Assembling a Computerized Adaptive Testing Item Pool as a Set of Linear Tests

Wim J. van der Linden
Adelaide Ariel
Bernard P. Veldkamp
University of Twente

Test-item writing efforts typically result in item pools with an undesirable correlational structure between the content attributes of the items and their statistical information. If such pools are used in computerized adaptive testing (CAT), the algorithm may be forced to select items with less than optimal information, that violate the content constraints, and/or have unfavorable exposure rates. Although at first sight somewhat counterintuitive, it is shown that if the CAT pool is assembled as a set of linear test forms, undesirable correlations can be broken down effectively. It is proposed to assemble such pools using a mixed integer programming model with constraints that guarantee that each test meets all content specifications and an objective function that requires them to have maximal information at a well-chosen set of ability values. An empirical example with a previous master pool from the Law School Admission Test (LSAT) yielded a CAT with nearly uniform bias and mean-squared error functions for the ability estimator and item-exposure rates that satisfied the target for all items in the pool.

Keywords: assembling multiple test forms, computerized adaptive testing, item-pool design, mixed integer programming, shadow-test approach

Ideally, an adaptive testing program should offer each examinee a short test that is informative at his or her ability level and meets all content specifications. Also, particularly in high-stakes testing, to avoid risking the security of its items and in view of the costs involved in item-pool regeneration, the selection of the items should prevent overexposure of some of them. Item-selection algorithms have been developed to realize these goals in the best possible way. Current algorithms typically select items that maximize Fisher’s information in the test or minimize the posterior variance of the examinee’s ability (van der Linden & Pashley, 2000). To balance test content across examinees, they spiral item selection over subsets of item in the pool (Kingsbury & Zara, 1991), minimize weighted deviations from bounds on content categories (Swanson & Stocking, 1993), or realize constraints on test

This article was written while the first author was a Fellow of the Center for Advanced Study in the Behavioral Sciences, Stanford, CA. He is indebted to the Spencer Foundation for a grant awarded to the Center to support his Fellowship.
content through a shadow test approach (van der Linden, 2000). Also, a powerful method to control item-exposure rates is by adopting a random component in the selection procedure (Sympson & Hetter, 1985; van der Linden & Veldkamp, 2004).

Nevertheless, although these methods are potentially powerful, their actual success is ultimately constrained by the composition of the item pool. For example, ability estimation may be less accurate than anticipated if several of the most informative items in the pool happen to have attributes that are excluded by some of the content specifications. Also, the items in the pool may have the majority of the required content attributes but several of them may miss one or two important attributes. If such cases occur, the item pool looses its efficiency because the computerized adaptive testing (CAT) algorithm has to compromise between the requirements to select tests that meet all content specifications, have high information at the examinees’ ability level, and do not overexposure a small set of items.

**Basic Idea**

Intuitively, an optimal item pool would consist of a maximal number of combinations of items that (a) meet all content specifications for the test and (b) are most informative at a series of ability levels reflecting the shape of the distribution of the ability estimates for a population of examinees. Observe that the second requirement is not only necessary to make the adaptive test informative for each examinee but also to yield exposure rates for the items in the pool that are below a realistic target value.

To optimize the composition of item pools for use in CAT, the idea of an item pool consisting of a set of fixed tests optimal at a distribution of ability levels reflecting the population of examinees is investigated. Although seemingly a step backward to more rigid traditional group-based paper-and-pencil testing, it is instructive to compare these two requirements with the practice of assembling linear test forms in testing programs that have such a format. In such programs, rather than assembling one individual test form at a time, it is common to assemble a set of parallel test forms from a master pool at the beginning of a new planning period to serve all of its testing windows and locations. If the program is item response theory (IRT) based, these forms are parallel in that each of them meets the same content specifications and the same target for the test information function. The set of these forms can be viewed as the operational “item pool” during the planning period of the program.

