Research Report

A Comparison of Two Procedures for Constrained Adaptive Test Construction

Frédéric Robin

Wim J. van der Linden

Daniel R. Eignor

Manfred Steffen

Martha L. Stocking

Research & Development

February 2005
RR-04-39

A Comparison of Two Procedures for Constrained Adaptive Test Construction

Frédéric Robin
ETS, Princeton, NJ

Wim J. van der Linden
University of Twente, The Netherlands

Daniel R. Eignor, Manfred Steffen, and Martha L. Stocking
ETS, Princeton, NJ

February 2005

As part of its educational and social mission and in fulfilling the organization's nonprofit charter and bylaws, ETS has and continues to learn from and also to lead research that furthers educational and measurement research to advance quality and equity in education and assessment for all users of the organization's products and services.

ETS Research Reports provide preliminary and limited dissemination of ETS research prior to publication. To obtain a PDF or a print copy of a report, please visit:

www.ets.org/research/contact.html

A Comparison of Two Procedures for Constrained Adaptive Test Construction

Frédéric Robin
Educational Testing Service

Wim J. van der Linden
University of Twente

Daniel R. Eignor
Manfred Steffen
Martha L. Stocking
Educational Testing Service

January 2005

Abstract

The relatively new shadow test approach (STA) to computerized adaptive testing (CAT) proposed by Wim van der Linden is a potentially attractive alternative to the weighted deviation algorithm (WDA) implemented at ETS. However, it has not been evaluated under testing conditions representative of current ETS testing programs. Of interest was whether STA would, under typical high-stakes on-demand testing situations, produce tests of comparable or better psychometric quality as those produced by the current weighted deviation algorithm. Based on simulated data, we found that the STA performed as well or slightly better than the WDA on two of the three commonly accepted testing objectives: measurement and content. The WDA appeared to perform slightly better than the STA when the issue is security or item exposure control. The paper provides a review of the rationale that led to the specific testing objectives employed, an outline of the test construction steps common to the two procedures investigated, and a description of the specific models and algorithms employed by both procedures. Detailed description of the simulation study conducted and the results obtained from both of the procedures for one of the pools are also provided. Results are summarized and further research needs in particular areas discussed.

Key words: Computerized adaptive testing, automated test assembly, test security, item exposure, item response theory

Introduction

The introduction of the College Board Computerized Placement Tests (CPTs) in 1986 signaled the beginning of operational computerized adaptive testing (CAT) activities at Educational Testing Service (ETS) (see Ward, 1988). While this implementation was to occasion much interest on the part of ETS staff and outside measurement practitioners, an almost 10-year period passed before the next operational implementation of CAT with an ETS test, in this case the Graduate Record Examinations® (GRE®) General Test, was to take place. While the CPTs had a fairly simple system of item content constraints that could be addressed easily with existing item selection procedures (see Kingsbury & Zara, 1989), other tests for which computerized adaptive testing was being considered, such as the GRE, did not possess these simple item content constraint systems. With tests like the GRE General Test, content constraints were not mutually exclusive and a selected item could satisfy a number of different constraints. A procedure for dealing with such constraints was needed. Moreover, if such a system were to parallel item selection manually done by test development specialists, it would need to take into account the additional issue that all constraints are not equally relevant and the fact that it is typically not absolutely critical that all constraints be exactly met in constructing a test.

The breakthrough at ETS that allowed the consideration of computerized adaptive test construction for tests having complex constraint systems occurred in 1992 with the development of the weighted deviations algorithm (WDA) (see Swanson & Stocking, 1993; Stocking & Swanson, 1993). The GRE General Test was the first test with a complex constraint system to then be considered for adaptive administration, although the Graduate Management Admission Test® (GMAT®) and the Praxis I™ were soon to follow. For all tests, fairly extensive periods of field testing preceded actual operational administration.

When the weighted deviations algorithm was being piloted with these tests, a number of concerns surfaced. When item pools of what was thought to be appropriate size were constructed to support these CATs, the algorithm did not make full use of the complete set of items constituting a particular pool. In fact, certain items in the pools were never selected for administration. These items came to be known as *bottom feeders* and occasioned a good deal of concern on the part of test development staff. It is costly to develop items that end up never being used. Also, at that time a misunderstanding was to occur about the algorithm, which in the technical literature had been grouped (not entirely appropriately) with a number of other

approaches that are collectively referred to as *greedy heuristics* (see Swanson & Stocking, 1993). Reference to the algorithm as being greedy was taken to mean that the algorithm unnecessarily overused or selected too quickly the items in the pool that were able to satisfy multiple content constraints, or that had desirable statistical properties, and left the less desirable items completely unselected. The above concern and misunderstanding led to a consideration of whether alternative procedures to the weighted deviations algorithm might be available.

At the point when interest in possible alternative procedures for constructing CATs subject to large constraint systems peaked, a viable alternative to the weighted deviations algorithm was introduced in the literature. The approach was based on the phased construction of a series of tests (called shadow tests) from which the items for the CAT were selected (see van der Linden & Reese, 1998; van der Linden, 2000a). After some discussion of the shadow test approach at ETS, van der Linden was contacted by ETS researchers to see if he would be willing to participate in a comparison experiment. Of interest was whether the shadow test approach would, under typical high-stakes on-demand testing situations, produce tests of comparable or better psychometric quality as those produced by the current weighted deviations algorithm and, at the same time, make better use of the items available in pools. Work on a series of comparison experiments was undertaken when van der Linden agreed to take part in the experiment.

The purpose of this paper is to describe the outcomes of the comparison experiments that made use of the weighted deviations algorithm and the shadow test algorithm (STA). Three item pools having varying degrees of item set structure were selected for the comparisons. These pools had been constructed so that the WDA was able to create CATs that satisfied testing program objectives. One of the pools was made up exclusively of discrete items, one had a combination of discrete items and item sets, and the final pool was made up exclusively of items appearing in item sets. A number of different criteria were used in comparing the results of application. These were, for the most part, the standard criteria used in evaluating the outcome of CAT construction when applying one particular algorithm. The resulting CATs that were considered were constructed via simulation techniques, using standard practice at ETS (see Eignor, Stocking, Way, & Steffen, 1993).

The remainder of this paper is organized into four main sections. The first section provides a review of the rationale that led to the specific testing objectives employed. The second section outlines the test construction steps common to the two procedures investigated and

describes in more detail the specific models and algorithms employed by both procedures. In the third section, detailed description of the simulation study conducted and the results obtained for one of the pools are provided. Finally, study results are summarized and the need for further research in particular areas is discussed.

Evaluation Methods

Over time, the testing objectives and their associated criteria used for evaluating the quality of CAT administrations have evolved to better take into account the full range of programs' expectations. As potential problems with existing tests were uncovered, new or improved test construction algorithms have been developed and more comprehensive testing objectives have been made explicit. The two algorithms compared here were developed to address the major testing objectives identified to date, which include overall measurement, content, and security objectives, to be evaluated over a large number of simulated or real test administrations (Stocking & Swanson, 1993; Davey & Parshall, 1995; van der Linden, 2000a). Because the item pools selected for use in the experiments had been built to allow the WDA to function adequately in satisfying the testing program objectives, the study was, in fact, stacked in favor of the WDA; the challenge to the STA was to do at least as well as the WDA, while making better use of limited and expensive item resources. In a separate study, the WDA and STA were compared using the content specifications of an existing large-scale testing program not run by ETS and an item pool developed independently of either method. For a report on this study, which resulted in different findings, see van der Linden (2003).

