DOCUMENT RESUME

ED 473 528

TM 034 736

AUTHOR	Veldkamp, Bernard P.; Ariel, Adelaide
TITLE	Extended Shadow Test Approach for Constrained Adaptive Testing. Research Report.
INSTITUTION	Twente Univ., Enschede (Netherlands). Faculty of Educational Science and Technology.
REPORT NO	RR-02-07
PUB DATE	2002-00-00
NOTE	30p.
AVAILABLE FROM	Faculty of Educational Science and Technology, University of Twente, P.O. Box 217, 7500 AE Enschede, The Netherlands. E-mail: Fox@edte.utwente.nl.
PUB TYPE	Reports - Research (143)
EDRS PRICE	EDRS Price MF01/PC02 Plus Postage.
DESCRIPTORS	*Adaptive Testing; Computer Assisted Testing; Item Banks; *Test Construction; Test Items
IDENTIFIERS	*Constraints

ABSTRACT

Several methods have been developed for use on constrained adaptive testing. Item pool partitioning, multistage testing, and testlet-based adaptive testing are methods that perform well for specific cases of adaptive testing. The weighted deviation model and the Shadow Test approach can be more generally applied. These methods are based on different ideas about how to optimize the performance of computerized adaptive testing (CAT), and they have different advantages. In this paper, both ideas are combined, and a new approach for selecting items in CAT, the Extended Shadow Test approach is presented. The performances of the Extended Shadow Test Approach, the Shadow Test approach, and the weighted deviation model are compared using simulation and existing item pools. The Extended Shadow Test approach shows accurate performance. Recommendations are given about its use. (Contains 4 tables and 22 references.) (Author/SLD)

Extended Shadow Test Approach for Constrained Adaptive Testing

Research Report
02-07

Bernard P. Veldkamp
Adelaide Ariel

faculty of
EDUCATIONAL SCIENCE
AND TECHNOLOGY

University of Twente

Department of
Educational Measurement and Data Analysis

Extended Shadow Test Approach for Constrained Adaptive Testing

Bernard P. Veldkamp

Adelaide Ariel

Abstract

For Constrained Adaptive Testing several methods have been presented. Item Pool Partitioning, Multi-Stage Testing, and Testlet-based Adaptive Testing are methods that perform well for specific cases of adaptive testing. The Weighted Deviation Model and the Shadow Test Approach can be more generally applied. Both methods are based on different ideas about how to optimize the performance of CAT, and these methods have different advantages. In this paper, both ideas are combined and a new approach for selecting the items in CAT, the Extended Shadow Test Approach, is presented. The performances of the Extended Shadow Test Approach, the Shadow Test Approach, and the Weighted Deviation Model are compared. The Extended Shadow Test Approach shows accurate performance. Finally, recommendations about the use are given.

Keywords. Computerized Adaptive Testing, Constraints, Item Selection, Shadow Test Approach, Weighted Deviation Model.

Introduction

The challenge to make computerized adaptive testing (CAT) possible for a test with constraints on the test content has been around for some time. In the early days of adaptive testing, the algorithms mainly concentrated on psychometric aspects of test construction. But when CAT became technically possible, the question became how to incorporate all kinds of test specifications in the assembly process.

One way of implementing the test specifications is to do Multi-Stage Testing (Lord, 1980, Adema, 1990, van der Linden, & Adema, 1998, and Luecht, & Nungester, 1998). In multi-stage testing examinees go through a sequence of subtests. Based on their ability estimates after finishing a subtest, a more difficult subtest or an easier subtest is presented. The subtest are assembled in such a way that all kinds of test specifications will be met. A somewhat similar approach is described in Wainer & Kiely (1987). They propose not to select individual items, but to select intact units of items, the so-called testlets. These testlets are assembled in advance and the examinees should answer the items in the testlet in a pre-specified order. The way examinees proceed to the next testlet can be fixed in advance, like in Multi-Stage testing, or can be left open. Recently some new psychometric theory for Testlet-Based adaptive testing has been developed (Wainer, Bradlow and Du, 2000, Glas, Wainer and Bradlow, 2000). A third strategy to deal with test specifications is presented in Kingsbury and Zara (1991). The general idea underlying this method is to partition the item pool according to the item attributes. During a CAT session the number of items in the different categories are registered. The next item is selected from the categories where the largest number of items still has to be chosen from. The same approach has been applied to the CAT-ASVAB (Segall, Moreno, Bloxom, & Hetter, 1997).

One of the main advantages of Multi-Stage testing and Testlet-based adaptive testing is that a content expert can review the stages or testlets before they are administered. In this way content validity can be guaranteed. However, optimal assembly of stages or testlets is a very complicated optimization problem. Besides, both approaches are not fully adaptive. The main disadvantage of Item Pool Partitioning is that this method can only handle constraints that categorize the items. So, none of these methods is really suitable for all cases of constrained adaptive testing.