If the same pool of items were to be used for a CAT version of the program, its composition as a set of tests would guarantee the pool to have a sufficiently large number of combinations of items meeting the content specifications. In fact, because the CAT algorithm would mix and match between the different tests, the actual number is much larger than the number of individual tests in the pool. However, because these combinations would tend to be optimal at ability levels near the center of the population distribution, the CAT algorithm would only be able to produce ability estimates with the same accuracy for examinees toward the tails of the pop-
ulation distribution if it compromised the content of the test and/or the exposure rates of some of the items in the pool. Intuitively, the need to compromise would be minimized if the set of individual tests in the pool kept its composition but had maximum information at a distribution of ability levels approximating the shape of the population of examinees.

The idea of an item pool consisting of a set of fixed tests optimal at a distribution of ability levels reflecting the population of examinees is elaborated for the case of CAT based on the shadow-test approach (van der Linden, 2000; van der Linden & Reese, 1998). The same type of CAT pool assembly is possible in combination with any other type of CAT algorithm, but the basic ideas are better explained and motivated for the shadow-test approach because this is also based on the use of linear tests. We expect the benefits of the type of pool to be similar for the two other main CAT algorithms, the weighted-deviation method (Swanson & Stocking, 1993) and the method with spiraling over content categories (Kingsbury & Zara, 1991). These methods will also find it easier to compromise between the maximum-information criterion and the content specification along the ability continuum for a pool assembled using the proposed method.

**Shadow-Test Approach**

In the shadow-test approach, the items in the CAT are selected not from the pool but from shadow tests assembled before the selection of each new item. Shadow tests are linear tests that (a) meet all of the content specifications for the adaptive test, (b) contain all items already administered to the examinee, and (c) are optimal at the examinee’s current ability estimate. The item that is actually administered is the most informative item in the shadow test among those that the examinee has not seen before. Adaptive testing with shadow tests guarantees test administrations that always meet all content specifications (i.e., are feasible with respect to the content constraints). Also, at each step the item is selected to be optimal for ability estimation (subject to these content constraints). The assembly of shadow tests in real time has become possible by the speed of modern computers and the use of software packages for linear integer programming (IP), such as the CPLEX package used in the empirical example later in this article (ILOG, 2002).

The process of CAT with shadow tests is illustrated graphically for a fictitious examinee in Figure 1. As usual, \( \theta \) denotes the ability parameter in the IRT model, for example, the three-parameter logistic (3PL) model in Equation 1, and \( \hat{\theta} \) is an estimate of the examinee’s value for the ability parameter. The first shadow test is assembled to be optimal at the initial ability estimate, \( \hat{\theta}_0 \). The first item administered is the item in the shadow test optimal at \( \hat{\theta}_0 \). After the response to the first item, the ability estimate is updated, and the second shadow test is assembled to be optimal at this updated ability estimate, \( \hat{\theta}_1 \), retaining the item that was already administered. The process is continued until the last free item in the shadow test, which was selected to be optimal at \( \hat{\theta}_n \), is administered. Observe that if the CAT were administered from a pool of linear tests, the set of shadow tests in Figure 1 would not be a subset of the
tests in the pool. For one thing, the tests in the pool are disjoint, whereas the set of shadow tests shows overlap. More importantly, the set of shadow tests would be the result of the process of mixing and matching between the individual tests in the item pool by the CAT algorithm discussed earlier. It is exactly this process that is expected to lead to less violation of the target for item-exposure rates and/or less loss of accuracy of ability estimation if the distribution of ability values at which the tests in the pool are optimal reflects the population of examinees.

The process in Figure 1 allows us to make the notion of “an item pool reflecting the population of examinees” more precise. Each shadow test is optimal at a series of estimates of approaching the examinees’ true values of $\theta$. The choice is, therefore, between the distribution of all $\hat{\theta}$s during the adaptive tests or the distribution of the true value of $\theta$ in the population of examinees. As a matter of fact, because the distribution of the actual exposure rates of the items depends on the former but the accuracy of the ability estimator on the latter, this choice amounts to one between the relative weight put on item security and the accuracy of the ability estimator. Our personal choice would be to guarantee item security and choose the distribution of $\hat{\theta}$, if necessary with a slight increase in test length to compensate for loss of accuracy in ability estimation. In the examples in this article, the consequences of the choice between the two distributions is examined for an item pool assembled from a previous master pool from the Law School Admissions Test (LSAT).
Earlier Methods

