DOCUMENT RESUME

ED 450 132 TM 032 322

AUTHOR	van der Linden, Wim J.
TITLE	Optimal Stratification of Item Pools in a-Stratified Computerized Adaptive Testing. Research Report.
INSTITUTION	Twente Univ., Enschede (Netherlands). Faculty of Educational Science and Technology.
REPORT NO	RR-00-07
PUB DATE	2000-00-00
NOTE	26p.
AVAILABLE FROM	Faculty of Educational Science and Technology, University of Twente, TO/OMD, P.O. Box 7500 AE Enschede, The Netherlands.
PUB TYPE	Reports - Research (143)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	*Adaptive Testing; *Computer Assisted Testing; *Item Banks; Linear Programming; *Test Construction; Test Items
IDENTIFIERS	*Stratification

ABSTRACT

A method based on 0-1 linear programming (LP) is presented to stratify an item pool optimally for use in "alpha"-stratified adaptive testing. Because the 0-1 LP model belongs to the subclass of models with a network-flow structure, efficient solutions are possible. The method is applied to a previous item pool from the computerized adaptive testing (CAT) version of the Graduate Record Examinations Quantitative Test. The results indicate that the new method performs well in practical situations. It improves item exposure control, reduces the mean squared error in the theta estimates, and increases test reliability. (Contains 2 figures and 25 references.) (Author/SLD)

Optimal Stratification of Item Pools in *a*-Stratified Computerized Adaptive Testing

Research Report

00-07

Wim J. van der Linden

faculty of
EDUCATIONAL SCIENCE
AND TECHNOLOGY

University of Twente

Department of
Educational Measurement and Data Analysis

Optimal Stratification of Item Pools in a-Stratified Computerized Adaptive Testing

Wim J. van der Linden

Abstract

A method based on 0-1 linear programming (LP) is presented to stratify an item pool optimally for use in *a*-stratified adaptive testing. Because the 0-1 LP model belongs to the subclass of models with a network-flow structure, efficient solutions are possible. The method is applied to a previous item pool from the CAT version of the GRE Quantitative Test. The results indicate that the new method performs well in practical situations. It improves item exposure control, reduces the MSE in the θ estimates, and increases test reliability.

Keywords: computerized adaptive testing; *a*-stratified adaptive testing; item pool stratification; 0-1 linear programming

Introduction

In computerized adaptive testing (CAT), items are sequentially selected according to the examinee's estimated proficiency (θ). In traditional CAT methods, items are selected to maximize Fisher's item information at the current θ estimate. As a result, some items will tend to become overexposed while others are seldom or never touched by the CAT algorithm. It is well known that the items that tend to get the highest exposure rates typically have the highest values for their discrimination parameter, whereas those with lower exposure rates still have acceptable values for this parameter from a measurement point of view.

Remedies to control for high exposure rates have been proposed by Davey & Parshall (1995), McBride & Martin (1983), Sympson and Hetter (1985), Stocking and Lewis (1995), Thomason (1995), van der Linden (1998a) and others. Among these methods, the most popular one is due to Simpson and Hetter (SH). The general idea of the SH method is to put a probabilistic "filter" between item selection and administration--that is, an item that is selected by the CAT algorithm may not be administered with a probability beyond a certain value. So, on the one hand the CAT algorithm still selects items to have maximum information, and hence high values for the discrimination parameter. On the other hand, upon selection, a probability experiment is run to decide whether the item is actually administered. Due to this probability experiment, the actual exposure rates of the popular items are reduced. The price to be paid, however, is that CAT with SH item exposure control tends to be less efficient than CAT based solely on Fisher's information criterion.

The *a*-stratified (ASTR) method of adaptive testing (Chang & Ying, 1999) takes a different approach. In this method, the item pool is divided into a number of strata, based on the values of the items for the discrimination parameter. During the test items are selected for administration based on these strata. Early in the test, items are administered from the stratum with the lowest value for the discrimination parameter. However, as the test progresses, strata with higher values are used. Within each stratum, the item with the value for the discrimination parameter closest to the examinee's current estimate of θ is selected for administration. As a consequence the *a*-stratification forces a more balanced exposure for all items. However, the price paid by the method will be low: Since estimation of θ tends to be quite inaccurate early in the test, it is more appropriate to use low-discriminating items at this point (Parshall, Hogarty, & Kromrey, 1999). Likewise, items with high discrimination can better be saved for used later in the test when the θ estimate has stabilized (Chang & Ying, 1996).