In order to do fully adaptive testing with constraints two methods have been proposed in the literature. The Weighted Deviation Model (Stocking & Swanson, 1993) is the oldest method. The heuristic for selecting the items is described in Swanson & Stocking (1993). It has been successfully applied to many CAT assembly problems. For example, Luecht (1996) applied the method to the problem of multidimensional constrained adaptive testing. The philosophy of the method is that the constraints are considered to be desired properties, and the item that provides most information and minimizes the deviation from the desired bounds is selected. The other method is the Shadow Test Approach (van der Linden, & Reese, 1998, van der Linden, 2000). In this method the constraints are a starting point. In order to guarantee content validity, the constraints should be met. Within these limitations an optimal shadow test is assembled from which the next item is selected. This method has also been successfully applied to several problems of constrained CAT. In Veldkamp, & van der Linden (2002), an application to the problem of multidimensional adaptive testing with constraints on test content is described, in Veldkamp (submitted) the method is applied to the problem of assembling constrained polytomous CAT, and in van der Linden, & Chang (submitted) it is applied to the problem of alpha-stratified adaptive testing. However, the STA might encounter problems in case of a non-linear optimization criterion, and in case of variable-length CAT.

Two different philosophies underlie both suitable methods for constrained adaptive testing. For the WDM the focus is on optimally selecting the next item. For the STA the focus is more on guaranteeing that all constraints will be met. The goal of the present research is to find an algorithm that combines both philosophies, and will be applicable to all different forms of computerized adaptive testing.

Theoretical Evaluation of Existing Methods.

Both philosophies handle the constraints in different ways. According to van der Linden (1998), three kinds of constraints can be distinguished. The first category consists of categorical constraints. These constraints can be applied to categorize the items. Examples of categorical constraints are constraints regarding item content, item type, gender orientation, or minority orientation. Quantitative constraints are constraints on a function of the attributes. For quantitative constraints one could think of response time, or word count constraints. The third type of constraints are constraints on dependencies between items. Sometimes, these constraints are called logical constraints. Constraints on enemy sets or item sets are the most common examples. When a formulation of an LP-model is given, these three kinds of constraints will be distinguished. First a description of the Weighted Deviation Model (WDM) is given, then the Shadow Test Approach (STA) is introduced.

Weighted Deviation Model

The WDM was first described in Swanson & Stocking (1993), and applied to the problem of adaptive testing in Stocking & Swanson (1993). In this method, test specifications are not formulated as constraints but as desired properties. These properties are part of the objective function. The advantage of this method is that test practitioners will never run into problems of infeasibility. Whenever some constraints might conflict, the model gives the opportunity to violate one of the constraints. The sum of weighted violations are minimized. The opportunity to weight the different desired properties enables the test assembler to control which constraints are most important and which can be relaxed. The model can be formulated as follows:

$$\min \sum_{j=1}^{J} w_j d_j \tag{1}$$

subject to:

$$\begin{aligned} \sum_{i=1}^{I} I_i(\widehat{\theta}) x_i + d_I &= \infty \qquad (2) \\ \sum_{i=1}^{I} x_i - d_c &\leq n_c \qquad (3) \end{aligned}$$

$$\sum_{i=1}^{I} a_{iq}x_i - d_q \leq n_q \tag{4}$$

$$\sum_{i \in S_e} x_i - d_e \leq 1 \tag{5}$$

$$\sum_{i=1}^{I} x_i = n \tag{6}$$

$$x_i \in \{0,1\}, d_j \geq 0. \tag{7}$$

where w_j denote the weights of the deviations d_j. Equation 2 maximizes the information in the test, and Equations 3 to 5 denote the different kinds of constraints, a_{iq} is the contribution of item i to quantitative constraint q. Equation 6 denotes the test length. The decision variables x_i indicate whether an item is in the test or not. In order to find the next item, mixed integer linear programming techniques can be applied, but it saves time to use a greedy heuristic. It essentially consists of three steps:

1. Compute the deviation of the constraints for every item that is not in the test yet.
2. Sum the weighted deviations across all constraints
3. Select the item with smallest sum of deviations.

Evaluation.

In the WDM method, the focus is on selecting the next item. The item that minimizes the sum of weighted deviations is selected. A deliberation is made between maximizing the information and minimizing violation of constraints. An advantage of the method is that each subsequent item is chosen optimally, given the information in the previously administered items. Especially after the first few items, when the ability estimate is quite stable, this locally optimum item selection is almost globally optimal. Besides, the greedy heuristic is a very fast method. So, even large assembly problems can be solved very quickly. However, one of the drawbacks of this method is the lack of content validity. In CAT different tests are administered to different examinees. In order to make sure that all tests in the same examination program are comparable, content validity needs to be guaranteed. It should be remarked that for item pools of good quality, the problem is small. These item pools have enough items to fulfil the constraints. Choosing proper weights can also enable test assemblers to overcome this drawback.

Shadow Test Approach

The STA was first described in van der Linden & Reese (1998). The basic concept of the method is that the selection of each new item is preceded by an on-line assembly of a shadow test. The shadow test is comparable with a test of full length that meets all the constraints. The shadow test performs optimally at the estimated ability level, and contains all the previously administered items. The next item is selected from the unadministered items in the shadow test. The following pseudo algorithm for the STA is given in van der Linden 2000:

1. Initialize the ability estimator.
2. Assembly a shadow test that meets the constraints and has maximum information at the current ability estimate.
3. Administer the item in the shadow test with maximum information at the current ability estimate.
4. Update the ability estimate.
5. Return all unused items to the pool.
6. Adjust the constraints.
7. Repeat Steps 2 - 6 until n-items have been administered.