The method proposed in this article differs from the methods studied in van der Linden, Veldkamp, and Reese (2000) and Veldkamp and van der Linden (2000). The major difference is that the methods in these references are methods for item pool design, whereas the current method is a method for the assembly of an item pool from a master pool. In an item pool \textit{design} problem, no actual items are available yet, but an optimal blueprint for an item pool is produced, where a blueprint is defined as a plan for the distribution of numbers of items over the space with all possible combinations of the relevant statistical and content attributes of the items. In van der Linden, Veldkamp, and Reese (2000), an integer-programming method is used to calculate an optimal blueprint for an item pool for a testing program with linear forms that minimizes the item-writing costs. In Veldkamp and van der Linden, an optimal blueprint for a CAT item pool is found by simulating adaptive test administrations over the design space.

Previous literature on the problem of item pool \textit{assembly} for CAT from a given master pool is scarce. Way, Steffen, and Anderson (2002) discuss several problem included in the development and maintenance of a master pool, called item vat by them. They also discuss constraints (or “docking rules,” as they call them) that may have to be imposed on a CAT pool assembled from it. But they do not provide any insight into algorithms that could be used to actually assemble such pools. However, a problem explored in somewhat more detail in the literature is the one of assembling systems of rotating item pools from a mastery pool. Ariel, Veldkamp, and van der Linden (2004) and Stocking and Swanson (1998) present methods for finding an optimal split of a master item pool into a set of (possibly overlapping) smaller pools such that rotating these pools among the examinees would result in favorable item-exposure rates.

Item Pool Assembly Method

The method is formulated for the distribution of estimates \( \hat{\theta} \) for the population of examinees. The distribution is denoted as \( g(\hat{\theta}) \). A method for the distribution \( g(\theta) \) is obtained if \( \theta \) is substituted for \( \hat{\theta} \). The question of how to get an estimate of \( g(\hat{\theta}) \) or \( g(\theta) \) is addressed later in the article.

\textit{Method}

The method consists of the following steps:

\textit{Step 1.} Determining the required item pool size, \( m \). It is assumed that \( m \) is a multiple of the length of the adaptive test, \( n \). The item pool is thus assembled to consist of \( m/n \) different linear tests.

\textit{Step 2.} Choosing an ability interval, \( [\theta_l, \theta_u] \), that is large enough to neglect possible parts of \( g(\hat{\theta}) \) not covered by it.

\textit{Step 3.} Partitioning the interval \( [\theta_l, \theta_u] \) into a series of \( m/n \) adjacent intervals. The sizes of intervals should be chosen such that each interval contains an equal portion of \( g(\hat{\theta}) \). Because we have \( m/n \) intervals, each interval should contain a por-
tion of $g(\hat{\theta})$ equal to $n/m$. The midpoint of the $f$th is denoted as $\theta_f$. This partitioning results in a set of points $\theta_f$ with a distribution that reflects $g(\hat{\theta})$.

**Step 4.** Assembling a set of $m/n$ tests from the master pool, one for each point $\theta_f, f = 1, \ldots, F = m/n$, using the model for test assembly in the next section. The model requires each test to meet the content constraints and to have maximum information at its point $\theta_f$. The set of all tests constitutes the CAT item pool.

The choice of a fixed value $n$ for the length of the test is in line with the current practice in CAT. Although a test with a stopping rule based on a fixed level of accuracy of the ability estimator would statistically be more satisfying, a common composition of the test to give it the same content validity for each examinee, which is possible only for a fixed test length, is deemed more important. This dilemma between accuracy and content validity is not serious, though; $n$ is typically chosen large enough to have ability estimators with satisfactory accuracy.

The larger the number of intervals, $m/n$, the better the set of midpoints $\theta_f$ can reflect the population distribution of the examinees. Because the test length $n$ is fixed, and $m$ is subject to practical considerations, we usually do not have much leeway when selecting the number of intervals. In the empirical sample below, the CAT specifications were taken from an existing program for a length $n = 50$. To keep the example realistic, the number of intervals and pool size, were equal to 10 and 500, respectively. The results for this number of intervals were promising, though. In practice, the length of a CAT is typically 25–35, and we are able to use the method with a larger number of intervals. We expect our results, therefore, to be conservative relative to those for real-life CAT programs.

