DOCUMENT RESUME

ED 467 367 TM 034 293

AUTHOR	van der Linden, Wim J.; Chang, Hua-Hua
TITLE	Implementing Content Constraints in Alpha-Stratified Adaptive Testing Using a Shadow Test Approach. Research Report.
INSTITUTION	Twente Univ., Enschede (Netherlands). Faculty of Educational Science and Technology.
SPONS AGENCY	Law School Admissions Council, Newtown, PA.
REPORT NO	RR-01-01
PUB DATE	2001-00-00
NOTE	24p.
AVAILABLE FROM	Faculty of Educational Science and Technology, University of Twente, P.O. Box 217, 7500 AE Enschede, The Netherlands.
PUB TYPE	Reports - Research (143)
EDRS PRICE	EDRS Price MF01/PC01 Plus Postage.
DESCRIPTORS	*Adaptive Testing; *Computer Assisted Testing; *Item Banks; Selection; *Test Items
IDENTIFIERS	*Constraints; Stratified Sampling

ABSTRACT

The methods of alpha-stratified adaptive testing and constrained adaptive testing with shadow tests are combined in this study. The advantages are twofold. First, application of the shadow test allows the researcher to implement any type of constraint on item selection in alpha-stratified adaptive testing. Second, the result yields a simple set of constraints that can be used in any application of the shadow test approach to reduce overexposure and underexposure of the items in the pool. An example from the Law School Admission Test is used to demonstrate the advantages. (Contains 20 references and 3 figures.) (Author/SLD)

Implementing Content Constraints in Alpha-Stratified Adaptive Testing Using a Shadow test Approach

Research Report 01-01

Wim J. van der Linden

Hua-Hua Chang
National Board of Medical Examiners

faculty of
EDUCATIONAL SCIENCE AND TECHNOLOGY

University of Twente

Department of
Educational Measurement and Data Analysis

Implementing Content Constraints in Alpha-Stratified Adaptive Testing Using a Shadow Test Approach

Wim J. van der Linden

Hua-Hua Chang
National Board of Medical Examiners

This study received funding from the Law School Admissions Council (LSAC). The opinions and conclusions contained in this paper are those of the other and do not necessarily reflect the policy and position of LSAC. The author are most indebted to Wim M.M. Tielen for his computational assistance. Requests for reprints should be sent to W.J. van der Linden, Department of Educational Measurement and Data Analysis, University of Twente, P.O. Box 217, 7500 AE Enschede, THE NETHERLANDS. Email: w.j.vanderlingen@edte.utwente.nl

Abstract

The methods of alpha-stratified adaptive testing and constrained adaptive testing with shadow tests are combined. The advantages are twofold: First, application of the shadow test approach allows us to implement any type of constraint on item selection in alpha-stratified adaptive testing. Second, the result yields a simple set of constraints that can be used in any application of the shadow test approach to reduce overexposure and underexposure of the items in the pool. An example from the Law School Admission Test is used to demonstrate the advantages.

Key words: alpha-stratification; computerized adaptive testing; item-exposure control; content constraints; shadow test approach

Implementing Content Constraints in Alpha-Stratified Adaptive Testing Using a Shadow Test Approach

Among the practical problems emerged since the first applications of computerized adaptive testing (CAT) in real-life testing programs, the problems of item exposure control and content balancing are most urgent. Adaptive tests that capitalize too much on the presence of a few items in the pool and ignore the others are not only cost ineffective but also bound to run into security problems. Also, if adaptive test administrations show too much variation in content, they are likely to violate important test specifications and the testing program looses its content validity.

Two promising procedures to deal with these problems are alpha-stratified adaptive testing (Chang & Ying, 1999) and constrained adaptive testing with shadow tests (van der Linden, 2000; van der Linden & Reese, 1998). The proposal of alpha-stratified adaptive testing was suggested by the observation that in CAT with maximum-information item selection (van der Linden & Pashley, 2000) the first items typically have high local discrimination, whereas, because of relatively large errors in the θ estimate, lower discrimination over a broader interval would be better (Chang & Ying, 1999). Alpha-stratified adaptive testing forces the CAT algorithm to select items with lower discrimination at the beginning of the test, saving the items with high discrimination for the end of it.

Constrained adaptive testing with shadow tests is a general method to introduce constraints on the item selection process. Though developed originally to implement content constraints on item selection (van der Linden & Reese, 1998), the method is capable to deal with any type of constraint for which a computer algorithm is available. Examples of others than content constraints are response-time constraints to control for differential speededness among examinees in adaptive testing (van der Linden, Scrams, & Schnipke, 1999), constraints on the moments of the item-score distributions to equate observed scores between adaptive tests or an adaptive and a paper-and-pencil test (van der Linden, 2001), and constraints to select among dimensions in mutidimensional adaptive testing (Veldkamp & van der Linden, submitted).

This paper combines the two methods of adaptive testing. The combination turns out to have two advantages. The use of the shadow test allows us to implement virtually any type of constraint on item selection in alpha-stratified adaptive testing. In addition, the constraints needed to model alpha-stratified adaptive testing constitute a simple set of mathematical (in)equalities. This set can be used in any other application of the shadow test approach to reduce overexposure and underexposure of the items in the pool.