However, the ASTR method has been criticized for the following four aspects:

(1) There is a concern on how to stratify the item pool. In particular, the method chosen to stratify the item pool may interfere with a positive correlation between the item difficulty and discrimination parameters, and therefore the CAT procedure may have suboptimal operating characteristics (Stocking, 1998).

(2) The method does not guarantee that the exposure for every item will be below a specified rate (Stocking, 1998; Parshall, Hogerty & Kromrey, 1999; Lueng, Chang, & Hau, 1999).

(3) The method does not incorporate any device for handling constraints on test content (Stocking, 1998);

(4) There are no guidelines on the number of strata to use as well as the number of items to administer from each strata (Stocking, 1998).

While all four issues are important, the current paper will only address the first one; the other issues are addressed elsewhere (van der Linden & Chang, submitted).

It is important to note that in practice the item discrimination and difficulty parameters are often positively correlated (Lord & Wingersky, 1984). Figure 1 displays pairs of values for the discrimination (*a*) and difficulty parameter (*b*) of three hundred and sixty items from a GRE quantitative test. It clearly shows a positive correlation. A comparable plot for the Arithmetic Reasoning Test in the ASVAB is given in van der Linden, Scrams, and Schnipke (1999). However, in order to make the ASTR method to perform well, a crucial requirement is that the examinee's θ estimate be matched closely with the value of the item difficulty parameter, particularly at the later stages of the test. This requirement implies that the distribution of difficulty parameter not be influenced by the stratification on the discrimination parameter, or, equivalently, that *a* and *b* be uncorrelated. If the two parameters do correlate, items with certain values for the difficulty parameter may be missing and others selected more frequently. Indeed, as Parshall, Hogarty and Kromrey (1999) and Ban, Wang & Yi (1999) report, the exposure rates for some items can be very high when the ASTR method is used with operational item banks.

[Figure 1 about here]

To overcome this problem, Chang, Qian and Ying (1999) developed the method of *a*-stratified CAT design with *b*-blocking (BASTR) which balances the distributions of *b* values among all strata. The BASTR method first partitions the item bank according to *b* values and then implements the *a*-stratification. A simulation study showed that BASTR, which can be thought of as a hybrid of Chang & Ying's *a*-stratification and Weiss's *b*-stratification (1976) methods, performs better that the original ASTR. It improved item exposure rates control, reduced mean squared errors (MSE), and increased test reliability.

The present paper addresses the same issue of item pool stratification. Its basic idea is to use the technique of 0-1 linear programming (LP) to stratify optimally an item pool for application in *a*-stratified CAT. The stratification is optimal in the sense that an even distribution of the items is approximated both across the strata of the discrimination parameter and across the difficulty parameter within the strata. This twofold goal is realized by formulating a model for the optimization problem with a special objective function. At the same time, the model is given constraints to govern the numbers of items assigned to each combination of values for the discrimination and difficulty parameter. Because the model appears to have the simple structure of a network-flow problem (Armstrong, Jones & Wang, 1995; Nemhauser & Wolsey, 1988), fast algorithms to calculate an optimal item pool stratification are available. The technique of 0-1 LP has been applied earlier to assemble test forms from an item pool to meet various specifications with respect to, for example, the information function or the content of the test. For a review of such applications, refer to van der Linden (1998b).

Basic Idea and Notation

To formulate the model, target values are specified for the values of *a* for each stratum as well as for the distributions of *b* values within the strata. In the empirical application below, these target values are chosen to be evenly distributed over *a* and *b*, but other choices are possible. The system of target values serves as the design for the stratified item pool. The items in the pool are then assigned to these target values such that their actual parameter values approximate the design as closely as possible.

