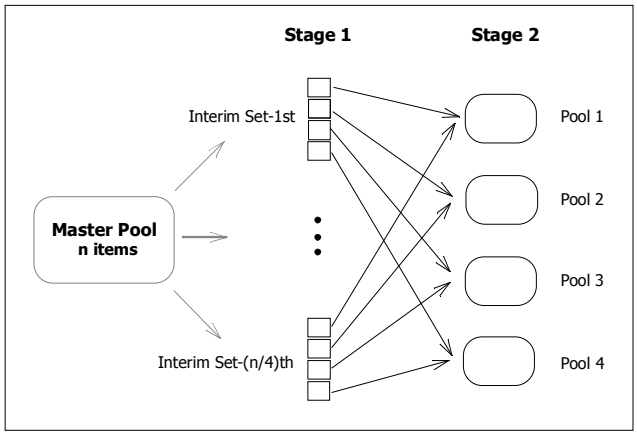
Figure 3. Item-exposure rates for CAT from master pool and nonoverlapping pools.

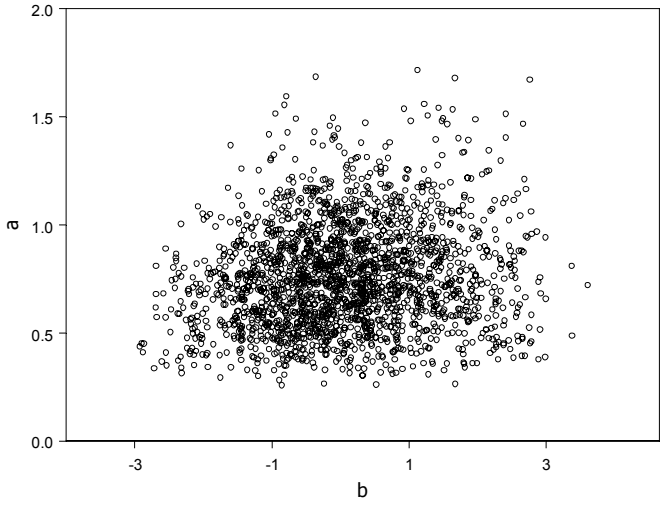
Figure 4. Bias and MSE functions for CAT from master pool and nonoverlapping pools.

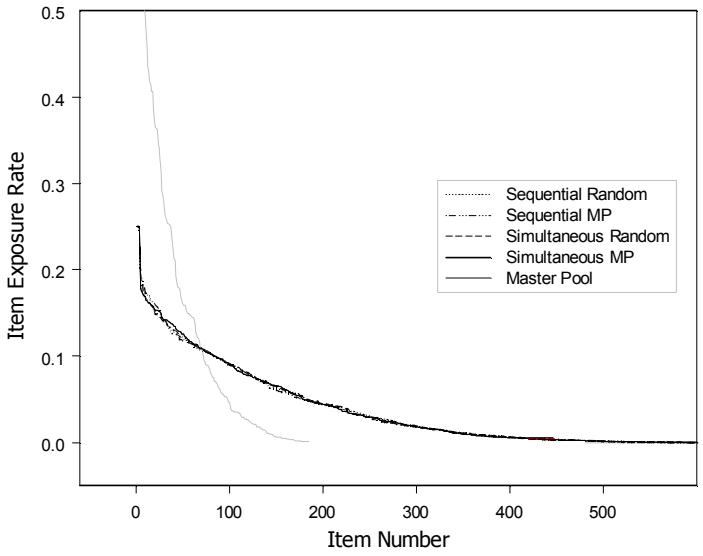
Figure 5. Items-exposure rates for CAT from master pool and overlapping pools.