Alpha-Stratified CAT

The fact that highly-discriminating items may be suboptimal in the presence of errors in the estimates of θ has been ignored in much of the literature on CAT. Nevertheless, the phenomenon was already known in classical test theory (CCT) under the name of "attenuation paradox", where it was shown that an increase in item-criterion correlation may imply a paradoxical decrease in the predictive validity of the tests if the items are unreliable. The analogy with the current problem arises when noticing the relations between item reliability (CCT) and item information (IRT) and between item validity (CCT) and item-ability correlation (item discrimination parameter in IRT) (Lord & Novick, 1968, 16.5).

Using an item-selection algorithm in CAT that always picks items with maximum discrimination at *all* θ estimates has in fact three disadvantages: (1) As already argued, the choice is likely to be suboptimal at the beginning of the test where the larger errors in the estimates of θ occur; (2) When the θ estimate converges towards the end of the test, selection with maximum discrimination becomes optimal, but then some of the best items in the pool are likely to have already been used; (3) Selecting items with maximum discrimination tends to capitalize on estimation errors in the discrimination parameter, with potentially serious effects on the estimation of θ even for calibration samples of moderate sizes (van der Linden & Glas, 2000).

In alpha-stratified adaptive testing, the item pool is stratified on the values of the item discrimination parameter. Suppose that R different strata are used, each indexed by a value of $r = 1, ..., R$, where a lower value of r indicates a stratum with lower values for the discrimination parameter. Further, suppose that the test consists of n items and that n_r items are selected from stratum r ($\sum_r n_r = n$). The order of the strata from which the items are selected is then $1, ..., R$. Within each stratum, the items are selected to have the smallest distance between the value of their difficulty parameter, b_i, and the current estimate of θ.

Observe that the order in which the strata are used leads towards more uniform exposures rates of the items, particularly if the strata in the item pool are chosen to have equal size and $n_r \equiv n/R$. Alpha-stratified adaptive testing thus has the potential of more favorable item-exposures rates in combination with a statistically more natural item selection criterion. This expectation has been confirmed in studies, for example, by Chang and Ying (1999) and Parshall, Hogarty and Kromrey (1999).

Though generally low and tending to uniformity, the exposures rate of the items alpha-stratified adaptive testing do not automatically meet a previously set upper bound. An unfavorable combination of size of pool, distribution of the item parameter values, number of strata, and test length may lead to higher than desirable exposure rates for some of the items.

In practice, the principle of alpha-stratified adaptive testing can therefore be used to increase the effectiveness of the Sympson-Hetter (1985) method of exposure control. The success of the latter, which is further described below, also depends on the size and composition of the pool. In addition, even for this method and a favorable pool of items, no formal proof exists of the exposure rates converging to values below a previously set bound for each item (see further below). In practice, however, with the possible exception of an occasional item, the method has been proven to be meet reasonable bounds for reasonable item pools, especially if the version conditional on θ proposed by Stocking and Lewis (1998, 2000) is applied.

Application of the principle of alpha-stratification improves the results by the Sympson-Hetter method for two reasons: (1) The Sympson-Hetter method does not address the problem of the large number of underused items in the pool, whereas alpha-stratification does; (2) The method eliminates all items that are selected from the pool but not administered. As a result, in a typical application with the maximum-information criterion, at the end of the test the number of highly discriminating items left near the examinee's true value of θ may have been reduced by a factor 3-5. However, if the Sympson-Hetter method is applied in combination with alpha-stratified CAT, all best items are still available when the last section of the test is reached.

Two remaining problems for alpha-stratified adaptive testing are how to stratify the item pool and balance test content across examinees (Stocking, 1998). The first problem is addressed in a companion paper (Chang & van der Linden, submitted), where the technique of network-flow programming is used to assign items optimally to strata, the objective being uniform distributions both of the discrimination parameter between strata and the difficulty parameter within each stratum. The second problem is addressed in the remainder of this paper.

Constrained CAT with Shadow Tests

The key idea underlying the shadow test approach is that items are not selected directly from the pool but from a shadow test. Shadow tests are a full-size tests assembled prior to each item in the adaptive test that have the following properties: (1) they contain all items already administered to the examinee; (2) they are optimal at the current θ estimate of the examinee; and (3) they meet all specifications the adaptive test has to meet. The item that is actually administered to the examinee is the one in the shadow test that has not yet been administered and is optimal at the θ estimate. After the item is administered, the shadow test is returned to the pool, the θ estimate is updated, and the procedure is repeated.

The only modification of the traditional CAT algorithm needed to execute a shadow test approach is a call to a test assembly algorithm prior to the selection of the item. Nevertheless, this modification guarantees two important features of the adaptive test. First, because each shadow test meets all test specifications, the adaptive test always meets all specifications. Second, because each shadow test is assembled to be optimal at the current $\widehat{\theta}$, and each item actually administered is the one in the shadow test optimal at the same $\widehat{\theta}$, the adaptive converges to optimality at the true θ value of the examinee. Observe that these features hold generally, that is, independent of the set of test specifications and the criterion of optimality chosen. For a more complete introduction to the shadow test approach, technical aspects of its implementation, and applications to item pools from large-scale testing programs, see van der Linden (2000).