To formalize the idea the following notation will be used:

Item index	: i=1,...,I;
Index for strata of a	: s=1,...,S;
Index for grid of b values	: k_s=1,...,K_s;
Target value for a_i parameter	: a_s^*;
Target value for b_i parameter	: b_{ks}^*;
Number of items assigned to target (k,s):	: n_{ks}.

Observe that, for the sake of generality, the target values for the item difficulty parameter have been chosen to be stratum dependent. However, a common set of target values, b_k^*, k=1,...,K, across all strata will be appropriate in most applications.

In addition, decision variables x_{iks} are needed that take the value one if item i is assigned to target value b_k^* in stratum s, and the value zero otherwise. The objective function and constraints in the optimization model are formulated using these variables.

Model

Because large numbers of decision variables will be involved in applications to CAT with item pools and sets of target values of realistic sizes, the current problem is modeled as a problem belonging to a subset of 0-1 linear programming (LP) problems known as (semi-assignment) network-flow problems (Nemhauser & Wolsey, 1988). These problems have a special structure allowing for the use of a simplified version of the simplex algorithm able to deal with thousands of variables in seconds. The software package CPLEX 6.5 (ILOG, 1999), used in the empirical example below, has a very efficient optimizer for network-flow problems.

In a network-flow interpretation of our problem, we "ship" or assign items from supply nodes to demand nodes. The supply nodes are the individual items i=1,...,I. The demand nodes are the combinations of target values for the item parameters, (b_k^*,a_s^*) .The total number of items assignment to each demand node (b_k^*,a_s^*) is n_{ks} items. The assignment is such that the "distance" between the items and their target values if minimized.

An attractive measure for the distance between the parameter values of item i and the combination of target values (b_k^*,a_s^*) is the following Euclidean measure:

$$\delta_{iks} = \sqrt{(b_i - b_k^*)^2 + \lambda^2(a_i - a_s^*)^2}\ , \qquad \textbf{(1)}$$

where the weight

$$\lambda = (b_{\max} - b_{\min})/(a_{\max} - a_{\min})$$

is introduced to remove the scale differences between b_i and a_i in the item pool. However, the weights could also be defined to express our belief that one of the two item parameters is more important than the other.

The model is as follows:

$$\text{minimize} \sum_{i=1}^{I}\sum_{k=1}^{K}\sum_{s=1}^{S} \delta_{ijks}x_{iks} \qquad \textbf{(2)}$$

subject to

$$\sum_{k=1}^{K}\sum_{s=1}^{S} x_{iks} = 1,\quad i=1,\ldots,I; \qquad \textbf{(3)}$$

$$\sum_{i=1}^{I} x_{iks} = n_{ks}, \quad k=1,\ldots, K, \quad s=1,\ldots,S; \tag{4}$$

$$0 \leq x_{iks} \leq 1, \quad i=1,\ldots,I, \quad k=1,\ldots,K, \quad s=1,\ldots,S. \tag{5}$$

Observe that the objective function in (2) minimizes the sum of the distances between the actual and target values for the item parameters. The constraints in (3) guarantee that each item is assigned to exactly one combination of target values. Likewise, the constraints in (4) require that each combination of target values receive n_{ks} items.

Empirical Example

The model was applied to stratify an item pool for the CAT version of the Graduate Record Exams (GRE) consisting of 360 quantitative items calibrated according to the 3PL model (Birnbaum, 1968). Figure 1 contains a scatter plot of the *a* values vs. *b* values for the 365 items in the GRE pool. Clearly, the values are positively correlated (r=0.44).

Item Pool Stratification.

Two methods were used in the item bank stratification: the original method (ASTR) and the 0-1 LP method proposed in this paper (0-1 ASTR).

For ASTR, the item bank was partitioned into four equally large strata in ascending order of the *a* values. The first stratum consisted of items with the smallest *a* values, the next consists of items with the next smallest *a* values, etc.

For 0-1 ASTR the following steps were made:

1. The set of target values (b_k^*, a_s^*) was chosen to have four evenly distributed values for a_i (0.55; 0.74; 0.95; and 1.28) were chosen. Because the values of b_i were scaled to have a mean of zero, the values -2.0, -1.0, .0, 1.0, and 2.0 were chosen as targets for the within-stratum distributions of this parameter.