**Models**

In the empirical example below, the three-parameter IRT model for the items in the master pool is

$$p_i(\theta) = \Pr(U_i = 1|\theta) = c_i + (1 - c_i)\left\{1 + \exp[-a_i(\theta - b_i)]\right\}^{-1},$$

where $U_i$ is a binary variable for the response of the examinee to item $i$, $\theta \in (-\infty, \infty)$ represent the examinee’s ability level, and $a_i \in (0, \infty), b_i \in (-\infty, \infty)$, and $c_i \in [0, 1]$ are the discriminating power, difficulty, and guessing parameter for item $i$, respectively (Birnbaum, 1968).

To present the model for the simultaneous assembly of multiple tests, the following notation is needed. The individual items in the master pool are denoted by $i = 1, \ldots, I$ and the individual tests to be assembled by $f = 1, \ldots, F$. The model is formulated using decision variables $x_{if}$, which take the value one if item $i$ is assigned to test $f$, and the value zero otherwise. The model will have constraints to prevent each item from being assigned to more than one test. As for the content specifications, constraints for only one set of categorical and one quantitative attribute are formulated; in a real-life application, more constraints of these types will be needed to represent the total set of content specifications. Generally, cate-
gorical constraints require tests to meet upper bounds \( n_c^{(u)} \) and lower bounds \( n_c^{(l)} \) on the numbers of items with content attributes \( c = 1, \ldots, C \). The sets of items in the master pool that have these attributes are denoted by \( V_c \). Quantitative constraints require tests to meet upper bounds \( n_q^{(u)} \) and lower bounds \( n_q^{(l)} \) on the sums of the values of the items on a quantitative attribute \( q \) (e.g., expected response time or word count). Finally, Fisher’s information in test \( f \) at \( \theta_j \) is denoted as \( I_I(\theta_j) \).

The general shape of the model is

\[
\text{maximize } y \quad \text{ (2)}
\]

subject to

\[
\sum_{i=1}^{f} I_i(\theta_f) x_{if} \geq y \quad f = 1, \ldots, F
\]

\[
\sum_{i=1}^{f} x_{if} = n \quad f = 1, \ldots, F
\]

\[
\sum_{i \in V_c} x_{if} \leq n_c^{(u)} \quad f = 1, \ldots, F, \ c = 1, \ldots, C
\]

\[
\sum_{i \in V_c} x_{if} \geq n_c^{(l)} \quad f = 1, \ldots, F, \ c = 1, \ldots, C
\]

\[
\sum_{i=1}^{f} q_i x_{if} \leq n_q^{(u)} \quad f = 1, \ldots, F
\]

\[
\sum_{i=1}^{f} q_i x_{if} \geq n_q^{(l)} \quad f = 1, \ldots, F
\]

\[
\sum_{f=1}^{F} x_{if} \leq 1 \quad i = 1, \ldots, I
\]

\[
x_{if} \in \{0, 1\} \quad i = 1, \ldots, I, \ f = 1, \ldots, F
\]

\[
y \geq 0.
\]

The first set of constraints in Equation 3 defines a lower bound \( y \) to the information in each of the tests at the points \( \theta_f \). The objective function in Equation 2 maximizes this lower bound. As a result, the tests are assembled to approximate a uniform distribution of information over \( \theta_f, f = 1, \ldots, F \), at maximum height. The second set of constraints in Equation 3 defines the length of the tests. The next four sets are needed to require both the numbers of items with categorical attributes and the sums of the values of the items on the quantitative attribute to be between upper and lower bounds. The two last sets of constraints guarantee that items are assigned to one test at most and the decision variables take only admissible values, respectively.
Solution

A solution to the optimization problem in Equations 2 and 3 can be calculated using one of the commercially available software packages for linear programming with an optimizer for mixed integer problems. The authors’ favorite is CPLEX, of which another version with remarkably improved power has been released recently (ILOG, 2002). A solution to the model in Equations 2 and 3 consists of a set of values for the decision variables $x_i$ and $y$; the former denote the items that have been assigned to the tests, the latter is the maximum value possible for the master pool given the constraints in the model.