The following text contains detailed descriptions of the testing objectives that guided the development of CATs using both the WDA and STA algorithms and the associated criteria that were used to compare their respective performances under the same highly constrained conditions.

Overall Testing Objectives and Criteria

Early work on adaptive testing focused essentially on statistical considerations (Lord, 1980; Weiss, 1983). By delivering to examinees only items with a difficulty level closely matching their ability, two benefits could be derived. First, by employing a variable length test and only terminating the test when some specified level of score precision had been obtained, examinees could be assessed with equivalent precision across a wider range of ability than it is

possible with linear test forms. Second, by avoiding the administration of items that are markedly too easy or too difficult for the examinee and by choosing more informative items, the number of items required per examinee could typically be reduced to approximately one-half that of linear test forms.

A major concern about early adaptive tests was that there was little to no control over the content used. Many testing programs have observed a relationship between item difficulty and item content. Some item types more readily produce difficult items than do others. When adaptively constructing tests, this can lead in extreme cases to some examinees being administered items of only one type. This raises important issues such as construct validity fairness. The two algorithms evaluated in this paper allow for the inclusion of a complex content blueprint in the item selection process. Another feature that both algorithms share is that, as currently implemented, they both assume/require that test length be fixed. That is, in order to provide more control over the content balance of the test (and, coincidentally, to facilitate the management of seating time in testing centers), the equal score precision benefit obtained from variable length testing has been given up.

A much discussed positive side effect of adaptive testing is that, because items are selected in response to examinee performance, each examinee almost certainly experiences a different collection of items. Some security concerns (e.g., answer sheet copying) seemed to have been eliminated. However, this perceived security gain was to be short lived. In practice, to significantly reduce test length, the most informative and well-targeted items amongst a limited size pool had to be sought. As a result, examinees demonstrating similar levels of performance tended to be delivered very similar collections of items. When operational test sessions from the same item pool are distributed over periods of time, test takers are afforded the opportunity to communicate between test sessions and gain pre-knowledge that is detrimental to the security of the test items. This, in turn, poses a threat to the validity of the scores. Item level exposure control was added to the item selection process in order to offset this security concern. But, while effective at limiting the rate at which items may be used over test administrations, exposure control can have dramatic consequences on the other properties of the tests produced. By restricting the availability of items and interfering with the item selection process, exposure control tends to lower measurement efficiency and/or increase the need for informative items, which in turn may decrease the sustainability of a testing program.

To facilitate the evaluation of test construction algorithms, the testing objectives that correspond to the issues discussed can be classified into measurement, content, and security categories. This classification seems particularly useful, as it sets apart objectives that may compete with each other. Under highly constrained conditions (high-stakes on-demand testing) such as those studied here, compromises or trade-offs may have to be made with algorithms like WDA in order to reach the best possible outcomes on all counts. Operational definitions of the testing objectives used in this study follow.

Measurement

With CAT, measurement objectives typically include notions of measurement efficiency and measurement precision, respectively characterized by overall measures of test length, score bias, and standard error conditional on a number of relevant ability levels. For operational reasons (see Eignor et al., 1993) most CAT programs, including the ones studied here, set test length to a fixed number of items. Thus the measurement efficiency objective was a given. With test scores for the testing programs involved being reported operationally as scaled scores resulting from the creation of expected number correct scores on a reference test (Eignor et al., 1993; Stocking, 1996), the measures used to evaluate measurement precision here were bias (BIAS) and standard error of measurement (SEM).

Under simulated conditions, these quantities can be estimated through a large number (n) of test replications over examinees at each one of $h = 1,...,H$ ability levels, as follows:

$$BIAS_h = \frac{1}{n}\left[\sum_{r=1}^{n}(\hat{\tau}_{rh} - \tau_h)\right]$$

and

$$SEM_h = \left[\frac{1}{(n-1)}\sum_{r=1}^{n}(\hat{\tau}_{rh} - \tau_h)^2\right]^{1/2}$$

where τ_h and $\hat{\tau}_{rh}$ represent an examinee's true and estimated number correct scores on the reference scale. The $\hat{\tau}_{rh}$ estimates are obtained from the corresponding IRT-based $\hat{\theta}_{rh}$ maximum likelihood estimates through the nonlinear test characteristic curve transformation (Hambleton, Swaminathan, & Rogers, 1991; Stocking, 1996).

In the context of our experiments, the measurement objectives were small BIAS values (within ± 0.5 on the τ scale used and for all *h* ability levels) and SEM values at or below the typical values judged acceptable by the testing programs (the *h*-values obtained with the WDA).

Content

Content blueprints and other content related specifications are designed to control multiple features or attributes of test forms (Swanson & Stocking, 1993; Stocking & Swanson, 1993; van der Linden & Reese, 1998; van der Linden & Glas, 2000). To the extent that these specifications are satisfied, all examinees may be provided with tests that are consistent in terms of content domain and item type representation. In order to avoid dependencies, such as one item providing clues to help answer another, item overlap constraints may also be specified to prevent the presence of some items together in the same test.

The scoring procedures and both of the adaptive algorithms evaluated in this paper rely on or are based on a unidimensional IRT model (e.g., the 3PL model). The benefits derived from that model are best realized when the responses to all items depend on only a single trait or skill. Although scales are seldom truly unidimensional, much of the power of the model can still be obtained by keeping the balance of dimensions constant across tests. The primary vehicles used to accomplish this are constraints related to the major content areas and the item types for a test.

Additionally, different item types can require markedly different amounts of time to answer. Thus, in order to standardize the workload across examinees, it is important for all examinees to be administered the same number of each type of item. For example, consider a verbal measure that contains two item types: Analogies and Reading Comprehension. On average, examinees spend four times as long answering an Analogy item compared to a Reading Comprehension item (taking into account the time required for reading the associated passage). If some examinees were administered an additional reading passage and associated set of items (and thus fewer Analogy items), it is reasonable to expect that their test would be a good deal more time-consuming.

A critical aspect of test fairness, when working with a fixed length adaptive test with a defined time limit, is that the expected workload should not be different across examinees. In a variable length/variable time limit environment, this is less of an issue. However, it can still be a concern. Not only would we expect that a test consisting of 20 Analogy items would be

measuring something different than a test of 20 Reading Comprehension items, it would be grossly unfair to expect one examinee to have to spend four times the amount of time as another to complete the same number of items to produce scores of equal precision. In a fixed length/fixed time limit environment, it is important that examinees have some expectation about test composition in order to budget their time. This allows them an optimal opportunity to complete the test. Surprises with respect to test composition could also unsettle an examinee and introduce another undesirable source of score variation.

Given that these considerations are well represented in the set of strict (STA) and/or relaxed (WDA) constraints imposed on item selection, the measures used to evaluate content objectives were

1. The proportion of content constraint violations generated and, when deviations occur
2. The extent to which content deviates from the lower or upper bounds over test administrations

Typically, with the programs investigated, content can be divided into various levels according to the importance accorded to each particular content attribute. Generally, with the WDA, the primary attributes are assigned the highest weights while secondary ones are assigned lower weights. In the context of our experiments, zero content violation was the target; however, minimally acceptable proportions of violations as well as minimal deviations from the bounds (or targets) specified were tolerated.