Though any test assembly algorithm or heuristic could be used, this paper focuses on the class of algorithms based on a 0-1 linear (LP) or mixed integer programming (MIP) approach to test assembly. Key in the approach is the definition of decision variables for the selection of the items in the test. In 0-1 LP-based test assembly, typically variables x_i are defined to be equal to one if item i is selected in the test and equal to zero if it is not, where $i = 1, ..., I$ is the set of indices denoting the items in the pool. Constraints on the item selection process are linear equalities and/or inequalities imposed on the values of the decision variables. Content constraints mostly take one of two possible forms, depending on whether the attributes of the items that need to be constrained are categorical or quantitative. If the attributes are categorical (e.g., as a content classification, learning taxonomy, or behavioral description) the set of attributes introduces a partition in the item pool that can be denoted as the class of sets V_g, $g = 1, ..., G$ and the constraints take the form

$$\sum_{i \in V_g} x_i \gtreqless n_g,\ g = 1, ..., G. \tag{1}$$

If the attributes are quantitative parameters q_i (e.g., response times, word counts, item information), each constraint takes the form

$$\sum_{i=1}^{I} q_i x_i \gtreqless n. \tag{2}$$

In addition, an objective function is defined on the variables that is maximized or minimized during the item selection process. For example, if the objective is to maximize Fisher's information in the test at the examinee's current estimate, $\widehat{\theta}$, the objective function is

$$\max \sum_{i=1}^{I} I_i(\widehat{\theta}) x_i, \tag{3}$$

where $I_i(\widehat{\theta})$ is the information in the response to item i at $\widehat{\theta}$.

The model can be solved for optimal values of the decision variables using one of the algorithms available in software packages for LP. The package used by the authors to solve the examples later in this paper was CPLEX 6.6 (ILOG, 2000), one of the fastest packages currently available to solve test assembly problems for item pools of the size typically used in large-scale testing programs. For a review of the various test assembly problems that can be solved using 0-1 LP and the technical details of their solutions, the reader should refer to van der Linden (1998).

Modeling Alpha-Stratified CAT

The item response theory (IRT) model used in the examples later in this paper was the three-parameter logistic (3PL) model

$$p_i(\theta) = \Pr\{U_i = 1\} \equiv c_i + (1 - c_i)\frac{\exp[a_i(\theta - b_i)]}{1 + \exp[a_i(\theta - b_i)]}, \tag{4}$$

where U_i is the response variable for item i, with $U_i = 1$ for a correct and $U_i = 0$ for an incorrect response, $\theta \in R$ is the ability of the examinee, and $a_i \in (0, \infty)$, $b_i \in R$, and $c_i \in [0, 1)$ are the discrimination, difficulty, and guessing parameter for item i, respectively.

Let i_k be the index of the item in the pool administered as the kth item in the adaptive test ($k = 1, ..., n$). Assume that $k - 1$ items have already been administered and that stratum r is active when item k is selected. The estimator of θ after $k - 1$ items is denoted as $\widehat{\theta}_{k-1}$. The shadow test assembled for the selection of the kth item is denoted as $(i_1, ..., i_{k-1}, i'_k, ..., i'_n)$, where $C_{k-1} \equiv \{i_1, ..., i_{k-1}\}$ is the set of items already administered and $F_k \equiv \{i'_k, ..., i'_n\}$ is the set of free items. The kth item is selected from the set $Q_r \cap F_k$.

In alpha-stratified adaptive testing the kth item is selected to have a value for the difficulty parameter, b_i, closest to $\widehat{\theta}_{k-1}$. Thus, a natural objective for the shadow test is to selects the set of n_r items from Q_r that have minimum distance to $\widehat{\theta}_{k-1}$. This objective is realized by requiring this set to have b_i values in the interval $(\widehat{\theta}_{k-1} - y, \widehat{\theta}_{k-1} + y)$, where y a nonnegative real-valued decision variable that is minimized.

The model becomes for the kth item becomes:

$$\min y \tag{5}$$

subject to

$$(b_i - \widehat{\theta}_{k-1})x_i \leq y, \; i \in Q_r, \tag{6}$$

$$(b_i - \widehat{\theta}_{k-1})x_i \geq -y, \; i \in Q_r, \tag{7}$$

$$\sum_{i \in Q_r} x_i = n_r,\ r = 1, ..., R, \tag{8}$$

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

$$\sum_{i \in V_g} x_i \gtreqless n_g,\ g = 1, ..., G, \tag{10}$$

$$\sum q_i^h x_i \gtreqless n_h,\ h = 1, ..., H, \tag{11}$$

$$y \geq 0, \tag{12}$$

$$x_i \in \{0, 1\},\ i = 1, ..., I. \tag{13}$$

The interval $(\widehat{\theta}_{k-1} - y, \widehat{\theta}_{k-1} + y)$ for the items in Q_r is defined in (6)-(7), whereas the size of the interval is minimized in (5). The constraints in (8) require the solution to have n_r items from each stratum r. The decision variables of the items already selected are set to one in (9). The constraints in (10)-(11) represents the sets of categorical and quantitative content constraints to be imposed on the item selection process. Finally, in (12)-(13) the ranges of possible values for the decision variables are defined.