In order to assembly the optimal shadow test in Step 2 of the algorithm a mixed integer linear programming approach is used. The problem of assembling a shadow test can be formulated in the following way:

$$\max \sum_{i=1}^{I} I_i(\widehat{\theta}) x_i \tag{8}$$

subject to:

$$\sum_{i \in S_{k-1}} x_i = k - 1 \tag{9}$$

$$\sum_{i=1}^{I} x_i = n \tag{10}$$

$$\sum_{i=1}^{I} x_i \leq n_c \tag{11}$$

$$\sum_{i=1}^{I} a_{iq} x_i \leq n_q \tag{12}$$

$$\sum_{i \in S_e}^{I} x_i \leq 1 \tag{13}$$

$$x_i \in \{0, 1\} \tag{14}$$

where the objective (Equation 8) is to maximize Fisher's information at the current ability estimate. All previous administered items S_{k-1} should be in the test (Equation 9). In Equation 10 the test length is defined. Equations 11 - 13 denote the different kinds of constraints, and the decision variables x_i denote whether an item is in the test or not.

Evaluation

According to van der Linden (2000), the ideal in constrained adaptive testing is that a test is feasible (meets all the constraints) and has an optimal value for the objective function at the estimated ability level of the examinee. For selecting the next item, a shadow test is selected first. In this way it is guaranteed that when the next item is chosen, there always exists a full test that meets all the constraints. So, it can be guaranteed that a feasible test is found and content validity is ensured for different examinees. The second aim of constrained adaptive testing, optimum information at the true ability level of the examinee, is less easy to achieve. It has been shown (van der Linden, 2000) that the tests will converge to the optimal value for the information function at the true ability of the examinee. However, in most adaptive tests the test length is only small, and this result holds in the long run. Therefore, the STA makes sure that content validity holds, and optimal measurement precision can almost be guaranteed.

About applicability to different forms of CAT, some remarks have to be made. The STA is applicable to all major test programs like the LSAT, the GRE, and the GMAT. However, the approach is based on Linear Programming optimization techniques. When the optimization criterion is a non-linear function of the items, like in multidimensional

CAT (e.g. see Segall, 1996), these techniques can no longer be applied and a linear approximation of the criterion has to be applied. Besides, in variable-length CAT the set of constraints changes during the CAT administration. Some constraints are based on an estimate of the ability parameter that changes after every next item. Because of this, the shadow test in one iteration is not always a feasible shadow test in the next iteration. And it can no longer be guaranteed that all constraints will be met,

Extended Shadow Test Approach.

When the existing methods are evaluated, different problems occur both for the WDM and the STA. A new approach is developed to overcome these problems. The approach needs to guarantee that all constraints will be met. Besides, it should be applicable in case of a non-linear objective function, and in case of variable-length CAT. The new approach, the Extended Shadow Test Approach (ESTA), is a combination of both approaches mentioned above. It will select the item that performs optimally at the current ability estimate, and it will assemble a shadow test that contains this items. In this way it will combine the strength of both methods, content validity will hold and measurement precision will increase. The general idea underlying the approach is that the technique of linear programming will be applied in order to guarantee that a full adaptive test that meets all constraints exists. Sofar it looks like the STA. However, the new objective function is not to select a shadow test that performs optimally for the estimated ability level, but to select the single item that performs best for this ability level, like in the WDM. This will make the approach applicable in case of a non-linear optimization criterion, and of a variable-length CAT.

A pseudo algorithm for the Extended Shadow Test Approach is

1. Initialize the ability estimator.
2. Assembly a shadow test that meets the constraints and *contains an item with maximum information at the current ability estimate.*
3. Administer *this item.*
4. Update the ability estimate.
5. Return all unused items to the pool.

6. Adjust the constraints.
7. Repeat Steps 2 - 6 until n-items have been administered.

The only differences between the new method and the STA are in Step 2 and 3 of the pseudo algorithm. The new model for selecting a shadow test in Step 2 is quite different and can be described in the following manner:

$$\max_{i \in I - S_{k-1}} I_i(\widehat{\theta}) x_i \tag{15}$$

subject to

$$\begin{aligned}
\sum_{i \in S_{k-1}} x_i &= k-1 && (16)\\
x_i &\leq y_i && (17)\\
\sum_{i=1}^{I} y_i &= n && (18)\\
\sum_{i=1}^{I} y_i &\leq n_c && (19)\\
\sum_{i=1}^{I} a_{iq} y_i &\leq n_q && (20)\\
\sum_{i \in S_e}^{I} y_i &\leq 1 && (21)\\
x_i, y_i &\in \{0,1\}. && (22)
\end{aligned}$$

Although there seem to be much similarities between this model and the model in Equation 8 - 14, the following differences should be mentioned. The objective function in Equation 15 now selects the k-th item with maximum Fisher's information from the set of unadministered items $I - S_{k-1}$. Besides, in Equation 17 - 21, the variable y_i denotes whether an item is in the shadow test or not, and x_i denotes whether the item is selected to be in the test. When the model in Equation 15 - 22 is solved, it results in a single item. In step 3 of the pseudo algorithm this item is administered to the examinee.