Security

Previous studies (Stocking & Lewis, 1995, 1998; Chang & Twu, 1998) and experience with operational testing programs have shown that, in a high-stakes on-demand CAT environment, the security of the items and consequently the validity of the test delivery can only be maintained if sufficiently low conditional (on ability) exposure rates are maintained for each item. Considering conditional item exposure data (i.e., the rate at which an item is used over test administrations associated with equivalent final ability estimates) the measures used to evaluate the security objective were

1. Conditional maximum item exposure rates (determined over all available items)
2. Conditional numbers of items with exposure rate above a specified upper limit

With programs specifying the acceptable upper limit, the security objectives were to obtain all conditional maximum item exposure rates below that value. As with content, minimally acceptable deviations from the upper limit of were acceptable.

Test Construction Process

The management of the item resources needed to operate a high-stakes CAT program plays an important role in realizing overall testing objectives. Currently, the consideration and management of the complete set of items available to be used for CAT construction, or *vat* as it is referred to at ETS, is done periodically. As a result of each vat management operation, new item pools made up of relatively small subsets of the vat are developed and then sent to testing centers to replace the pools that have been in use. Operational pools provide the resources from which tests are to be selected. Therefore, the test construction process is the result of not only the selection of items from an item pool at a particular time or from pools over time obtained through the use of item selection algorithms such as the WDA or STA, but also the result of the process of selecting items to form pools through the application of vat management and pool assembly procedures. If enough items with the required characteristics (in terms of measurement and content) available in the vat are included in the pool, item selection algorithms may function optimally. However, if the pool is too small or has deficiencies in some content areas, the functioning of test construction algorithms may be hampered. In fact, we will then see remarkable differences between the WDA and STA in how they cope with such conditions.

The following section contains a brief description of the vat management and pool assembly approach that was used in creating the pools used in our simulation experiments; the test assembly process common to both the WDA and STA algorithms; and, described in detail, the unique features of each algorithm.

Vat Management and Pool Assembly

The role of vat management is to maintain up-to-date information on each existing or newly created item with respect to its psychometric characteristics, usage history, and availability status. Based on possible drift in psychometric characteristics over time and based on past usage rates, items may become temporarily retired (i.e., *docked*) or permanently retired from the vat. Then, given the item resources available at a particular time, pools are assembled to best support the item selection process. For the test assembly process to satisfy the measurement and

content testing objectives and ensure test security, each pool produced must provide enough items with enough information in all content areas, at all relevant ability levels. In any practical situation, vat size and vat replenishment rates are limited, so pools must also be assembled in such a way that the use of items is spread across pools and over time and, roughly, the same amount of item resources remain available each time a new pool is needed. Detailed descriptions and references on the process used for vat management and pool assembly can be found in papers by Way and Steffen (1998), Mills and Steffen (2000), and Way, Swanson, Steffen, and Stocking (2001).

Typically, the process used to select the pool(s) to be used for test assembly can be described as follows:

Step 1. After taking into account past item exposures and item monitoring and quality control analyses, determine the list of items that can be made available for administration.

Step 2. Extract a large enough subset of the available items from which one or more pools may be assembled (generally avoiding items with lowest IRT a-parameters within each content category).

Step 3. Assemble one or more pools to meet pool specifications, including number of items and pool information targets according to a content-by-difficulty classification.

Step 4. Test the appropriate functioning of the pool(s) for either validation for operational use or revision and rerunning of the selection process (Steps 1 to 3) until obtaining the appropriate pool(s)—where the testing of the pool(s) is done by simulating the administration of a large number of tests to test takers representative of the examinee population, using the test assembly algorithm and the settings and specifications that will be used operationally.

In the cases studied, the specifications driving the selection of the pool(s) (Step 3) and the specifications driving the WDA test assembly from a pool (Step 4) have evolved over time. During the initial transition from paper-and-pencil to CAT, the pool was simply the collection of all available item resources (result of Step 1). Other than having enough items for each major content area, little attention was paid to the balance of resources (information as well as number of items) across content areas or across difficulty levels and to possible adverse effects of

dependencies such as those that may exist between item enemies (i.e., items that for various reasons cannot appear together in the same test administration). Configuring a pool amounted to running a series of simulations (Steps 3 and 4) in which the specifications were modified in order to try to produce CATs that met all of the content constraints in addition to the other testing goals. As pool development became more of a production activity, and the vat of item resources became large enough to afford some flexibility, pool selection became more purposeful.

Many goals were considered in developing pool specifications; however, two of these were deemed to be of the utmost importance. To minimize some of the compensatory solutions that had been observed in early pools (e.g., information at lower levels of difficulty coming in large part from more difficult items), the first major goal was to have each major content stratum contain as full a range of item difficulty as possible, with information goals within each combination of content and difficulty to be met. The second major goal was to reduce the need to custom configure each pool. Essentially, this meant finding a set of pool specifications that was likely to function on the first attempt. With the heuristic underlying the WDA, this was most easily accomplished by having relatively tight pool specifications ensuring a high degree of similarity across pools and, also, a high degree of similarity across tests and across pools at each stage of test construction. Although an "ideal" pool could have been defined, the VAT resources could not have sustained those specifications for any extended period of time. Through a series of trial-and-error experiments, a set of pool specifications that was deemed most likely to be sustainable were developed and appropriate settings for the WDA (i.e., the content deviation weights described below) were determined. The "average" of the specifications used to configure these pools was then taken as the starting point for all future pools. As a result of this approach, most pools were then developed and validated in the first trial.

Although the pool specifications were developed in light of operating with the WDA, they were by no means *optimized* to function with the WDA. They were developed to form the best compromise between perceptions of an ideal pool and the practical realities of available item resources and production schedules. For the purpose of this study, we assumed that because the pools used in our experiments were tested and validated with the WDA and, because the content, psychometric, and item exposure objectives were the same for both algorithms, the content of the pools used would not unduly hinder the STA.

Test Assembly Process

Given a pool of calibrated items, CAT test assembly is generally done through a sequential selection process where items are selected and administered one at a time until the examinee has responded to the required number of items and testing ends. In this way each item selection decision can be made with all the information available (items previously administered and examinee responses collected so far) taken into account. In assembling tests, both the WDA and STA go through the following steps:

Step 1. Initialize the pool (all the items are set to be available for selection) and set the initial ability estimate before starting a new test administration.

Step 2. Find the best possible subset of *l* items (item selection list) from the available pool, given the items already administered and the current interim ability estimate.

Step 3. Apply exposure control to the item selection list and select the next item for administration (which also results in the item selected and the items that were passed over because of exposure control to become unavailable).

Step 4. Administer the selected item and score the response provided by the examinee.

Step 5. Update the interim ability estimate using the item responses already provided (and possibly some prior information).

Step 6. Check the test termination rule(s) and either go back to Step 2 to continue testing or go to Step 7.

Step 7. Score the test and go back to Step 1 to start the next test.

Figure 1 illustrates these steps. In particular, it shows how the size of the item pool shrinks (suggested by the smaller dotted oval) with more items becoming unavailable for selection as testing progresses. With the WDA, these items remain unavailable until the end of the current test and the pool is re-initialized to begin the next administration (Step1). With the STA, the items that have passed over because of exposure control (Step 2) are also set aside, but in a spare pool, which may be released back into the pool if an infeasibility (i.e., the algorithm is unable to select the next item) is to arise before the ongoing test is finished.