The kth test selected in the adaptive test is

$$i_k \equiv \arg\min_i \left\{ \left| b_i - \widehat{\theta} \right| \mid i \in Q_r \cap F_k \right\}. \tag{14}$$

Modifications of Sympson-Hetter Method

The Sympson-Hetter method of exposure control (1985) is based on a distinction between the events of selecting item i for administration from the pool and actually administering the item. We denote these events as S_i and A_i, and their probabilities as $P(S_i)$ and $P(A_i)$, respectively. Because A_i implies S_i, it holds that

$$P(A_i) = P(A_i, S_i) = P(A_i \mid S_i) P(S_i). \tag{15}$$

For a given CAT procedure it is thus possible to lower exposure rate of item $P(A_i)$ relative to $P(S_i)$ by choosing $P(A_i \mid S_i) < 1$. The idea can be implemented by ordering the items according to their value for the item-selection criterion at $\widehat{\theta}_{k-1}$, selecting the first item, and conducting a probability experiment that determines with probability $P(A_i \mid S_i)$ if the item will be administered. If the item is not administered, it is removed from the pool during the rest of the test. In principle, it may be necessary to run a long list of experiments before an item is administered. Stocking and Lewis (1998) proposed an equivalent probability experiment that picks one item for administration from a list of fixed length with probabilities with sizes relative to those of the control parameters.

To adjust $P(A_i \mid S_i)$ to a rate lower than a maximum rate r_i selected by the test administrator, an iterative series of simulation studies is run in which the probabilities $P(S_i)$ and $P(S_i)$ are estimated and the values of the control parameters $P(A_i \mid S_i)$ adjusted. Let $P^{(t)}(S_i)$ and $P^{(t)}(A_i)$ denote the probabilities at Step t. The values of $P(A_i \mid S_i)$ for the next step are then adjusted by the following rule:

$$P^{(t+1)}(A_i \mid S_i) = \begin{cases} 1 & \text{if } P^{(t)}(A_i) \leq r\ , \\ r/P^{(t)}(S_i) & \text{if } P^{(t)}(A_i) > r. \end{cases} \tag{16}$$

Observe that the equality in (15) only holds within Step t, but that (16) is based on the assumption of the same equality for the probabilities between steps. However, the assumption is invalid; for example, the actual value of $P(A_i)$ does depend not only the values of $P(A_j \mid S_j)$ and $P(S_j)$ in the previous step for item $j = i$, but also on those for items $j \neq i$. For this reason, convergence of the adjustments to values below r_i is not guaranteed. However, as already noted, in practice for a reasonable CAT procedure and item pool, the method shows convergence for nearly all of the items.

Two modifications of the Sympson-Hetter method are needed to apply the method to alpha-stratified CAT implemented through the shadow test approach. First, the list of items from which an item is picked for administration is now defined as the set of free items in the shadow test, F_k, ordered by the distance of their value for b_i to $\widehat{\theta}_{k-1}$. Second, because the Sympson-Hetter method removes all previously selected items not administered from the pool, it holds that for a combination of a poorly designed pool, tight sets of constraints in (10)-(11), and long adaptive tests with low maximum exposure rates r_i, the model in (6)-(13) may not always have a solution towards the end of the test for each examinees, that is, the test assembly problem may become infeasible. The problem is fixed by storing all items that are selected but not administered in a separate set. Let R_{k-1} denote this set if $k-1$ items have been administered. If infeasibility occurs when assembling the shadow test for item k, set R_{k-1} is added to the pool temporarily, and a solution always exist.

Simulation Study

A simulation study was conducted to assess the impact of the following choices both on the statistical properties of the final estimator, $\widehat{\theta}_n$ and the exposures rates of the items:

(1) Alpha-stratified CAT vs. maximum-information CAT;

(2) CAT without vs. with content constraints on item selection;

(3) CAT without vs. with Sympson-Hetter exposure control.

All possible combinations of choices were examined. The total number of conditions in the study was thus equal to 8.

Item Pool and Test Specifications

The item pool and test specifications were taken from the Law School Admission Test (LSAT). The item pool was a previous pool consisting of 753 items. In all, 65 categorical and quantitative constraints were needed to model the content specifications for the LSAT. The length of the adaptive test was set equal to 50 items, which is half the length of the current paper-and-pencil version of the LSAT. The right-hand side coefficients in the content constraints in (10)-(11) were reduced proportionally.

The item pool was divided into $R = 5$ strata of equal size with the 20% of the items with the lowest value for the discrimination parameter in Stratum 1, the next 20% in Stratum 2, etc. From each stratum $n_r = 10$ items were selected for the adaptive tests .

Adaptive Tests

In the conditions with alpha-stratified CAT, a test assembly model with the objective function in (5) and the associated constraints in (6)-(7) was used. For CAT with maximum-information item selection, the objective function and constraints were replaced by the objective function in (3). Maximum-information item selection was thus also implemented through a shadow test approach. The conditions with the content constraints were realized by added the set of 65 constraints from the LSAT in (10)-(13) to the test assembly model. Finally, the Sympson-Hetter method was used with the modifications described in the previous section and for all items a target exposure rate of $r_i = .20$.

Adaptive test administrations were simulated for θ =-2.0, -1.5, ..., 2.0, with 2500 replications for each θ value. The initial value of $\widehat{\theta}$ was set equal to 0. The next estimates were EAP estimates with a noninformative prior. The shadow tests were obtained through calls to the CPLEX 6.6 software referred to earlier.

Results

The bias and MSE functions of the ability estimator in the two main types of CAT in the study are displayed in Figure 1 and 2. Ideally, bias functions have negligibly small values uniformly over θ. This ideal was met for all functions in the conditions with alpha-stratified CAT. The same holds for maximum-information CAT, with the exception of the condition with Sympson-Hetter item-exposure control. In this case, after 20 items the lower end of the ability scale showed a negative bias, with considerable size at θ=-2.0. However, after the full test of 50 items in this condition bias was generally reduced to a very low level.