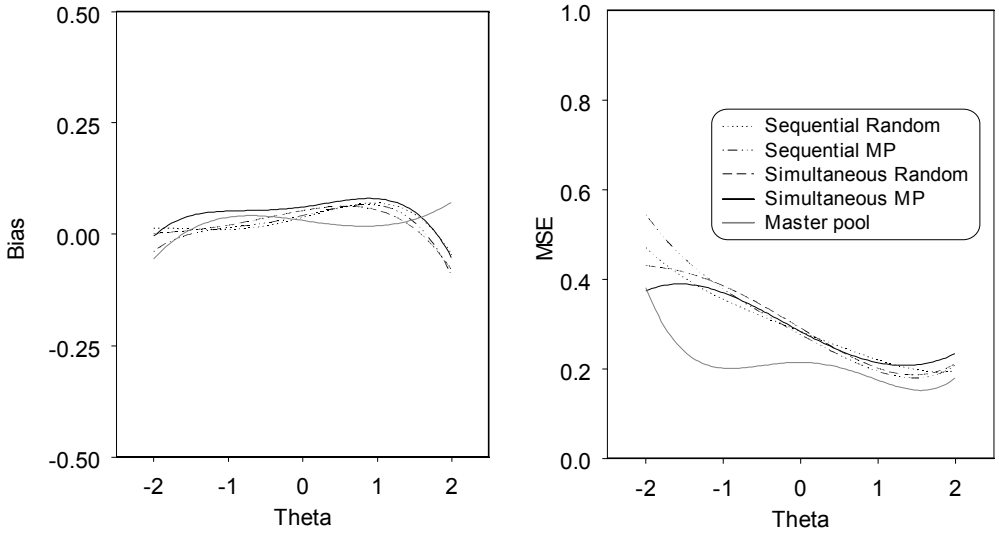
Figure 6. Bias functions for CAT from master pool and overlapping pools.

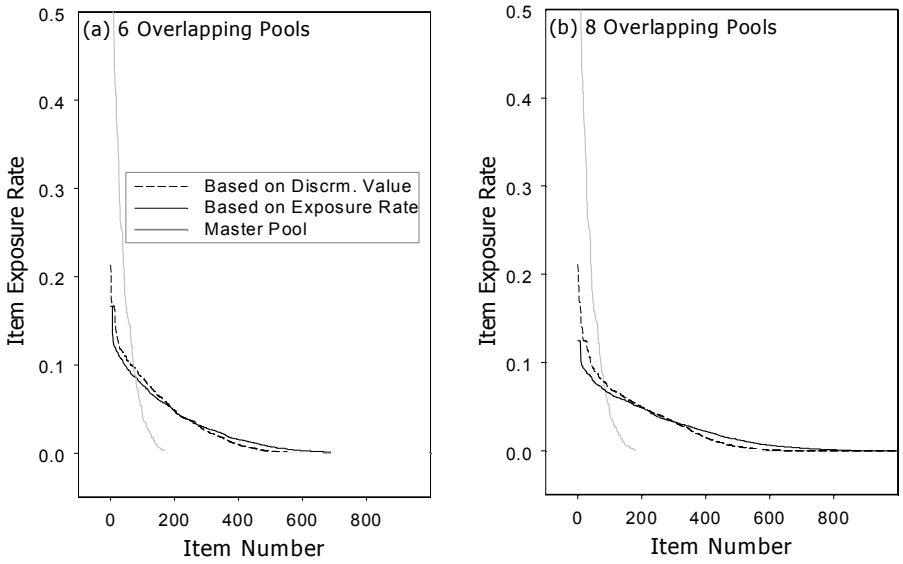
Figure 7. MSE functions for CAT from master pool and overlapping pools.