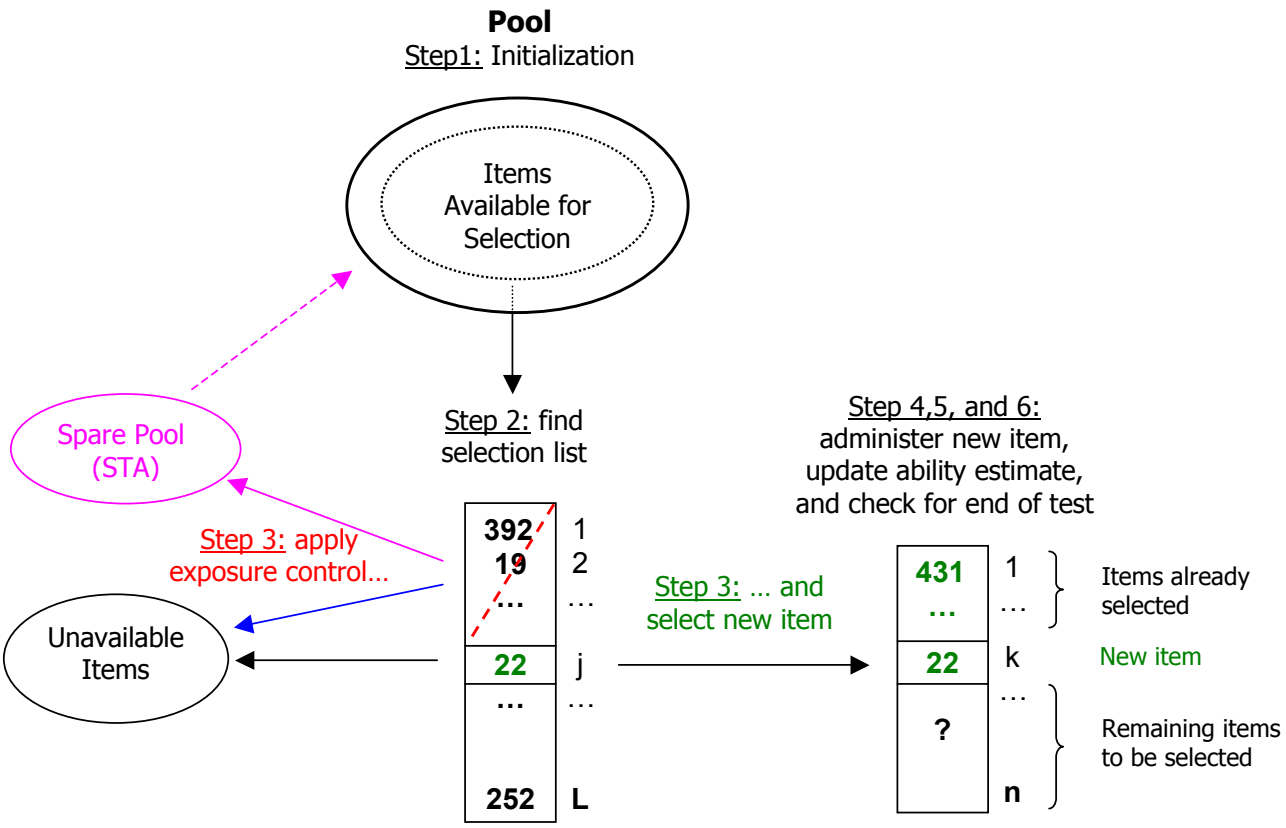

Figure 1. **Diagram of the item selection administration process. Although these steps are common to both the WDA and STA, important differences in the methods and algorithms employed to carry them out exist.**

Shadow Test and Weighted Deviations Algorithms

Both algorithms compared here—the weighted deviation algorithm (WDA) developed by Stocking and Swanson (Stocking & Swanson, 1993; Swanson & Stocking, 1993) and updated to include the multinomial conditional exposure control procedure developed by Stocking and Lewis (1995, 1998), and the shadow test algorithm (STA) developed by van der Linden (van der Linden & Reese, 1998; van der Linden, 2000a) and also updated to include the multinomial conditional exposure control procedure—follow the same steps described above. However, in Step 2, while the WDA employs a weighted deviations model and a heuristic algorithm to find a best-suited item list, the STA employs a maximum information model and a combinatorial algorithm to find that list. In Step 3, both procedures use the same conditional exposure control

mechanism, which was applied to their respective lists obtained as a result of Step 2. However, because the nature of the WDA and STA selection lists differ, their implementations of the exposure control mechanism also differ (see following text for more details). In Step 5, while the WDA employs maximum likelihood estimation (MLE) techniques to obtain the interim ability estimates, the STA employs expected a posteriori (EAP) estimation with a noninformative (-5.0, 5.0) uniform prior. Both algorithms used the same initialization, administration, test termination, and scoring procedures (Steps 1, 4, 6, and 7).

Potential advantages and drawbacks of the MLE and EAP interim estimation procedures (Step 5) have been investigated (Wang & Vispoel, 1998; van der Linden, 2000b; Yi, Wang, & Ban, 2001). Early in test administrations, MLEs may diverge towards extreme positive or negative values that need to be replaced by acceptable upper or lower bounds. EAPs, on the other hand, can always be determined but tend to produce estimates that are regressed towards the mean of the prior distribution provided (with a uniform prior the regression effect is minimum). However, as more item responses are gathered, the two estimators tend to converge towards similar values. Although studies conducted under essentially unconstrained conditions have shown a somewhat faster convergence of interim ability estimates towards the final estimate and slightly smaller standard errors of estimation as a result of using EAP, under highly constrained conditions, where pool limitations, content and item exposure constraints, and best item information available at the current ability estimate are all important factors in item selection, the potential benefits of EAP may not be significant. The operating choices made by the algorithms' authors are likely to respectively produce the best outcomes under the experimental conditions studied; thus, we followed these (although using the same ability estimation methods would have avoided a possible confounding of the comparison of the two selection algorithms).

The differences between the models and algorithms employed for finding the best-suited item list are more likely to provide and explanation for any differences in outcomes between the WDA and STA than the interim ability estimation procedure employed. Also to be noted are some subtle differences in their respective implementation of the conditional exposure control procedure employed.

Shadow Test Algorithm

With the STA, solutions to the problem of finding best-suited item lists are found based on a constrained (test length, content, and overlap) maximum test information model solved

using a full search combinatorial optimization algorithm (Nemhauser & Wolsey, 1988; Winston, 1991). For the experiments, STA best-suited item lists were set to include the two most informative and independent shadow tests that could be assembled from the currently available items (a shadow test is defined as a collection of items that, when added to the items already administered, forms a complete feasible test) (van der Linden & Reese, 1998; van der Linden, 2000a). The logic of selecting items from the most informative shadow test(s), rather than simply selecting the most informative items available at the current selection step, takes into account the future selection steps needed to complete the test; thus, selections that may appear best in the short term, but would hamper future selections, are avoided. Given reasonable pool size, finding a most informative shadow test or tests is a combinatorial optimization problem that can be solved efficiently (on the order of one second with a Pentium III processor) with the CPLEX commercial integer programming solver (ILOG, 2000).

The effectiveness of the multinomial conditional item exposure control and item selection process can be influenced by the length of the selection list. (Stocking & Lewis, 1995, 1998). To a certain extent, longer selection lists tend to improve exposure control but result in more items being made unavailable to the point of possibly depleting the pool towards the end of test administration. Shorter selection lists, on the other hand, may not permit the desired level of exposure control. Through extensive simulations, van der Linden (2001) found that selection lists, including two independent shadow tests, rather than one, was the best compromise for the testing situations investigated in this study.