[Figure 1-2 about here]

All MSE functions in Figure 2 run horizontally, with the exception of those for maximum-information CAT with Sympson-Hetter item-exposure control at n=20. The exception points at the bias component obtained for this condition already shown in Figure 1. As expected, the MSE functions at n=50 items were much lower than those at n=20. Also, the functions for maximum-information CAT were lower than those for alpha-stratified CAT. However, for n=50 items, both types of CAT showed satisfactory MSE. For the condition with alpha-stratified CAT at n=20, it should be noted that at this stage only the first two strata, with the items with the lowest discrimination in the pool, were covered. A genuine 20-item alpha-stratified CAT would have consisted of five different strata of five items each. Thus, the relatively large MSE in this condition should not come as a surprise.

Generally, imposing content constraints on an item selection process tends to produce poorer ability estimates than unconstrained item selection from the same pool. However, in spite of the large number of constraints for both types of CAT hardly any increase in MSE was observed. The most likely explanation for this phenomenon is the quality of the item pool. The items in this pool were carefully written according to the content specifications for the LSAT. Hence, the shadow test algorithm did not have to force item selection much to meet the constraints.

[Figure 3 about here]

In Figure 3, the empirical exposure rates of the items are presented in a decreasing order. For all conditions, the rates for alpha-stratified CAT were much more uniform than those for maximum-information CAT. The addition of Sympson-Hetter item-exposure control to the procedure had a favorable impact on maximum-information CAT, but the resulting rates were still much more unfavorable than those for alpha-stratified CAT.

Discussion

Large numbers of content constraints can easily be implemented in alpha-stratified CAT through a shadow-test approach. For a well-designed item pool, such as the one from the LSAT in the empirical study, imposing content constraints on the item selection do not need to have any disadvantageous impact on the statistical properties of the ability estimator. Relative to maximum-information CAT, alpha-stratification tends to result in much more favorable exposures rates for the items. The rates for the popular items are likely to be reduced considerably and, equally important, those for the unpopular items to go up to much more acceptable levels. The price to be paid for this result is a slight loss in the accuracy of the estimator. However, from a practical point of view, this loss can be compensated for by adding a few items to the test, whereas loss due to item compromise or inefficient item use is more difficult to compensate.

References

Chang, H., & van der Linden, W. J. (submitted). Optimal stratification of item pools in alpha-stratified computerized adaptive testing. *Applied Psychological Measurement.*

Chang, H., & Ying, Z. (1999). A-stratified multistage computerized adaptive testing. *Applied Psychological Measurement, 23,* 211-222.

Chang, H., Qian, J., & Ying, Z. (in press). A-stratified multistage CAT with b-blocking. *Applied Psychological Measurement.*

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications.* Boston: Kluwer-Nijhoff.

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

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.

Parshall, C., Hogarty, K., & Kromrey, J. (1999, June). *Item exposure in adaptive tests: An empirical investigation of control strategies.* Paper presented at the Annual Meeting of the Psychometric Society, Lawrence, KS.

Stocking, M. L. (1998). *A framework for comparing adaptive test designs* (Unpublished manuscript). Princeton, NJ: Educational Testing Service

Stocking, M. L., & Lewis, C. (1998). Controlling item exposure conditional on ability in computerized adaptive testing. *Journal of Educational and Behavioral Statistics, 23,* 57-75.

Stocking, M. L., Lewis, C. (2000). Methods of controlling the exposure of items in CAT. In W. J. van der Linden & C. A. W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 163-182). Boston: Kluwer.

Sympson, J. B., & Hetter, R. D. (1985, October). Controlling item-exposure rates in computerized adaptive testing. *Proceedings of the 27th annual meeting of the Military Testing Association* (pp. 973-977). San Diego, CA: Navy Personnel Research and Development Center.

van der Linden, W.J. (Ed.) (1998). Optimal test 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. 27-52). Boston: Kluwer.

van der Linden, W.J. (2001). Adaptive testing with equated number-correct scoring. *Applied Psychological Measurement, 25.* (In press)

van der Linden, W.J. & Glas, C.A.W. (2000). Capitalization on item calibration error in adaptive testing. *Applied Measurement in Education, 12*, 35-53.

van der Linden, W. J., & Pashley, P. J. Item selection and ability estimation in adaptive testing. In W. J. van der Linden & C. A. W. Glas (Eds.) (2000). *Computerized adaptive testing: Theory and practice* (pp. 1-25). Norwell, MA: Kluwer Academic Publishers.

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

van der Linden, W. J., Scrams, D. J., & Schnipke, D. L. (1999). Using response-time constraints to control for differential speededness in adaptive testing. *Applied Psychological Measurement, 23*, 195-210.

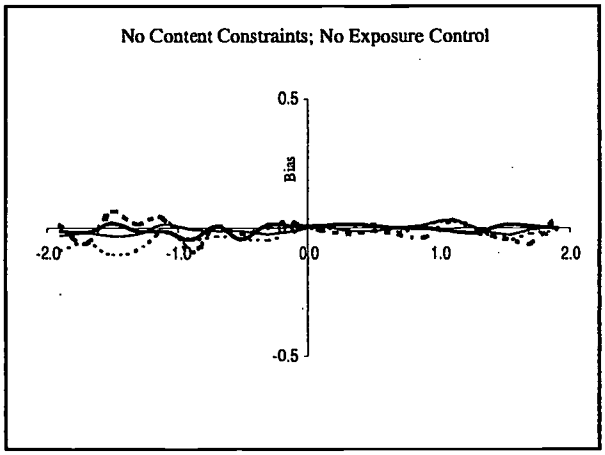
Veldkamp, B.P., & van der Linden, W.J. (submitted). Multidimensional adaptive testing with constraints on test content. *Psychometrika*.