Computational complexity

When the ESTA model is applied, many variables and constraints have to be added. For every item two variables are in the model; x_i denotes whether item i is selected for the CAT, and y_i denotes whether item i is in the shadow test. Besides, for every item i a constraint has to be added to make sure that item i is in the shadow test, when it is selected (Equation 17). The impact of this extension will be shown in the numerical examples.

In general, an increase in the number of variables and constraints results in much more computation time for finding the optimal solution. For the ESTA this is not the case, because of the special structure of the problem. This structure enables implementation of a very fast algorithm. The algorithm consists of the following steps. First, order the items x_i with respect to the amount of information they contribute, from most informative to least informative. Since only one item is selected in every iteration, the first x_i in this row with a feasible shadow test, can be proven to be the optimal solution. Therefore the second step is to solve the following LP for the subsequent items in the row

$$\max y_{i*} \qquad (23)$$

subject to

$$\begin{aligned} \sum_{i \in S_{k-1}} y_i &= k-1 && (24) \\ \sum_{i=1}^{I} y_i &= n && (25) \\ \sum_{i=1}^{I} y_i &\leq n_c && (26) \\ \sum_{i=1}^{I} a_{iq} y_i &\leq n_q && (27) \\ \sum_{i \in S_e}^{I} y_i &\leq 1 && (28) \end{aligned}$$

$$y_i \in \{0,1\}. \tag{29}$$

where i^* denotes the item for which it is checked whether a shadow test exists. When the resulting value for y_{i*} equals 1, a shadow test exists, and the optimal solution is found. Moreover, the complexity of the model in (23)-(29) is very low. As a result, the next item in CAT can be found easily when the ESTA is applied.

Comparison of Methods.

Three case studies were conducted in order to evaluate the use of the new method. In the first study, the STA and the ESTA were applied to assemble a fixed-length CAT, where items were selected based on Fisher information, and a number of constraints had to be met. This is the most general form of CAT. In the second case study, a multidimensional CAT was simulated. A non-linear optimization criterion was applied. The third case, was an example of variable-length CAT.

CASE 1: fixed-length CAT

An item pool for an admission test consists of 2131 items. The item pool fitted the three-parameter logistic model. So, the probability that candidate j obtains a correct answer to item i is defined as

$$P_i(\theta_j) = c_i + (1-c_i)\frac{\exp[a_i(\theta_j - b_i)]}{1+\exp[a_i(\theta_j - b_i)]}, \qquad i = 1, ..., N, \qquad j = 1, ..., M. \tag{30}$$

where θ_j is the ability level of candidate j, a_i the discrimination parameter, b_i the difficulty parameter, and c_i the guessing parameter of item i. For each item, the item type and the word count are known. EAP-estimates are applied for estimating the ability parameters after every iteration. 200 examinees were simulated for every grid point in {-2,-1.5,...,2}. The problem of CAT assembly, can be formulated in general terms as

$$\max \text{ the information in the test}$$

subject to

lower and upper bounds for number of items for each item type

lower and upper bound for total word count

test length constraint

where this problem is solved iteratively, and in every iteration the ability of the examinee is estimated. The next item is selected based on the current ability estimate. For this problem LP models were defined for the STA and the ESTA. The characteristics of the models are described in Table 1.

Model	# variables	# constraints
Shadow Test Approach	2131	20
Extended Shadow Test Approach	4262	2151

Results for the methods are shown in Figure 1. Only small differences in performance were found between the STA and the ESTA. The average bias was at about the same for all grid points, and for low θ-values, the resulting MSEs were slightly smaller for the ESTA. Because of this, it can be concluded that the ESTA performed at least as well as the STA in this simulation study.

==================

Insert Figure 1 at about here

==================

CASE 2: Multidimensional CAT

An item pool from the ACT Assessment Program was used. The item pool consisted of 176 items calibrated under a two-dimensional version of the 2PL model

$$P_i(\boldsymbol{\theta}_j) = \frac{e^{(\mathbf{a}_i \cdot \boldsymbol{\theta}_j + d_i)}}{1 + e^{(\mathbf{a}_i \cdot \boldsymbol{\theta}_j + d_i)}}, \tag{31}$$

where $P_i(\boldsymbol{\theta}_j)$ is the probability that a person $j = 1, ..., J$ with ability vector $\boldsymbol{\theta}_j$ gives a correct answer to an item $i = 1, ..., I$, $\mathbf{a}_i$ is the vector of discrimination parameters of item i along the components of ability $\boldsymbol{\theta}_j = (\theta_{j1}, \theta_{j2})$, and d_i is the scalar parameter representing the easiness of item i. The calibration was carried out using the program NOHARM (Fraser & McDonald, 1988). The items in the pool were classified according to the six content and three skill categories used in the ACT Assessment Program to formulate the test specifications. The problem can be formulated as

max determinant of Fisher's Information Matrix

subject to

lower and upper bounds for number of items for each content class

lower and upper bounds for number of items for each item type

test length constraint

where Fisher's information matrix is of the following form

$$I(\boldsymbol{\theta}) = \begin{bmatrix} \sum_{i=1}^{n} a_{1i}^2 P_i Q_i & \sum_{i=1}^{n} a_{1i} a_{2i} P_i Q_i \\ \sum_{i=1}^{n} a_{1i} a_{2i} P_i Q_i & \sum_{i=1}^{n} a_{2i}^2 P_i Q_i \end{bmatrix}. \qquad (32)$$

A linear approximation of the determinant of Fisher's information matrix was used for the STA (See also, Veldkamp, 2002). The characteristics of the models are described in Table 2.