Unfortunately, in controlling item exposure, a number (which depends on the multinomial probabilistic experiment) of the top items that belong to the item selection list are made unavailable for the current and the next selections of an ongoing test administration. Therefore, as the pool gets smaller, the guarantee of always finding a feasible test can be lost. To be able to avoid infeasibility, the items that are made unavailable as a result of exposure control are set-aside in a spare pool. Then, in the rare occasion where the pool of available items becomes too depleted before the end of a test, the spare pool can be used to replenish the pool and solve the infeasibility problem (van der Linden, 2001). Under the conditions studied, the spare pool was needed in less than 5% of the simulated test administrations.

Weighted Deviations Algorithm

With the WDA, solutions to finding best-suited item lists are obtained based on a more complex formulation of the problem; this is solved using a heuristic algorithm (Stocking & Swanson, 1993). Instead of formulating the content specifications into constraints to be strictly enforced, they are formulated as goals or targets. Deviations from the content targets are then weighted and incorporated in the objective function together with test information, so that the best solution to be found maximizes information and minimizes the deviations from the content specifications. Considering solutions that violate content specifications avoids the feasibility issue, and new trade offs between content and information can be considered. A potential drawback with this model is that, as a result of not strictly enforcing all the content constraints (item enemy constraints are still enforced), nearly all possible *k*-items subsets from the pool (*k* corresponding to the number of items that remain available for selection) may have to be evaluated in order to find an optimal solution. Even with a commercial integer programming solver such as CPLEX, this is not likely to be possible in a reasonable amount of time. Rather than using full combinatorial optimization techniques, the heuristic algorithm developed to solve the weighted deviations model focuses on estimating the potential of each available item to contribute to the test total weighted deviations. To do so, the IRT information and the collection of attributes accumulated through the items already administered, as well as estimates of the IRT information and the collection of attributes that are likely to be obtained from the next item selections to be made to complete the test, are taken into account (see "provisions for avoiding local optimality" in Stocking & Swanson, 1993, p. 281). The best-item selection list is then simply made up of the items estimated to have the most favorable contribution to the test being assembled.

Concerning exposure control with the WDA, the length of the selection list was fixed to 10 items (a value that has generally been found to work well in practice). Because that number is small enough relative to test length and pool size and because item overlap lists can be kept small enough when the pool is constructed, the pool can never be depleted. Because the content constraints are treated as goals rather than strict enforcements, tests can always be assembled; therefore, feasibility is not an issue and, consequently, setting items aside in a spare pool as they become unavailable is not needed.

Summary

Summarizing comparison of the methodologies used with the WDA and the STA, it should be clear that, although the test construction process is done sequentially (one item at a time), both the weighted deviations and shadow test algorithms are designed to optimize the item selection at the test level. Hence, both algorithms improve upon procedures that do not take into account possible adverse effects of the item selection at hand on the item selections that remain to be done before testing is finished—procedures that would technically be referred to as *greedy* (Nemhauser & Wolsey, 1988; Winston, 1991). Differences between the two approaches reside in part in the respective optimization models used. The STA allows for a tradeoff between the precision of ability estimation and the item exposure rates and the WDA allows for a tradeoff between the precision of ability estimation, the item exposure rates, and the deviations from content specifications. The STA uses a commercial solver that provides an exact solution to the STA optimization problem; while the WDA uses a specialized heuristic that is very fast in terms of computing time and provides an approximation to the WDA optimization problem. Another difference is that item selection lists are set up in a different manner. With the WDA, the length of the item selection list remains fixed. The STAs varies as testing progresses and shadow tests become shorter, changing the rate at which items may be excluded as a result of exposure control.

Ultimately, finding out the strengths and weaknesses of each algorithm is a matter of empirical evidence that can be gathered from the solutions they provide in the context of typical testing situations. The next section presents extensive simulations conducted to gather evidence and the results obtained. For a second study with different results, see van der Linden (2003).

Simulation Study

Data were generated to compare the performance of the WDA and STA algorithms under a range of high-stakes testing conditions represented by item pools from three different operational ETS CAT programs. One pool was made up entirely of discrete items; the second pool, of a combination of discrete and set-based items; and the final pool, of set-based items. The set-based item selections were also done one item at a time; but, once an item belonging to a set is selected, the next items must be selected from the collection of items associated with the same stimulus until the specified number of items to be selected for that set is reached. Because the collection of items available for a particular item-set selection is almost always greater than the fixed number of items to be selected, the composition of a set to be administered may vary.

Each experimental condition (i.e., each pool from which CATs were to be constructed) was defined by its testing objectives (specific to the testing program they represent) previously defined in terms of measurement, content, and security. Each pool was chosen from among those used for actual administration of the test to real test takers in 1998. Analyses of the simulated data were conducted to evaluate the extent to which each testing objective was satisfied and to compare the WDA and the STA test construction algorithms' performance.

Detailed results for the testing situation represented by the pool containing only discrete items are provided. The pool used in this testing situation will be referred to as Pool A. As it is most often the case, this pool was selected on the first trial (through the process described earlier) without modification of the standard content weights and, thus, can be considered representative of the pools typically used. Because the two experiments conducted in the situations represented by the two other pools (and referred to as Pools B and C) led to very similar findings, the detailed results are not provided. A table of summary results across the all three testing situations simulated is provided to highlight the respective performance and the strengths and weaknesses of each test construction procedure.

Experiments

Each simulation experiment was executed by replicating 500 test administrations to examinees at each of several ability levels ranging from extremely low to extremely high. Unconditional results whenever computed were obtained by aggregating conditional results using weights representative of typical population densities at each ability level (Mislevy, 1984).

Each test administration was initialized with the first interim ability estimate set to 0.0. Item selections were then done using the WDA or the STA algorithm, item responses were generated according to items' 3PL models, interim ability was estimated using MLE (WDA) or EAP with noninformative uniform prior (STA), and, finally, scoring was done using MLE expected number right estimation.

Because of exposure control, simulation experiments were conducted in two steps (Stocking & Lewis, 1998). Preliminary runs were conducted to establish the most appropriate conditional item exposure parameters for the desired level of test security. Then, a final run mimicking actual administrations was executed generating the empirical data to be analyzed.

The three testing situations studied represent somewhat different high-stakes on-demand test construction challenges (Table 1); in particular, the tests constructed from Pool B used both

discrete items and set items and the tests constructed from Pool C used only sets. When the use of set items is specified, item selection is further constrained so that, whenever an item belonging to a set is selected, the next few items (generally 3 to 5 items, depending on the particular set) have to come from the same block of items available to form that set.

Table 1

Summary of Experimental Conditions Across Testing Programs

	Pool A	Pool B	Pool C
Test specifications/objectives			
Measurement:			
Bias (across ability levels)	<0.5 (on tau scale)	<0.5 (on tau scale)	<0.5 (on tau scale)
SEM (across ability levels)	Typical WDA results or better	Typical WDA results or better	Typical WDA results or better
Test specifications/objectives			
Content:			
Fixed test length (items within sets)	28 (0)	35 (26)	28 (28)
Number of content constraints to satisfy (minimal violations tolerated):	24	56	22
Security:			
Conditional item exposure (minimal violations tolerated):	< 0.29	< 0.22	< 0.20
Resources (item pool):			
Total number of items	397	474	626
Number of discrete items	397	84	0
Number of sets	0	60	97
Item statistics:			
a-parameters mean (and SD)	0.93 (0.30)	0.78 (0.21)	0.69 (0.22)
b-parameters mean (and SD)	0.27 (1.26)	0.19 (1.17)	-1.11 (1.16)
c-parameters mean (and SD)	0.17 (0.04)	0.17 (0.11)	0.14 (0.09)
Item overlap:			
Number of lists	48	14	0
Average list size	9.5	3.7	0

Detailed Outcomes With the Discrete Item Pool (Pool A)

In the Pool A testing situation, the expected number correct scoring or tau scale was established between 0.0 to 60.0 points. To thoroughly span that range, 11 ability levels from 0 to 60 points by increments of 5 were used.