It is important that the tests be assembled simultaneously (i.e., in one optimization run). If they are assembled sequentially, each next test would tend to be less informative at its $\theta_j$ than the tests assembled earlier at theirs. Because of the nature of the decision variables, the model in Equations 2 and 3 does assemble all tests simultaneously. The first to model the problem of simultaneous test assembly using this type of variable was Boekkooi-Timminga (1999).

For a given master pool, the number of variables $x_i$ increases exponentially in the number of tests to be assembled. If the number gets too large, it is recommended to run the optimizer until a solution is obtained with a value for the objective function differing from an upper bound to its optimum only by a small percentage. A reasonable upper bound can always be found by relaxing the integer constraints on $x_i$ in the last set of constraints in Equation 3, that is, replacing the domain of these variables by the interval $[0,1]$. Alternatively, an implementation of the model with the “dummy-test” approach proposed in van der Linden and Adema (1998) can be used.

Feasibility

Our assumption is that the testing program has sound inventory management and does not suffer from such problems as occasional depletion of the pool or substantial numbers of items written to wrong specifications. The item-inventory practice described in Way, Steffen, and Anderson (2002) satisfies these assumptions. For such programs the model in Equations 2 and 3 is feasible, even for smaller master pools.

One reason for the pool assembly problem in Equations 2 and 3 to become infeasible is a master pool short on items with some combinations of content attributes. If so, no combination of items may exist that meets all constraints. However, all constraints in Equation 3 (except those in the first set; see below) are on content attributes. If the items have been written to the correct specifications, infeasibility is therefore unlikely.

Another reason for infeasibility is a combinations of undesirable correlation between content and statistical attributes in the item pool and a model with constraints on both types of attributes. The only constraints on a statistical attribute are those in the first set in Equation 3. But these constraints define $y$ as a lower bound to the information values at $\theta_j$; because $y$ is a variable, it is impossible for these constraints to cause infeasibility.
If in an application the model happens to be infeasible, the only solutions are to replenish the master pool or reduce the number of linear forms, $F$, to be assembled.

**Empirical Study**

The goal of the empirical part of this study was to assess the impact of the distribution of the ability levels $\theta_i$ on the quality of the adaptive test. In particular, the difference in impact between sets of tests assembled at a distributions of $\theta_i$ matching $g(\theta)$ and $g(\hat{\theta})$ was studied. As a baseline, adaptive testing from a more traditional item pool was used. How testing companies actually assemble their CAT pools is one of their secrets, but from the pools we have been able to investigate, it is clear that the distribution of the items invariably reflects the distribution of examinees. We realized this by assembling our baseline pool as a set of parallel tests with information functions reflecting the population distribution $g(\theta)$. In particular, the following CAT pools were present:

1. A set of tests assembled using the method proposed in this article to be optimal at a distribution of $\theta_i$ chosen to reflect $g(\hat{\theta})$ (Condition 1);
2. A set of tests assembled using the method proposed in this article assembled to be optimal at a distribution of $\theta_i$ chosen to reflect $g(\theta)$ (Condition 2); and
3. A baseline consisting of a set of parallel tests each assembled to have an information function reflecting $g(\theta)$ (Condition 3).

Each of these pools was evaluated using CAT simulations with and without item-exposure control. The quality of the CAT administrations was determined using the bias and mean-squared error (MSE) functions of the ability estimators and the actual distribution of the item-exposure rates in the pool. It was not necessary to record the number of times the content constraints on the test were met. A sufficient condition for all shadow tests to be feasible is to have at least one feasible combination of items in the pool (van der Linden & Reese, 1998), and this condition is automatically met for item pools consisting of a set of tests each assembled to meet the content specifications.

The differences in quality of measurement between the conditions were expected to be smaller near the middle of the ability distribution than in the tails. As for the tails, we expected the pools in Conditions 1 and 2 to outperform the one in Condition 3, but hoped for small differences between the pools in Conditions 1 and 2. In practice, an estimate of the distribution of ability estimates, $g(\hat{\theta})$, for a population of examinees can easily be obtained by running CAT simulations from an item pool of comparable quality as the one that is assembled. It is more difficult, however, to infer a good estimate of the true ability distribution, $g(\theta)$, from a set of response data. Small differences between Conditions 1 and 2 would suggest that we can use $g(\hat{\theta})$ instead of $g(\theta)$.