Model	# variables	# constraints
Weighted Deviation Model	195	19
Shadow Test Approach	176	10
Extended Shadow Test Approach	352	186

The determinant is a function of two continuous variables. A grid of nine points was applied to discretize the two-dimensional space. For each of these grid points a constraint was added in the WDM, therefore the number of constraints for the WDM is equal to the number of constraints of the STA plus nine. The number of constraints for the ESTA is equal to the number of constraints for the STA plus the number of variables. The results of simulating 1000 examinees for every gridpoint are denoted in Table 3.

==================

Insert Table 3 at about here

==================

The STA performed much better than expected. A worse performance was expected because of the application of an approximate optimization criterion. The results are comparable to the results for the ESTA. The ESTA performed slightly better for the first ability parameter, the STA performed slightly better for the second ability parameter. In this simulation study, the WDM performed the best. However, the differences in performance were negligible.

CASE 3: variable-length CAT

The same item pool was applied as for the first case. A bound on the amount of information was used as a stopping criterion. For every grid point in {-2,-1.5,...,2}, the amount of information had to exceed I_{min}. Therefore the CAT assembly problem was slightly changed:

min the number of items in the test

subject to

lower bounds on information for grid points

lower bound on number of items for each item type

lower bound on the total word count

The STA could not be applied to this problem, because the constraints on the information in the test are dependent on the estimated theta. The estimate changes after administering each item. Therefore the set of constraints is different for every iteration, and a shadow test from a previous iteration might not be feasible anymore. The ESTA does not suffer from these problems. In every iteration, one item that provides most information is selected. Because of this, the information is part of the objective function instead of part of the constraints. The methods stop, when the amount of information in the test exceeds the stopping criterion for every grid point. The characteristics of the models are described in Table 4.

Model	# variables	# constraints
Weighted Deviation Model	2152	21
Extended Shadow Test Approach	4262	2141

The model for the ESTA is much more complicated than the model for the WDM. The results of simulating 1000 examinees for every gridpoint are shown in Figure 2.

==================

Insert Figure 2 at about here

==================

The ESTA needed fewer items than the WDM. Besides, application of the ESTA resulted in slightly smaller MSEs. So, it can be concluded that the ESTA outperformed the WDM in this specific simulation study.

Discussion.

Both the WDM and the STA have proven their use for administering constrained CAT. They are effective, although they are based on entirely different ideas. The main difference is the status of constraints. In the WDM, they are viewed as desired properties. A test assembler can variate the weights of the constraints. In this way it can be guaranteed that some constraints will never be violated, while less important constraints are formulated less strictly. On the other hand, the STA is based on the idea that all constraints are equally important, and all of them have to be met in order to guarantee content validity. In this paper, the ESTA is introduced in the same tradition as the STA. In every iteration, a shadow test that checks whether a test exists that meets all the constraints is assembled. However, the method could be changed easily in order to adopt the WDM point of view. When a penalty term is added to one of the constraints in (18) - (21), and this penalty term is added to the objective function with a small weighting factor, this constraint is considered to be of less importance. In fact, even a hybrid model could be used in practise.

The existing methods for constrained CAT encounter some problems. The WDM might result in occasional constraint violation, and several case exist where the STA cannot be applied directly. The ESTA can solve these problems. A shadow test is assembled in every iteration. Because of this, it is guaranteed that all constraints are met. Because the ESTA selects only one item in every iteration, it can also handle non-linear objective functions and variable-length CAT. Because of this, it can be concluded that the ESTA is the most general approach for administering constrained CAT.

When results for the different methods are compared, it can be concluded that the WDM performed the best. In this method, the constraints are less strictly formulated. Because of this, better results can be obtained. When the STA and the ESTA are compared, no real differences in resulting MSEs were found. This result was not expected for CASE 2, because the STA selected items based on an approximation of the optimization criterion. In Veldkamp (2002), the same item bank was applied to assemble linear tests from. In that study, a substantial difference between the WDM and an approach based on linearizing the criterion was found. Because item selection in CAT is based on an estimated ability parameter, the influence of the item selection method is probably smaller. From CASE 1, it can be concluded that the ESTA does not lose precision, although the method focuses on optimal selection of only the next item. In CASE 3 it is shown that the ESTA is very capable of dealing with variable-length CAT.

In the simulation studies, exposure rate control was not applied. The reason is that applying exposure control methods will limit the number of available items. This will result in even smaller differences between the three method. In practise however, these methods need to be applied to guarantee item bank security. In Stocking and Lewis (1997), the problem of exposure control conditional on the ability estimate for the WDM is described. Two methods for dealing with exposure control in the STA are presented in van der Linden and Veldkamp (2002a,2002b).

The ESTA has proven to be an elegant method for administering constrained CAT. It does not suffer from the problems of occasional constraint violation and is very generally applicable. The method seems more complicated, because of the additional variables and constraints in the model. However, it is a promising alternative for the existing methods.

References

Adema, J.J. (1990). The construction of customize two-stage tests. *Journal of Educational Measurement, 27,* 241-253.