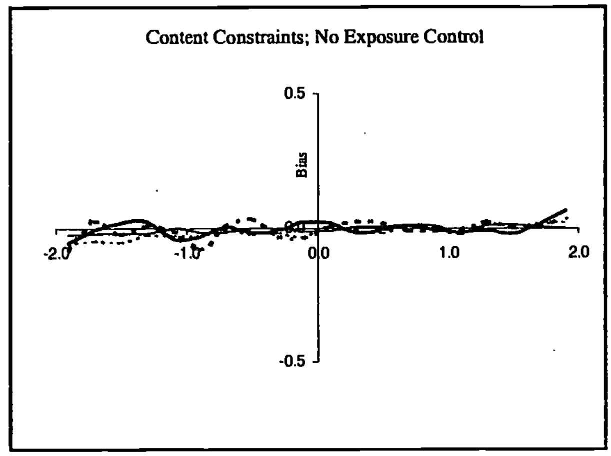
Figure Captions

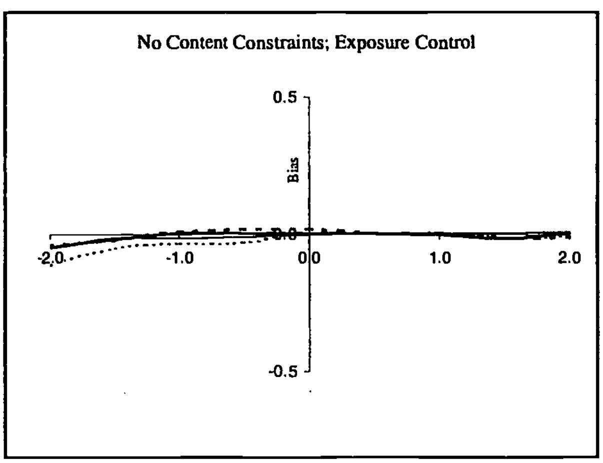
Figure 1. Bias functions for alpha-stratified (bold lines) and maximum-information CAT (thin lines) after n=20 (dashed lines) and n=50 items (solid lines) under the conditions with/without content constraints and with/without Sympson-Hetter item-exposure control.

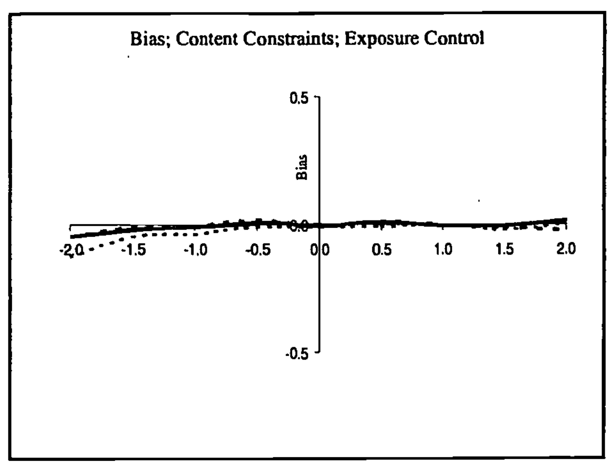
Figure 2. MSE functions for alpha-stratified (bold lines) and maximum-information CAT (thin lines) after n=20 (dashed lines) and n=50 items (solid lines) under the conditions with/without content constraints and with/without Sympson-Hetter item-exposure control.

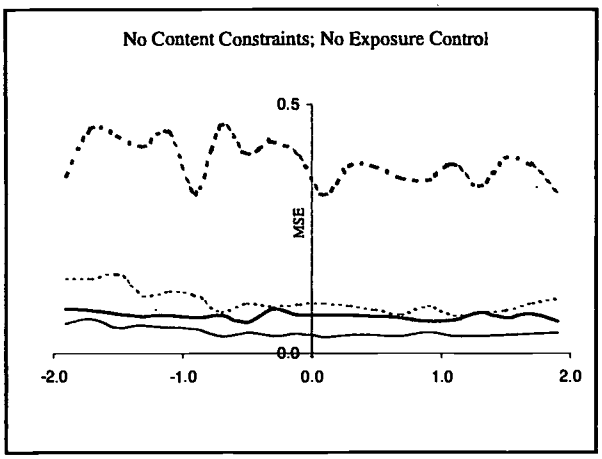
Figure 3. Item exposure rates for alpha-stratified (bold lines) and maximum-information CAT (thin lines) under the conditions with/without content constraints and with/without Sympson-Hetter item-exposure control.