**Method**

The master item pool was a previous pool from the LSAT with $I = 5,316$ items. This size of pool is huge for the test assembly problem involved in the method
in this paper. However, because the pool was a previous real-life mastery pool, it was decided to leave it intact. The sum of the information functions of all items in the pool is depicted in Figure 2. As the heavier right tail in this figure reveals, the pool had a tendency to be on the difficult side for the $N(0, 1)$ population assumed in this study.

The length of the adaptive test was set equal to 50 items, which is half the length of the current paper-and-pencil version of the LSAT. The bounds in the right-hand sides of the constraints in Equation 3 were reduced proportionally. The paper-and-pencil version of the LSAT does have an item-set structure with several specifications for the content of the common stimuli, which was ignored in this study though. The total number of categorical and quantitative constraints needed to model the content specifications of the LSAT in Equation 3 was 96.

The size of the operational CAT pools to be assembled from the LSAT pool was set to 500. Given the length of the adaptive test, the pools, therefore, had to consist of 10 individual tests each. This number of tests involved a version of the model in Equations 2 and 3 with $10 \times 5,316 + 1 = 53,161$ variables.

The true ability distribution $g(\theta)$ was assumed to be $N(0, 1)$. For the first condition, the distribution of $g(\hat{\theta})$ was obtained by simulating CAT administrations from the master item pool for 10,000 examinees randomly sampled from $g(\theta)$. The simulations were based on the shadow test algorithm previously described. The distributions of $g(\theta)$ and $g(\hat{\theta})$ are shown in Figure 3. As expected, the estimate of $g(\hat{\theta})$ was less peaked and, consequently, had more mass in its tails.

For the first condition, the interval $[\theta_l, \theta_u] = [-2.0, 2.0]$ was partitioned into 10 intervals using $g(\hat{\theta})$ according to Step 3 in the description of method. The intervals had midpoints $\theta_j$ that were equal to $\{-1.606, -1.018, -0.656, -0.361, -0.109, \ldots\}$.
0.143, 0.404, 0.696, 1.055, 1.627). For the second condition, the same method was used to partition the interval $[\theta_l, \theta_u] = [-2.0, 2.0]$ into 10 intervals using $g(\theta)$. These intervals had midpoints $\theta_f$ equal to $\{-1.593, -0.991, -0.648, -0.373, -0.122, 0.122, 0.373, 0.648, 0.990, 1.593\}$.

For the third condition, the pool of parallel tests with the shape of their individual test information function matching the shape of $g(\theta)$ was assembled using a modified version of the model in Equations 2 and 3. The modification was the replacement of the first set of constraints in Equation 3 by a larger set, where the test information functions were controlled at the values $\theta_k, k = 1, \ldots, K$. These sets of constraints require the test information function to have a relative shape determined by the weights $g(\theta_k)$ but maximizes its height. In this study, we chose to control the test information functions at $\theta_1 = -1, \theta_2 = 0, \text{ and } \theta_3 = 1$. This choice was motivated by our ample experience that, as information functions are smooth, well-behaved functions, control at only a few $\theta$ values is sufficient to get desired results.

For each of the conditions, the CAT simulations were repeated twice. The simulations were also based on the shadow test algorithm. In the first set, 1,000 CAT administrations were simulated for each value $\theta = -2.0, -1.5, \ldots, 2.0$. This set was used to evaluate the bias and MSE functions of the ability estimator. The same number of simulations was used at each of the $\theta$ values to get estimates of these functions with uniform accuracy along the scale. In the second set, 10,000 CAT administrations were simulated with $\theta$ values randomly sampled from $N(0, 1)$. This
set was used to evaluate the distributions of the actual exposure rates of the items in the three pools for the true population in this study.

The ability estimator was the expected a posteriori (EAP) estimator with uniform prior over $[-4.0, 4.0]$. For each simulated examinee, the ability estimator was initialized at $\hat{\theta} = 0$. In the conditions with item-exposure control, the method with probabilistic item-ineligibility constraints in van der Linden and Veldkamp (2004) was used with a target value for the exposure rates set at 0.25. This method conducts a probability experiment before each examinee who takes the test to determine which items are eligible for the examinee and which are not. The probabilities are updated on the fly to maintain the target as an upper bound on the actual exposure rates of the items.