Fraser, C. and McDonald, R.P. (1988). *NOHARM II: A FORTRAN program for fitting unidimensional and multidimensional normal ogive models of latent trait theory.* Armidale, Australia: University of New England, Centre for Behavioral Studies.Glas, C.A.W., Wainer, H., & Bradlow, E.T. (2000). MML and EAP estimation in testlet-based adaptive testing. In: W.J. van der Linden, & C.A.W. Glas (Eds.) *Computerized adaptive testing: Theory and practice*, (pp. 271-288). Boston, MA: Kluwer Academic Publishers.

Kingsbury, G.G., & Zara, A.R. (1991). 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: Erlbaum.

Luecht, R.M. (1996). Multidimensional computerized adaptive testing in a certification or licensure context. *Applied Psychological Measurement, 20,* 389-404.

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

Segall, D.O., Moreno, K.E., Bloxom, B.M., & Hetter, R.D. (1997). Psychometric procedures underlying the CAT-ASVAB. In: W.A. Sands, B.K. Waters, & J.R. McBride (Eds.) *Computerized Adaptive Testing: From inquiry to operation.* Washington DC: American Psychological Association.

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. (1998). Optimal assembly of psychological and educational tests. *Applied Psychological Measurement, 22,* 195-211.

van der Linden. W.J. (2000). Constrained adaptive testing with shadow tests. In: W.J. van der Linden, & C.A.W. Glas (Eds.) *Computerized adaptive testing: Theory and practice*, (pp. 27-53). Boston, MA: Kluwer Academic Publishers.

van der Linden, W.J., & Adema, J.J. (1998). Simultaneous assembly of multiple test forms. *Journal of Educational Measurement, 35,* 185-198 [Erratum in Vol. 36, 90-91].

van der Linden, W.J., & Chang, H-H. (2002). Implementing content constraints in alpha-stratified adaptive testing using a shadow test approach. *submitted*

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. (2002a). Sympson-Hetter exposure control in the shadow tests approach. *to be submitted*

van der Linden, W.J., & Veldkamp, B.P. (2002b). Constrained-based item-exposure control in computerized adaptive testing. *to be submitted.*

Veldkamp, B.P. (2002). Item selection in polytomous CAT. *submitted for publication*

Veldkamp, B.P. (2002). Multidimensional Constrained Test Assembly. *Applied Psychological Measurement, 26,* 133-146

Veldkamp, B.P., & van der Linden, W.J. (2002). Multidimensional adaptive testing with constraints on test content. *Psychometrika, 67,* in press

Wainer, H., Bradlow, E.T., & Du, Z. (2000). Testlet response theory: An analog for the 3PL model useful in testlet-based adaptive testing. In: W.J. van der Linden, & C.A.W. Glas (Eds.) *Computerized adaptive testing: Theory and practice*, (pp. 245-270). Boston, MA: Kluwer Academic Publishers.

Wainer, H., & Kiely, G.L. (1987). Item clusters in computerized adaptive testing: A case for testlets. *Journal of Educational Measurement, 24,* 185-201.

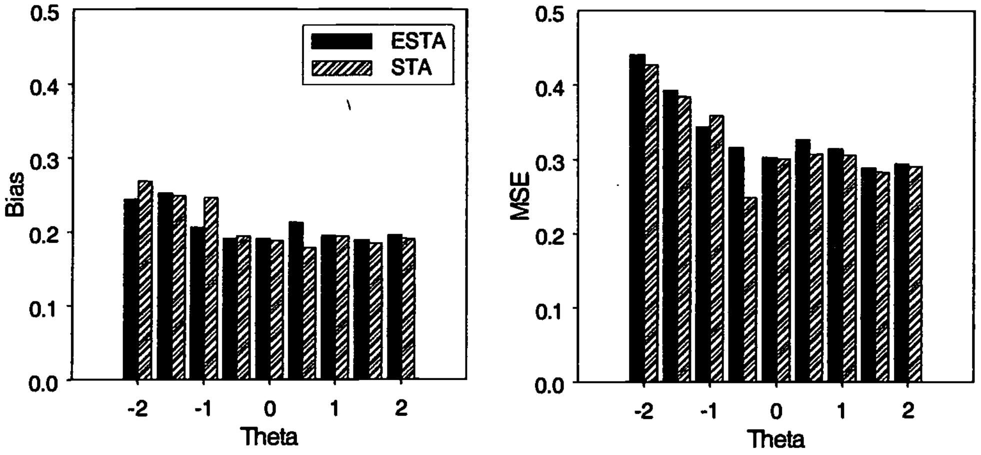
Figure Captions

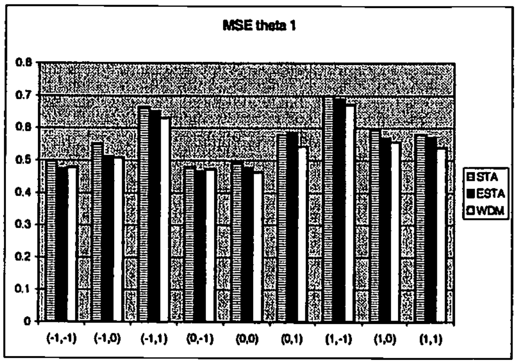
Figure 1. Bias and MSE for both the STA and the ESTA, applied to the problem of fixed-length CAT.