2. For each combination of target values, the number of items was set equal to n_{ks}=18.

3. The values for the distance measure δ_{iks} in (1) were calculated for each item in the pool.

4. The input file for the software package CPLEX 6.5 (ILOG, 1999) was prepared and an optimal solution for the model in (2)-(4), with the distance measure in (1), was calculated. The CPU time needed to solve the model was only 1.66 seconds on a 266 MHz Pentium II processor (64 KB RAM).

Table 1 gives some summary statistics for the two stratification methods. In both cases, the means of the *a*-values for the four levels are naturally ordered. For ASTR, the means of b values vary noticeably across strata. An important goal for 0-1 ASTR was to make the distributions of the *b* values more identical across all strata. A useful indicator of the realization of this goal is the extent to which the means and standard deviations of the *b* values per stratum are similar to the overall mean and standard deviation given by the first column of Table 1. The other columns show that 0-1 ASTR outperforms ASTR in realizing this goal. The only less satisfactory results are those for Stratum 4, where 0-1 ASTR performs better than ASTR but, due to the substantial correlation between the *a* and *b* values in the pool, even the optimal solution is less satisfying.

Table 1. Item Bank Statistics

	Total	Stratum 1		Stratum 2		Stratum 3		Stratum 4	
Method		*ASTR*	*0-1 ASTR*	*ASTR*	*0-1 ASTR*	*ASTR*	*0-1 ASTR*	*ASTR*	*0-1 ASTR*
# Items	360	90	90	90	90	90	90	90	90
Mean *a*	0.87	0.52	0.53	0.74	0.75	0.95	0.93	1.28	1.26
SD *a*	0.31	0.09	0.11	0.06	0.10	0.06	0.13	0.20	0.21
Min *a*	0.26	0.26	0.26	0.64	0.54	0.85	0.63	1.07	0.98
Max *a*	2.00	0.64	0.73	0.84	0.99	1.01	1.23	2.00	2.00
Mean *b*	0.14	-0.39	-0.10	-0.08	-0.03	0.30	0.13	0.74	0.54
SD *b*	0.99	1.19	1.18	1.02	1.03	0.65	0.92	0.57	0.60
Min *b*	-2.89	-2.89	-2.89	-2.47	-2.89	-2.19	-2.47	-0.73	-0.92
Max b	2.21	2.00	2.02	1.79	1.79	1.70	1.76	2.21	2.21

Simulation Study

A simulation study was used to compare the performance between the two pools in an application of the *a*-stratified CAT method of item selection in terms of efficiency of θ estimation, effectiveness in item pool usage, and maximum item exposure rates. The design and evaluation criteria in the simulation study were similar to those of Chang and Ying (1999).

A fixed test length of 40 items was used throughout the simulation study. Three thousand θ values were generated from the standard normal distribution, N(0,1). The method of maximum likelihood estimation (MLE) was used to estimate θ. Ten consecutive items were selected from each *a* stratum. The first three items were selected from the first stratum as described in Chang & Ying (1999). Actually, at each step two items with the closest b values were selected first, and then one of them was chosen randomly with probability of .50.

Evaluation Criteria

Reliability. The correlation coefficient between the true and estimated θ values was calculated. This reliability can be interpreted as the correlation ratio associated the observed and true scores on the test (Lord, 1980, p. 52).

Bias and mean squared error. These quantities were calculated as:

$$\text{Bias} = \frac{1}{3000}\sum_{i=1}^{3000}(\hat{\theta}_i - \theta_i) \qquad (6)$$

$$\text{MSE} = \frac{1}{3000}\sum_{i=1}^{3000}(\hat{\theta}_i - \theta_i)^2 \qquad (7)$$

Number of Under-utilized Items: From a practical point of view, items with very low exposure rate are useless. In this study, an item was considered as under-utilized if its exposure rate was below 5%.

Scaled Chi-squared Statistic: Mimicking Pearson's χ^2 statistic, Chang & Ying (1999) proposed a statistic to measure the skewness of the exposure rate distribution. It was defined as:

$$\chi^2 = \sum_{j=1}^{N}\frac{(r_j - L/N)^2}{L/N}, \qquad (7)$$

where r_j is the observed exposure rate for the jth item L is the test length, and N is the item pool size.