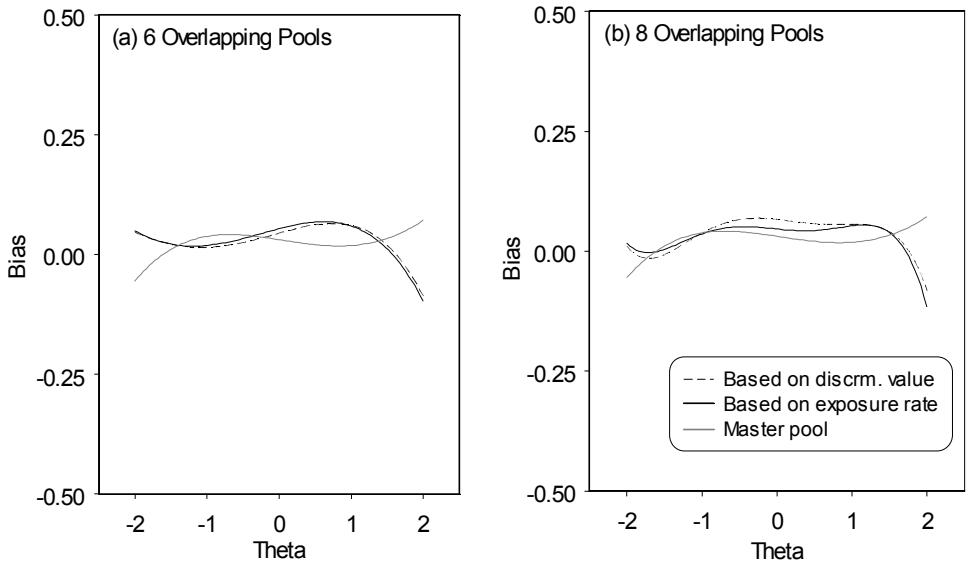

Stage 1
Stage 2
Interim Set-1st
Master Pool
n items
Interim Set-(n/4)th
Pool 1
Pool 2
Pool 3
Pool 4

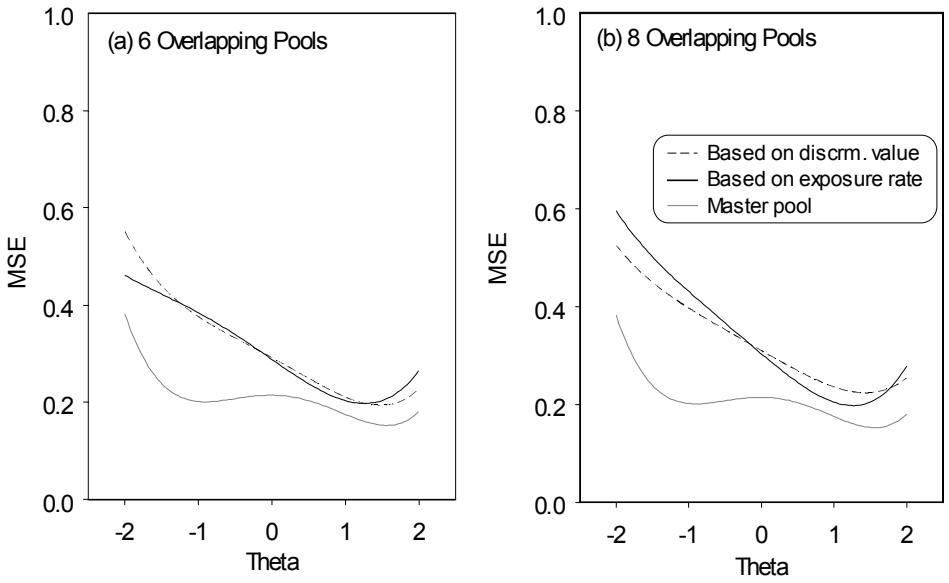

2.0
1.5
1.0
0.5
0.0
a
-3
-1
1
3
b


0.5
0.4
0.3
0.2
0.1
0.0
Item Exposure Rate
0
100
200
300
400
500
Item Number
Sequential Random
Sequential MP
Simultaneous Random
Simultaneous MP
Master Pool


0.50
0.25
0.00
-0.25
-0.50
Bias
-2
-1
0
1
2
Theta
1.0
0.8
0.6
0.4
0.2
0.0
MSE
Sequential Random
Sequential MP
Simultaneous Random
Simultaneous MP
Master pool
-2
-1
0
1
2
Theta


(a) 6 Overlapping Pools
(b) 8 Overlapping Pools
Based on Discrm. Value
Based on Exposure Rate
Master Pool
Item Exposure Rate
Item Number
0.0
0.1
0.2
0.3
0.4
0.5
0
200
400
600
800


(a) 6 Overlapping Pools
(b) 8 Overlapping Pools
Bias
Theta
0.50
0.25
0.00
-0.25
-0.50
-2
-1
0
1
2
Based on discrm. value
Based on exposure rate
Master pool


(a) 6 Overlapping Pools
(b) 8 Overlapping Pools
MSE
Theta
1.0
0.8
0.6
0.4
0.2
0.0
-2
-1
0
1
2
Based on discrm. value
Based on exposure rate
Master pool