Figure 2. Bias and MSE for the STA, the WDM, and the ESTA, applied to the problem of multidimensional CAT.

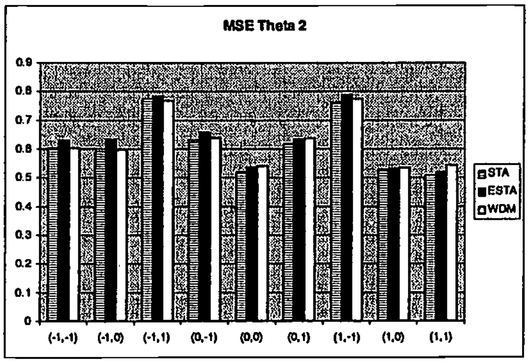
Figure 3. Bias and MSE for the WDM and the ESTA, applied to the problem of variable-length CAT.

Figure 4. Average test length when WDM and STA are applied to the problem of variable-length CAT.

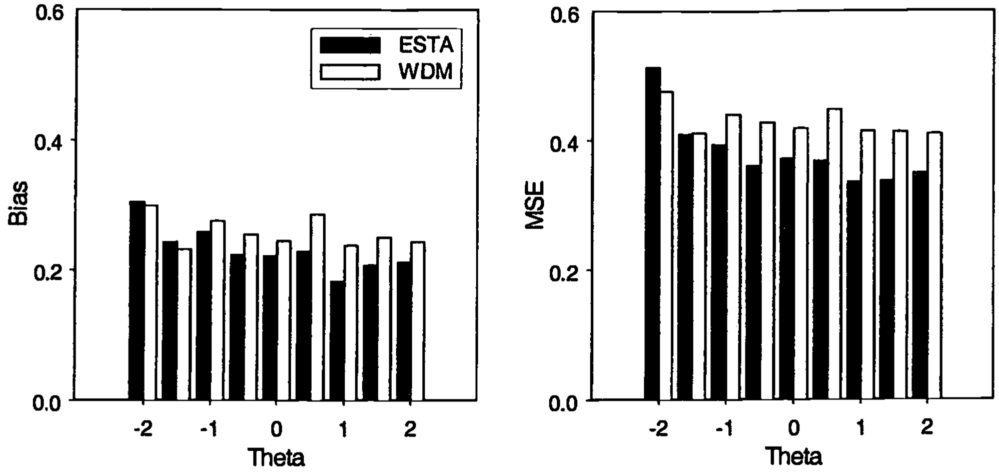

0.5
0.4
0.3
0.2
0.1
0.0
Bias
ESTA
STA
-2
-1
0
1
2
Theta
0.5
0.4
0.3
0.2
0.1
0.0
MSE
-2
-1
0
1
2
Theta

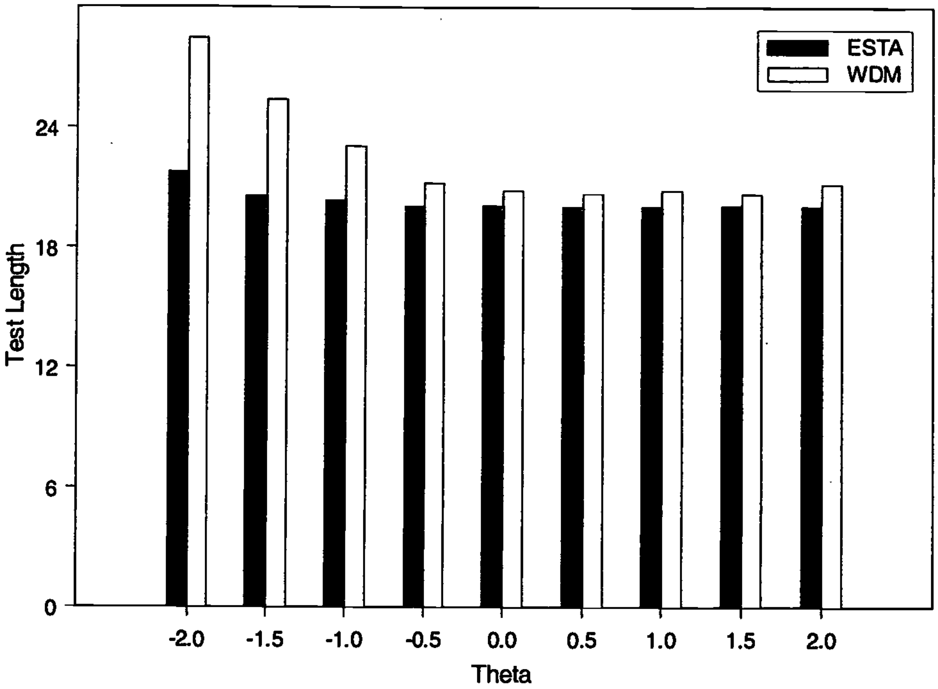


MSE theta 1



MSE Theta 2

0.6
0.4
0.2
0.0
Bias
ESTA
WDM
-2
-1
0
1
2
Theta
0.6
0.4
0.2
0.0
MSE
-2
-1
0
1
2
Theta


ESTA
WDM
Test Length
0
6
12
18
24
-2.0
-1.5
-1.0
-0.5
0.0
0.5
1.0
1.5
2.0
Theta

Titles of Recent Research Reports from the Department of Educational Measurement and Data Analysis. University of Twente, Enschede, The Netherlands.