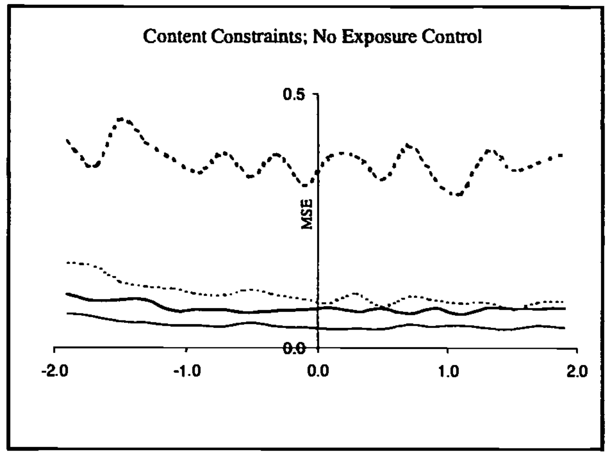

No Content Constraints; No Exposure Control
0.5
Bias
-2.0
-1.0
0.0
1.0
2.0
-0.5

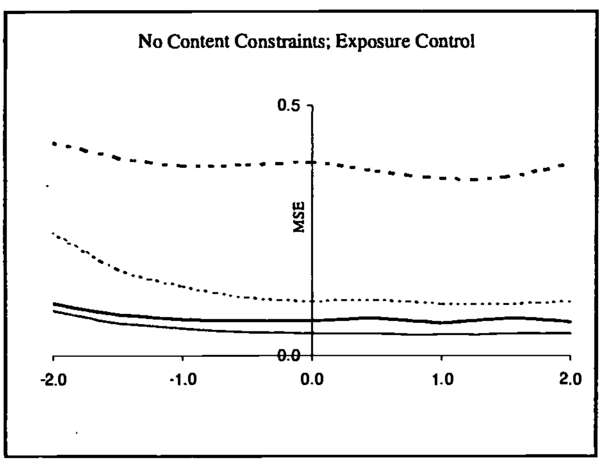

Content Constraints; No Exposure Control
0.5
Bias
-2.0
-1.0
0.0
1.0
2.0
-0.5

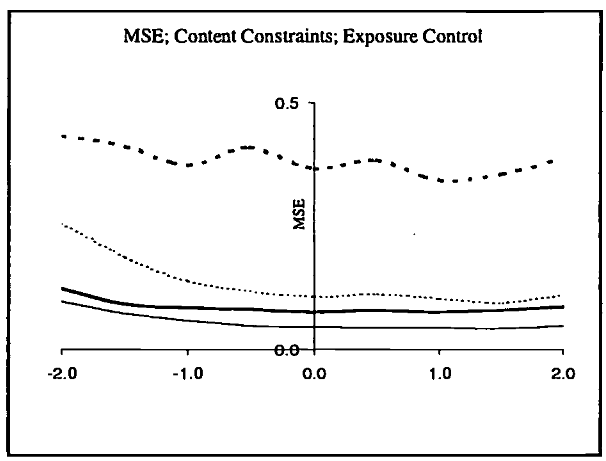

No Content Constraints; Exposure Control
0.5
Bias
0.0
-2.0
-1.0
0.0
1.0
2.0
-0.5

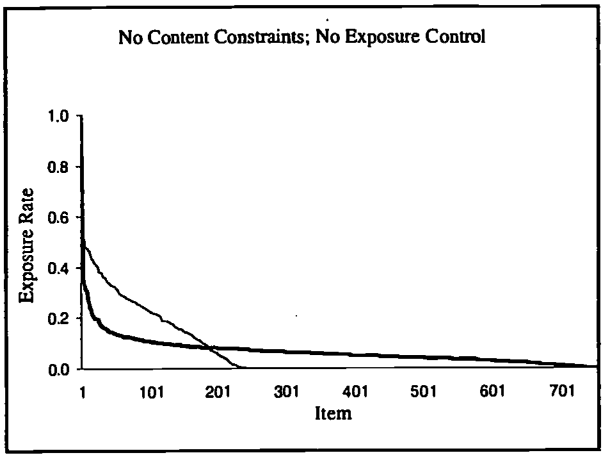

Bias; Content Constraints; Exposure Control
0.5
Bias
0.0
-2.0
-1.5
-1.0
-0.5
0.0
0.5
1.0
1.5
2.0
-0.5