Also with Pool A, noting that content constraints 21 to 24 (associated with low weights) were set as targets and that deviations were, to some extent, judged acceptable by test developers (in fact allowing the WDA to make tradeoffs between information and content that the STA would not be allowed to make as a result of enforcing them as bounds), two STA simulations were conducted. One simulation, in which the content bounds 21 to 24 were set to the target values (lower and upper bounds of 3 for all four of them), resulted in content results (unnecessarily) better than those of the WDA. One simulation, in which the content bounds 21–24 were set to the values that the WDA was able to obtain (lower and upper bounds of 2 and 4 instead of 3), actually set up different specifications than the ones used with the WDA in order to achieve content results similar to those of the WDA. As it turned out, to the credit of the STA, the more restrictive content specifications did not have a significant effect on the measurement and item exposure objectives we looked at. That being noted, only the less restrictive specifications more in line with what was demanded of the WDA are reported below.

Measurement

Figure 2 shows both the conditional measurement bias and SEM results for both procedures with Pool A. All WDA and STA bias values were close and acceptably low. Similarly, WDA and STA SEM values were close together and slightly above 4.0 on a wide range of very low to moderately high abilities. As expected, scale restrictions lead to smaller SEMs (down to about 2.0) towards highest abilities. This did not occur at the lowest ability levels (most likely because of uncertainty associated with examinees' guessing). Overall, very similar measurement outcomes were obtained between the WDA and STA.

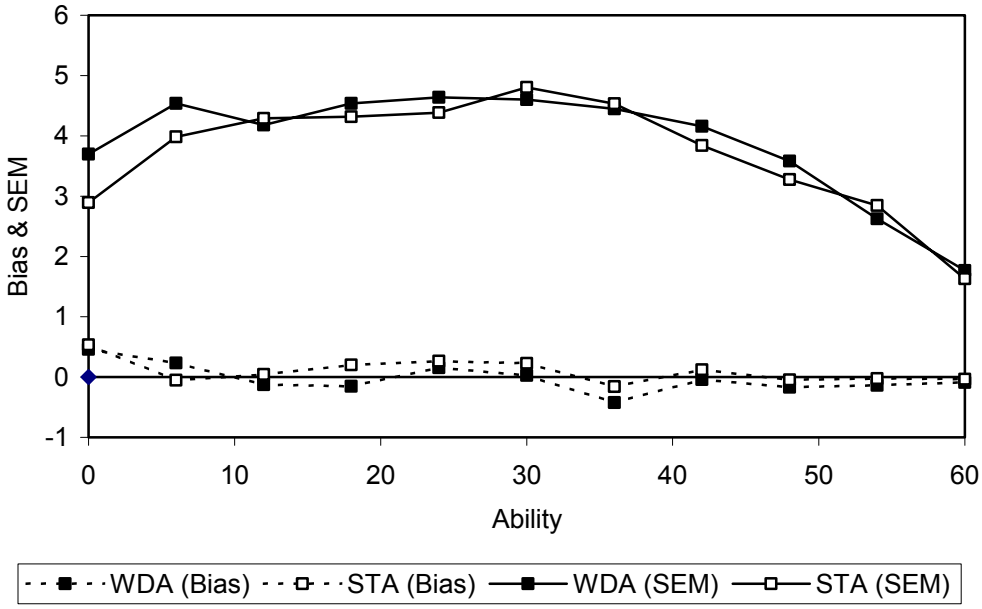

Figure 2. **Measurement outcomes for Discrete Pool WDA and STA simulations.**

Content

Table 2 shows the extent to which content specifications were met with Pool A. It includes the complete list of the 24 test content attributes (constraints) specified; the weights (cwts) used by the WDA to prioritize attributes and to realize appropriate trade-offs between the content attributes and between content and test information when compromises have to be made and lower (cblo) and upper (cbhi) bounds set for each attribute and used by both WDA and STA; the average number of items selected for each test that possess the attribute (cbar) and the percentage of tests that violate the content specifications (viol); and, for particular numbers of items administered, the percentage of the total number of CATs administered that contained that number for the examinee population estimated, based on the simulated data. Two rows per content attribute were used to report the results; the first one indicates the WDA results and the second one the STA results (between parentheses).

Table 2

Content Outcomes for Discrete Pool WDA and STA Simulations Including: Average Test Content, Percent of Tests* W*ith Content Violation, and Distribution of Content by Attribute

	Specifications			Results																		
				Summary		Number of Items with the Content Attribute																
Attr	cwts	cblo	cbhi	cbar	Viol	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
1	2.0	16	16	16	0																	100	
				(16)	(0)																	(100)	
2	2.0	12	12	12	0													100					
				(12)	(0)													(100)					
3	1.0	16	16	16	0.1																0	100	0
				(16)	(0)																	(100)	
4	1.0	9	9	9	0.1									0	100	0							
				(9)	(0)										(100)								
5	1.0	3	3	3	0				100														
				(3)	(0)				(100)														
6	1.0	15	15	15	0															0	100	0	
				(15)	(0)																(100)		
7	1.0	13	13	13	0													0	100	0			
				(13)	(0)														(100)				
8	3.0	10	10	10	0										0	100	0						
				(10)	(0)											(100)							
9	3.0	4	4	4	0				0	100	0												
				(4)	(0)					(100)													
10	3.0	2	2	2	0			100															
				(2)	(0)			(100)															
11	3.0	6	6	6	0.1						0	100											
				(6)	(0)							(100)											
12	3.0	5	5	5	0.1						100	0											
				(5)	(0)						(100)												
13	3.0	1	1	1	0		100																
				(1)	(0)		(100)																
14	5.0	8	8	8	0									100	0								
				(8)	(0)									(100)									
15	5.0	8	8	8	0								0	100									
				(8)	(0)									(100)									
16	6.0	7	7	7	0							0	100										
				(7)	(0)								(100)										
17	6.0	5	5	5	0						100	0											
				(5)	(0)						(100)												
18	6.0	1	2	1.9	0.5		6	94	0														
				(2)	(0)		(2)	(98)															
19	5.0	2	3	2.4	0.6		1	54	45	0													
				(2.8)	(0)			(22)	(78)														
20	4.0	4	4	4	1.7				2	98													
				(4)	(0)					(100)													
21	1.0	3	3	3.2	1.7			5	72	21	2	0	0										
				(2.9)	(0)				(100)														
		2	4	3.4	3.8		0	8	48	40	3	1	0										
				(3.3)	(0)			(22)	(29)	(49)													
22	1.0	3	3	2.9	0.1		0	17	81	2													
				(2.9)	(0)				(100)														
		2	4	2.4	1.7		2	62	29	7	0												
				(2.4)	(0)			(68)	(24)	(8)													
23	1.0	3	3	2.9	0.5	0	0	18	76	6	0												
				(2.9)	(0)				(100)														
		2	4				1	28	52	19	0												
				(2.9)	(0)			(37)	(35)	(28)													
24	1.0	3	3	3.1	0.1		0	8	76	16													
				(2.9)	(0)				(100)														
		2	4				1	15	40	43	1												
				(3.4)	(0)			(18)	(23)	(59)													

Note. Results between parentheses are for STA.