RR-02-07	B.P. Veldkamp & A. Ariel, *Extended Shadow Test Approach for Constrained Adaptive Testing*
RR-02-06	W.J. van der Linden & B.P. Veldkamp, *Constraining Item Exposure in Computerized Adaptive Testing with Shadow Tests*
RR-02-05	A. Ariel, B.P. Veldkamp & W.J. van der Linden, *Constructing Rotating Item Pools for Constrained Adaptive Testing*
RR-02-04	W.J. van der Linden & L.S. Sotaridona, *A Statistical Test for Detecting Answer Copying on Multiple-Choice Tests*
RR-02-03	W.J. van der Linden, *Estimating Equating Error in Observed-Score Equating*
RR-02-02	W.J. van der Linden, *Some Alternatives to Sympson-Hetter Item-Exposure Control in Computerized Adaptive Testing*
RR-02-01	W.J. van der Linden, H.J. Vos, & L. Chang, *Detecting Intrajudge Inconsistency in Standard Setting using Test Items with a Selected-Response Format*
RR-01-11	C.A.W. Glas & W.J. van der Linden, *Modeling Variability in Item Parameters in Item Response Models*
RR-01-10	C.A.W. Glas & W.J. van der Linden, *Computerized Adaptive Testing with Item Clones*
RR-01-09	C.A.W. Glas & R.R. Meijer, *A Bayesian Approach to Person Fit Analysis in Item Response Theory Models*
RR-01-08	W.J. van der Linden, *Computerized Test Construction*
RR-01-07	R.R. Meijer & L.S. Sotaridona, *Two New Statistics to Detect Answer Copying*
RR-01-06	R.R. Meijer & L.S. Sotaridona, *Statistical Properties of the K-index for Detecting Answer Copying*
RR-01-05	C.A.W. Glas, I. Hendrawan & R.R. Meijer, *The Effect of Person Misfit on Classification Decisions*
RR-01-04	R. Ben-Yashar, S. Nitzan & H.J. Vos, *Optimal Cutoff Points in Single and Multiple Tests for Psychological and Educational Decision Making*
RR-01-03	R.R. Meijer, *Outlier Detection in High-Stakes Certification Testing*
RR-01-02	R.R. Meijer, *Diagnosing Item Score Patterns using IRT Based Person-Fit Statistics*

RR-01-01 H. Chang & W.J. van der Linden, *Implementing Content Constraints in Alpha-Stratified Adaptive Testing Using a Shadow Test Approach*

RR-00-11 B.P. Veldkamp & W.J. van der Linden, *Multidimensional Adaptive Testing with Constraints on Test Content*

RR-00-10 W.J. van der Linden, *A Test-Theoretic Approach to Observed-Score Equating*

RR-00-09 W.J. van der Linden & E.M.L.A. van Krimpen-Stoop, *Using Response Times to Detect Aberrant Responses in Computerized Adaptive Testing*

RR-00-08 L. Chang & W.J. van der Linden & H.J. Vos, *A New Test-Centered Standard-Setting Method Based on Interdependent Evaluation of Item Alternatives*

RR-00-07 W.J. van der linden, *Optimal Stratification of Item Pools in a-Stratified Computerized Adaptive Testing*

RR-00-06 C.A.W. Glas & H.J. Vos, *Adaptive Mastery Testing Using a Multidimensional IRT Model and Bayesian Sequential Decision Theory*

RR-00-05 B.P. Veldkamp, *Modifications of the Branch-and-Bound Algorithm for Application in Constrained Adaptive Testing*

RR-00-04 B.P. Veldkamp, *Constrained Multidimensional Test Assembly*

RR-00-03 J.P. Fox & C.A.W. Glas, *Bayesian Modeling of Measurement Error in Predictor Variables using Item Response Theory*

RR-00-02 J.P. Fox, *Stochastic EM for Estimating the Parameters of a Multilevel IRT Model*

RR-00-01 E.M.L.A. van Krimpen-Stoop & R.R. Meijer, *Detection of Person Misfit in Computerized Adaptive Tests with Polytomous Items*

RR-99-08 W.J. van der Linden & J.E. Carlson, *Calculating Balanced Incomplete Block Designs for Educational Assessments*

RR-99-07 N.D. Verhelst & F. Kaftandjieva, *A Rational Method to Determine Cutoff Scores*

RR-99-06 G. van Engelenburg, *Statistical Analysis for the Solomon Four-Group Design*

RR-99-05 E.M.L.A. van Krimpen-Stoop & R.R. Meijer, *CUSUM-Based Person-Fit Statistics for Adaptive Testing*

RR-99-04 H.J. Vos, *A Minimax Procedure in the Context of Sequential Mastery Testing*

Research Reports can be obtained at costs, Faculty of Educational Science and Technology, University of Twente, TO/OMD, P.O. Box 217, 7500 AE Enschede, The Netherlands.

faculty of

EDUCATIONAL SCIENCE AND TECHNOLOGY

A publication by
The Faculty of Educational Science and Technology of the University of Twente
P.O. Box 217
7500 AE Enschede
The Netherlands