Results

For the first two conditions, the assembly of the pools took approximately 83 s on a PC with an Intel Pentium II 860 MHz processor with 128 MB of memory. For the third condition, it turned out to be difficult to reach optimality in realistic time. The process was, therefore, stopped when a solution with a value for the objective function differing no more than 5% from the relaxed solution was found. This result was reached in approximately 3 hr. The difference in solution times between Conditions 1, 2, and 3 can only be attributable to the presence of the additional constraints in Equation 4; all other CPLEX settings were identical. Apparently, if a problem reaches this size (over 50,000 variables), the reduction of the feasible area introduced by Equation 4 was too large for CPLEX to handle. It is very well possible that different settings or a slight reformulation of the model could remove this effect. However, as demonstrated by the results in Figure 4, a solution differing less than 5% from optimality works well for all practical purposes.

Item Pools

Figure 4 shows the information functions for the 10 tests as well as for the item pools in the three conditions. Each of the individual tests had its peak approximately at the required $\theta$ value. The only difference was a slight tendency of the peaks toward the lower end of the scale to be to the right of their required $\theta$ values. This tendency is believed to be a consequence of the fact that the master pool was slightly on the difficult side for the $N(0,1)$ ability distribution (see Figure 2). Observe that the peaks for the curves for Conditions 1 and 2 followed the ability distribution in the chosen population, whereas those for Condition 3 were all near the middle of the ability distribution. Also, the peaks for Condition 1, which were based on $g(\hat{\theta})$ instead of $g(\theta)$, showed a slightly larger spread. All results were thus nearly exactly as required.

It is interesting to note that, despite the differences between the first three plots in Figure 4, the last plot shows approximately the same shape of the total information functions for the three pools. Thus, it is possible that items pools, which are nearly identical on the surface, at a deeper level have an entirely different composition. The important difference between the three pools is the distribution of the
individual tests of which they are composed. For Conditions 1 and 2, this distribution guarantees numbers of items with both maximum information and the required combination of attributes available at \( \theta \) values in proportion to the density of the examinees in the population. The pool for Condition 3 lacks this feature.

**CAT Simulations**

The quality of the three pools was further assessed using CAT simulations. The bias and MSE functions estimated in this study are given in Figure 5. Both for the CAT with and without exposure control, the bias functions for Conditions 1 and 2 had never a difference larger than 0.009 over the full range of \( \theta \) values. Both functions also showed a slight inward trend typical of a Bayesian ability estimator. The MSE functions for Conditions 1 and 2 differed never more than 0.005. Also, they were flat for all practical purposes, except for a slight upward trend at the lower end of the scale, which is believed to be a consequence of the fact that the master pool was on the difficult side, and the lower end of the distributions of the tests in the pools for these two conditions were influenced by this. The curves for Condition 3 coincide with those for Conditions 1 and 2 for the \( \theta \) values in the middle of the scale. In fact, the only meaningful differences are between Conditions 1, 2, and 3 at the lower and upper ends of the scale, where both the bias and MSE in Condition 3 became approximately twice as large as in Conditions 1 and 2.

**Item-Exposure Rates**

Figure 6 shows the distributions of the actual item-exposure rates for the three pools, both for the cases with and without exposure control. In either case, all three distributions were nearly indistinguishable. For the conditions with exposure control, each item always had an exposure rate below the target value of 0.25. This result should not be ascribed to the composition of the item pools, but to the power of the exposure-control method used, which guarantees effective control of all items (for details, see van der Linden & Veldkamp, 2004). However, the fact that these results led only to small random variation in the bias or MSE functions was due to the composition of the pools: Because each pool consisted of a sufficiently large number of items with the required combinations of attributes, when the exposure-control method had to reject an item for administration to an examinee, it was always able to find another one with the same combination of attributes that had its maximum information at a point not too far from the item rejected. As demonstrated by the less uniform shape of the bias and MSE functions for Condition 3 in Figure 5, the only difference between the item pools was that for this condition both the rejected item and its substitute tended to be closer to the center of the ability distribution than desirable.