Considering attributes 1 to 24, WDA results showed only minor content violations (less than 2% of the tests for attribute 20 and, at most, 0.1% for the other attributes), which although not ideal were judged acceptable. Note that for attributes 21 to 24, results were reported under the two alternative sets of specifications discussed earlier (lower/upper bounds set to the initial target of 3, and relaxed to 2 and 4, respectively). However, in any case, outcomes were evaluated according to the same desired lower/upper values of 2 and 4, respectively. As expected, there was never any violation with the STA and specifications were followed to the letter. With the WDA, specifying a target of 3 lead to more concentrated outcomes around 3; however, either specification resulted in about the same minor violations.

In terms of deviations from the bounds (when violations occurred), they did not go beyond 1 for the WDA—except for attributes 23 and 21 when, on rare occasions (less than 0.3% of all tests), deviations of 2 and 3 occurred.

Security

Figures 3 and 4 provide details on the security objectives outcomes of the discrete pool (Pool A) WDA and STA simulations. Figure 3 shows that the maximum item exposures were similar across both algorithms. Figure 4 shows that the numbers of overexposed items by ability level were also similar across both algorithms, except at extreme ability levels.

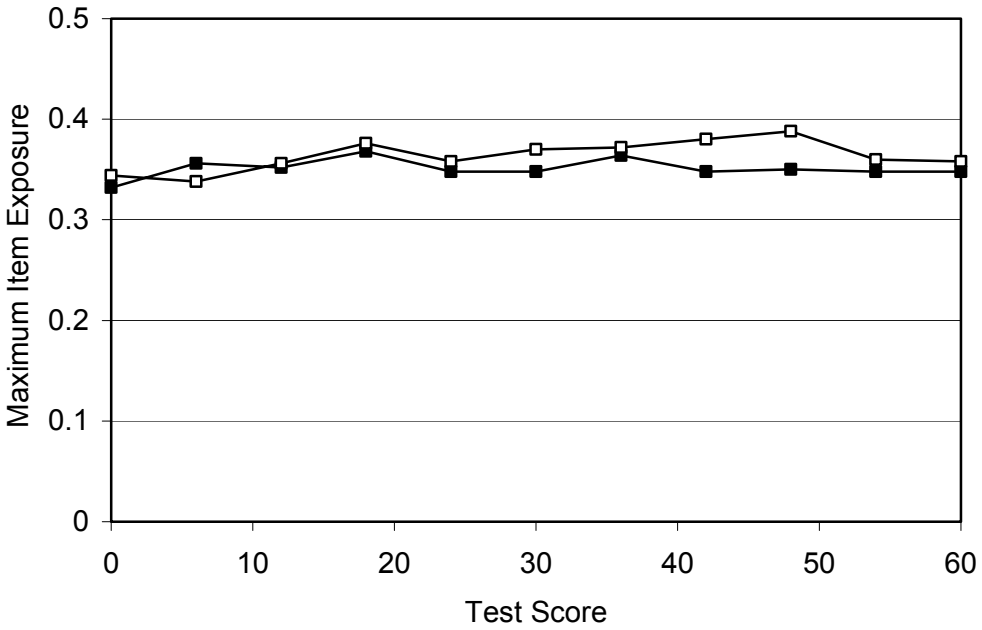

Figure 3. **Security outcomes for Discrete Pool A WDA and STA simulations—maximum item exposure by ability level.**

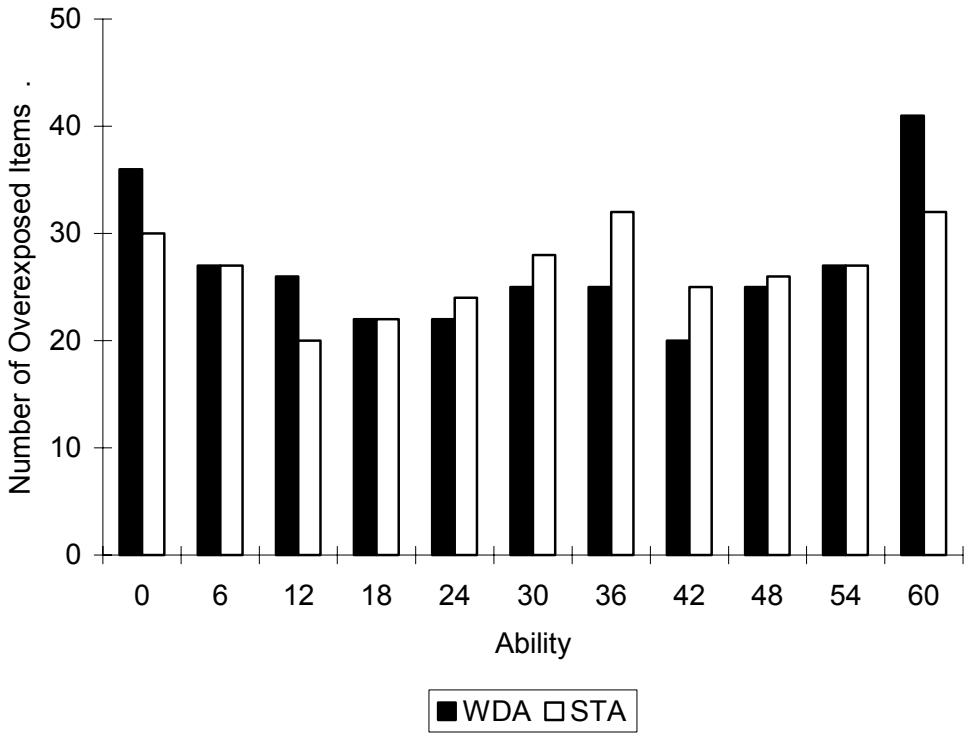

Figure 4. **Security outcomes for Discrete Pool A WDA and STA simulations—number of overexposed items by ability level.**

Summary of Testing Objective Outcomes for Pools A, B, and C

Table 3 summarizes the testing objective outcomes under all three test pools. Results are grouped by testing objective, with Pools A, B, and C on the first, second, and third lines, respectively. Conditional results that were obtained at various ability levels were grouped together into five summary levels (very low, low, medium, high, and very high ability) to be compared across pools. Absolute WDA results (least favorable value when grouping more than one ability level) and the amount by which STA results were higher (positive number) or lower (negative number) were reported. All STA results were reported between parentheses next to the WDA results.

As was the case with Pool A, WDA and STA Pool B and C outcomes were very similar. Overall, measurement results were slightly better with the STA, while the opposite was true with exposure control. By design, the STA ensured a tighter control of the content than the WDA. The two Pool A simulation experiments with the WDA indicated that a tighter control of the content was possible without noticeable degradation of measurement or exposure control results.