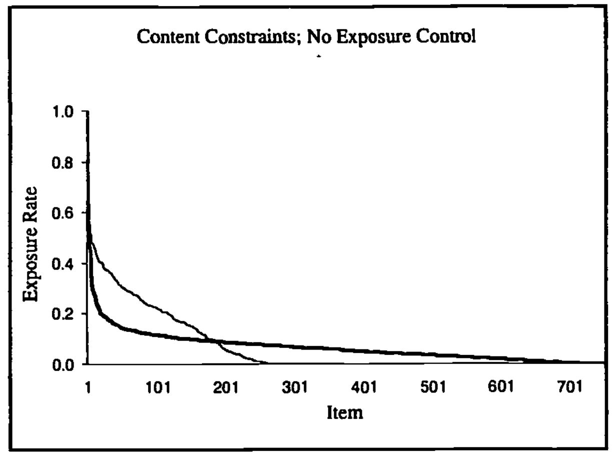

No Content Constraints; No Exposure Control
0.5
MSE
0.0
-2.0
-1.0
0.0
1.0
2.0

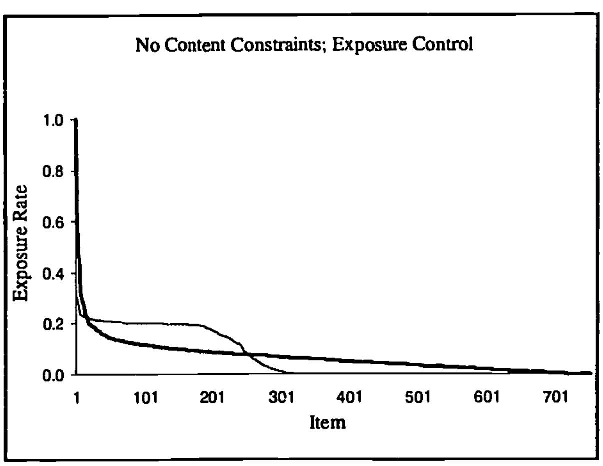

Content Constraints; No Exposure Control
0.5
MSE
0.0
-2.0
-1.0
0.0
1.0
2.0

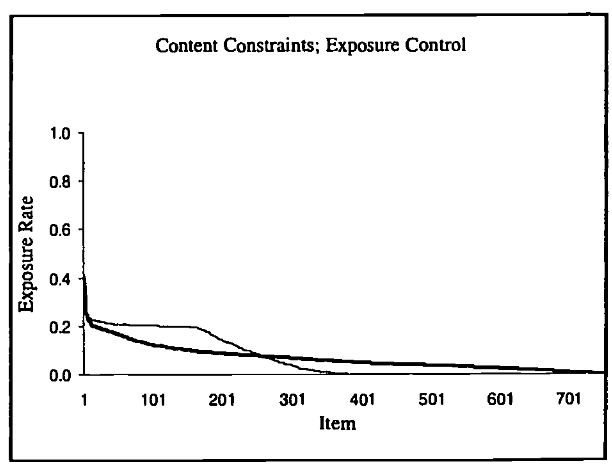

No Content Constraints; Exposure Control
0.5
MSE
0.0
-2.0
-1.0
0.0
1.0
2.0


MSE; Content Constraints; Exposure Control
0.5
MSE
0.0
-2.0
-1.0
0.0
1.0
2.0


No Content Constraints; No Exposure Control
Exposure Rate
1.0
0.8
0.6
0.4
0.2
0.0
1
101
201
301
401
501
601
701
Item


Content Constraints; No Exposure Control
Exposure Rate
1.0
0.8
0.6
0.4
0.2
0.0
1
101
201
301
401
501
601
701
Item


No Content Constraints; Exposure Control
Exposure Rate
1.0
0.8
0.6
0.4
0.2
0.0
1
101
201
301
401
501
601
701
Item


Content Constraints; Exposure Control
Exposure Rate
1.0
0.8
0.6
0.4
0.2
0.0
1
101
201
301
401
501
601
701
Item

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

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*
RR-99-03	B.P. Veldkamp & W.J. van der Linden, *Designing Item Pools for Computerized Adaptive Testing*

RR-99-02 W.J. van der Linden, *Adaptive Testing with Equated Number-Correct Scoring*

RR-99-01 R.R. Meijer & K. Sijtsma, *A Review of Methods for Evaluating the Fit of Item Score Patterns on a Test*

RR-98-16 J.P. Fox & C.A.W. Glas, *Multi-level IRT with Measurement Error in the Predictor Variables*

RR-98-15 C.A.W. Glas & H.J. Vos, *Adaptive Mastery Testing Using the Rasch Model and Bayesian Sequential Decision Theory*

RR-98-14 A.A. Béguin & C.A.W. Glas, *MCMC Estimation of Multidimensional IRT Models*

RR-98-13 E.M.L.A. van Krimpen-Stoop & R.R. Meijer, *Person Fit based on Statistical Process Control in an AdaptiveTesting Environment*

RR-98-12 W.J. van der Linden, *Optimal Assembly of Tests with Item Sets*

RR-98-11 W.J. van der Linden, B.P. Veldkamp & L.M. Reese, *An Integer Programming Approach to Item Pool Design*

RR-98-10 W.J. van der Linden, *A Discussion of Some Methodological Issues in International Assessments*

RR-98-09 B.P. Veldkamp, *Multiple Objective Test Assembly Problems*

RR-98-08 B.P. Veldkamp, *Multidimensional Test Assembly Based on Lagrangian Relaxation Techniques*

RR-98-07 W.J. van der Linden & C.A.W. Glas, *Capitalization on Item Calibration Error in Adaptive Testing*

RR-98-06 W.J. van der Linden, D.J. Scrams & D.L.Schnipke, *Using Response-Time Constraints in Item Selection to Control for Differential Speededness in Computerized Adaptive Testing*

RR-98-05 W.J. van der Linden, *Optimal Assembly of Educational and Psychological Tests, with a Bibliography*

RR-98-04 C.A.W. Glas, *Modification Indices for the 2-PL and the Nominal Response Model*

RR-98-03 C.A.W. Glas, *Quality Control of On-line Calibration in Computerized Assessment*

RR-98-02 R.R. Meijer & E.M.L.A. van Krimpen-Stoop, *Simulating the Null Distribution of Person-Fit Statistics for Conventional and Adaptive Tests*

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

NOTICE

Reproduction Basis

X	This document is covered by a signed "Reproduction Release (Blanket)" form (on file within the ERIC system), encompassing all or classes of documents from its source organization and, therefore, does not require a "Specific Document" Release form.
	This document is Federally-funded, or carries its own permission to reproduce, or is otherwise in the public domain and, therefore, may be reproduced by ERIC without a signed Reproduction Release form (either "Specific Document" or "Blanket").