**Discussion**

Although the method item pool assembly introduced in this article is based on a simple traditional principle, its impact on the composition of the item pool and,
FIGURE 5. Bias and MSE functions of ability estimators for adaptive tests from the three pools (with and without item-exposure control). A, Bias, no exposure control. B, MSE, no exposure control; C, bias, exposure control; D, MSE, exposure control.
hence, on the properties of the adaptive test in the empirical study was remarkably positive. The method succeeded in breaking down effectively the undesirable correlational structure between the points of maximum information and the content attributes of the items typically found in operational CAT pools. As a result, the CAT algorithm was never hampered by any lack of items with required combinations of attributes along the $\theta$ scale and able to move more freely from the location of the initial item to true ability level of the examinee. The nearly uniform MSE functions in Figure 5 witness this effect. A second result is that the item-exposure control mechanism had no difficulty replacing items rejected for an examinee with items of comparable statistical quality. The fact that, except for small random variation, the curves in Figure 5 have the same shape for the conditions with and without exposure control identical witnesses the second result.

The distribution of the item-exposure rates in Figure 6 reveals that without item-exposure control the item-selection procedure never selected about 200 items out
of the pool of 500. If item-exposure control was applied, the maximum exposure rates went below the target value of 0.25 for all items, but the number of items never selected was still about 80 for all three conditions. (The nonzero rates for items with rank number larger were smaller than 1% and are not visible in the figure.) The fact that a considerable portion of items is never selected in CAT is a necessary phenomenon. As shown in van der Linden (2003, appendix, property 1), for any item-selection procedure in CAT the sum of the exposure rates is always equal to the length of the test, \( n \). Graphically, the area below any of the curves in Figure 6 is thus equal to this number. The impact of item-exposure control is a only redistribution of the area: the exposure rates of the most popular items go down, and those for some of the others go up to compensate. The only way to use more of the items is to lower the target value for the exposure-control mechanism. However, the unexposed items are not necessarily bad. Because the CAT algorithm optimizes, items that are only slightly less informative than an item in their neighborhood may never get selected. Because they have never been exposed, they can be included in the master pool again and kept available for a later version of pool for the same test.

In the empirical example, we hardly found any differences between the pools for Condition 1 and 2 assembled for an application of the method with \( g(\hat{\theta}) \) and \( g(\theta) \), respectively. This result suggests that in practical applications there is no need to estimate \( g(\theta) \); instead we can use the much easier to estimate distribution \( g(\hat{\theta}) \).

Finally, in a sense, the method proposed in this article might only seem to have shifted the problem of item pool design from the operational pool to the master pool. In our empirical study, the master pool did not impose any severe constraint on the assembly of the operational pools. This does not necessarily hold for real-life situations with less-well-managed master pools. However, we do not recommend using our method for solving possible problems with the composition of master pools. These problems can only be solved effectively by imposing the constraints in Equations 2 and 3 directly on the item-writing practice.

References


van der Linden, W. J., Ariel, and Veldkamp


Authors

WIM J. VAN DER LINDEN is Professor of Measurement and Data Analysis, Department of Research Methodology, Measurement, and Data Analysis, University of Twente, P.O. Box 217, 7500 AE Enschede, The Netherlands; w.j.vanderlinden@utwente.nl. His areas of specialization are test theory, applied statistics, and research methods.

ADELAIDE ARIEL is Research Assistant, Department of Research Methodology, Measurement, and Data Analysis, University of Twente, P.O. Box 217, 7500 AE Enschede, The Netherlands; a.ariel@utwente.nl. Her areas of specialization are operations research and optimal test assembly.
Assembling a Computerized Adaptive Testing Item Pool

BERNARD P. VELDKAMP is Assistant Professor, Department of Research Methodology, Measurement, and Data Analysis, University of Twente, P.O. Box 217, 7500 AE Enschede, The Netherlands; b.p.veldkamp@utwente.nl. His areas of specialization are optimal test assembly, computerized adaptive testing, and operations research.

Manuscript received May 19, 2003
Revision received October 6, 2003
Accepted November 12, 2003