Table 3

Summary of Testing Objective Outcomes for Pools A, B, and C

	For Overall Examinee Population or Conditional on Ability				
	Very Low	Low	Medium	High	Very High
Pool A:	{0;6}	{12;18}	{24;30;36}	{42;48}	{54;60}
Pool B:	{10;15}	{20;25}	{30;35}	{40;45}	{49}
Pool C:	{10;15}	{20;25}	{30}	{35}	{38}
Measurement:					
1. BIAS	0.53 (-0.07)	0.20 (-0.05)	0.26 (0.16)	0.12 (0.04)	0.04 (0.10)
	0.08 (0.49)	0.09 (0.23)	0.09 (0.20)	0.02 (0.08)	0.02 (-0.01)
	0.04 (0.07)	0.06 (0.12)	0.05 (-0.03)	0.16 (-0.09)	0.01 (0.06)
2. SEM	4.5 (-0.6)	4.5 (-0.2)	4.2 (-0.3)	4.5 (-0.6)	4.5 (-0.6)
	2.8 (-0.1)	3.4 (-0.1)	3.3 (-0.1)	2.5 (0.1)	0.6 (-0.1)
	2.4 (-0.1)	2.8 (-0.1)	2.2 (0.0)	1.7 (-0.1)	1.0 (-0.1)
Content:					
1. Violations	Less than 2% violations on any attribute (none)				
	Almost none, except for one attribute with about 0.1% violations (none)				
	None for most attributes, few around 4%, and one at 12.5% (none)				
2. Deviations from bounds when violations occur	By 1 item maximum for most attributes, by 2 and 3 on two attributes (none)				
	By 1 item maximum (none)				
	By 1 item maximum for most attributes, by 2 on few attributes (none)				
Security:					
1. Max. Exposure; target =0.29	0.34 (0.01)	0.37 (0.01)	0.36 (0.01)	0.35 (0.03)	0.35 (0.10)
0.22	0.27 (0.01)	0.29 (-0.02)	0.28 (0.03)	0.28 (0.04)	0.27 (0.00)
0.20	0.26 (0.04)	0.25 (0.02)	0.23 (0.07)	0.24 (0.12)	0.27 (0.04)
2. Number of overexposed items	30 (6)	22 (4)	32 (-7)	26 (-1)	32 (9)
	70 (-4)	36 (20)	36 (12)	42 (18)	74 (-3)
	39 (12)	28 (4)	20 (5)	25 (13)	50 (1)

Note. The results between parentheses are for STA.

Conclusion

The purpose of this study was to compare the performance of the WDA and the STA using item pools from three existing CAT programs at ETS. Since the WDA has been used operationally to construct CATs with these pools and deemed to be successful, this goal translated into one of seeing whether the STA could do at least as well as the WDA, while making better use of limited item resources for the particular pool in question.

Three commonly accepted testing objectives were used to compare the performance of the two algorithms. These three had to do with measurement, content, and security. The STA performed as well or slightly better than the WDA on two of these three objectives: measurement

and content. The WDA appears to perform slightly better than the STA when the issue is security, or item exposure control. The comparability of results should provide some degree of comfort to the testing programs now employing the WDA for CAT construction purposes. Using at least these three testing objectives, the STA and the WDA produced very similar results.

At a more general level, one problem that the study clearly highlighted is the undermining effect of item exposure control on item selection. In response to this further research is being conducted at the University of Twente (see van der Linden, in press; van der Linden & Veldkamp, in press) and at ETS on alternative procedures for controlling item exposure that are independent of the item selection process.

As far as the general purpose of this study is concerned, the STA did not produce dramatically better results than the WDA. As expected, the STA met every content objective without exception where the WDA has low rates of violations for some minor constraints. In terms of psychometric quality and resource usage, the results were very similar. Thus, there is no compelling reason to believe that many of the practical issues that have arisen in the past few years (Stocking, Steffen, & Eignor, 2002) can be cured simply by switching algorithms. Even if a switch were to be considered important, there are at least two practical issues that would need to be investigated further by ETS prior to the switch. First, to date the STA software has been used primarily as a research tool. Some, possibly significant development time would be required to incorporate the STA methodology used into the current test delivery application. Second, unlike the WDA routines, which were written in house, the STA utilizes commercial software (ILOG, 2000) that carries some licensing fees for production applications.

The conclusions drawn here should not be overweighted. Comparisons were made for only a single pool for each measure, and pool composition was not varied. Resource utilization and the complexity of configuring pools have become major issues for CAT programs. It is possible that the differences in the two methods would be manifested more clearly in terms of things like the robustness to the specific composition of a pool or in the robustness of the quality of tests to aberrant responding behavior. Recent years have also seen an increase in the concern over the composition of specific tests. Thus, there are a variety of other criteria to be used and comparisons that can and should be made before firm conclusions are drawn about the utility of these two methods. Further research in this direction would most likely lead to new improvements for both the WDA and STA.

References

Chang, S., & Twu, B. (1998). *A comparative study of item exposure control methods in computerized adaptive testing* (Research Rep. No. 98-3). Iowa City, Iowa: American College Testing.

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

Eignor, D. R., Stocking, M. L., Way, W. D., & Steffen, M. (1993). *Case studies in computer adaptive test design through simulation* (ETS RR-93-56). Princeton, NJ: ETS.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage.

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

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

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Mills, C. N., & Steffen, M. (2000). The GRE computerized adaptive test: Operational issues. In W. J. van der Linden & C. A. W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 27–52). The Netherlands: Kluwer.

Mislevy, R. J. (1984). Estimating latent distributions. *Psychometrika, 49*, 359–381.

Nemhauser, G. L., & Wolsey, L. A. (1988). *Integer and combinatorial optimization*. New York: Wiley.

Stocking, M. L. (1996). An alternative method for scoring adaptive tests. *Journal of Educational and Behavioral Statistics*, *21*, 365–389.

Stocking, M. L., & Lewis, C. (1995). *A new method for controlling item exposure in computerized adaptive testing* (ETS RR-95-25). Princeton, NJ: ETS.

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., Steffen, M., & Eignor, D. R. (2002). *An exploration of potentially problematic adaptive tests* (ETS RR-02-05). Princeton, NJ: ETS.

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

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

van der Linden, W. J. (2000a). 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.

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

van der Linden, W. J. (2001, November). *Report on CAT algorithm comparison* (Internal Memorandum). Enschede, The Netherlands: Department of Research Methodology, Measurement, and Data Analysis, University of Twente.

van der Linden, W. J. (2003). *A comparison of item-selection methods for adaptive tests with content constraints* (Research Rep. No. 03-06). Enschede, The Netherlands: Department of Research Methodology, Measurement, and Data Analysis, University of Twente.

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., & Glas, C. A. W. (Eds.). (2000). *Computerized adaptive testing: Theory and practice.* Boston, MA: Kluwer Academic Publishers.

van der Linden, W. J., & Reese, L. M. (1998). A model for optimal constrained adaptive testing. *Applied Psychological Measurement, 22*, 259–270.

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

Wang, T., & Vispoel, W. P. (1998). Properties of ability estimation methods in computerized adaptive testing. *Journal of Educational Measurement*, *35*, 109–135.

Ward, W. C. (1988). The College Board computerized placement tests: An application of computerized adaptive testing. *Machine-Mediated Learning, 2*, 217–282.

Way, W. D., & Steffen, M. (1998, April). *Strategies for managing item pools to maximize item security*. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Way, W. D., Swanson, L., Steffen, M., & Stocking, M. L. (2001). *Refining a system for computerized adaptive testing pool creation* (Research Rep. No. 01-18). Princeton, NJ: Educational Testing Service.

Weiss, D. J. (Ed.). (1983). *New horizons in testing: Latent trait test theory and computerized adaptive testing*. (pp. 257–283). New York: Academic Press.

Winston, W. L. (1991). *Operations research: Applications and algorithms.* Boston, MA: PWS-Kent Publishing Company.

Yi, Q., Wang, T., & Ban, J-C. (2001). Effect of scale transformation and test-termination rule on the precision of ability estimation in computerized adaptive testing. *Journal of Educational Measurement*, *38*, 267–292.