Test-overlap Rate: Test overlap rate, which is the expected number of common items encountered by two randomly selected examinees divided by L, was also measured.

Results

Table 2 summarizes the results from the simulation. For both pools the correlation coefficients between θ and $\hat{\theta}$ had comparable values (around 0.96). However, 0-1 ASTR outperformed ASTR in terms of reducing both bias and MSE. 0-1 ASTR also made more efficient use of the item bank. Among 360 items, only 1 items had exposure rate below 5% when 0-1 ASTR was used, whereas there were 48 such items for ASTR. The χ^2 measure for 0-1 ASTR was smaller than that of ASTR: the F ratios $F_{01ASTR,ASTR}$ $= \chi_{01ASTR}{}^2 \ / \ \chi_{ASTR}{}^2$ = 0.56 (see Chang & Ying, 1999, for a detailed discussion of the *F* ratio index). Thus, there was about 44% reduction of skewness in 0-1 ASTR relative to ASTR. The test-overlap rates were 17.4% and 14.6% for ASTR and 0-1 ASTR, respectively.

Figure 2 exhibits a relationship between the item exposure rates and item parameter values for the pools stratified according to the two methods. For ASTR, there were several dozens items that had unacceptably high exposure rates. These items all came from Stratum 3 and 4 (high *a* values) and had low *b* values. As demonstrated by the positive correlation between the *a* and *b* values, the overexposure of these items was due to the lack of items with low *b* values within these strata. In fact, Table 1 shows that, though the mean *b* value for the item bank is 0.14, it becomes 0.30 for Stratum 3 and .74 for Stratum 4. However, in spite of the positive correlation between the *a* and *b* values in the GRE item pool, this overexposure in disappeared for Stratum 3 and was reduced for

Stratum 4 for the item pool stratification method based on the above network-flow model (0-1 ASTR).

Table 2. MSE, bias and other performance measures for the two methods.

Methods	**ASTR**	**0-1 ASTR**
MSE	0.081	0.073
BIAS	0.012	0.002
Overlap Rate	17.4%	14.6%
χ^2	22.769	12.646
Exposure rate<5%	48	1
$\rho_{\vartheta,\hat{\theta}}$	0.962	0.965

Discussion

In this paper, we have introduced a network flow model for application in item pool stratification in the ASTR method of item selection in CAT. The ASTR method was proposed originally to avoid high *a* items to be overly exposed and make more even and efficient use of all items in an item bank. ASTR performs well for ideal item banks where the *a* and *b* parameters are not correlated, but it could lead to problems when *a* and *b* are correlated. The new item pool stratification method (0-1 ASTR) provides an optimal solution for this case. It can be thought of as a preemptive measure to force balanced distributions of *b* values across strata. As a result, some of the strata formed by the method covers a wider range of *b* values than that for the original ASTR. Simulation results showed that the new method performs well in practical situations. It improved item exposure control, reduced the MSE in the θ estimates, and increased test reliability. The use 0-1 LP in *a*-stratified adaptive testing can be generalized to deal with many other practical issues in CAT designs, for example, balance test content (van der Linden & Chang, submitted).

An issue in the application of 0-1 LP model to optimal item pool stratification not yet touched is how to select the target values. The values in the empirical application in this paper were selected intuitively; other selections would have produced other results. A more formal criterion to choose target values would be welcome, in particular if the criterion would allow us to predict what would happen for empirical item pools with various distributions of item parameter values.

References

Armstrong, R.D., Jones, D.H., & Wang, Z. (1995). Network optimization in constrained standardized test construction. In K. D. Lawrence (Ed.), *Applications of management science: Network optimization applications* (Vol. 8) (pp. 189-212). Greenwich, CT: JAI Press.

Ban, J., Wang, T., & Yi, Q. (1999,June). *Comparison of the a-stratified method, the Sympson-Hetter method, and the matched difficulty method in CAT administration.* Paper presented at the Annual Meeting of the Psychometric Society, Lawrence, KS.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick, *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.

Chang, H., & Ying, Z. (1996). Global information approach to computerized adaptive testing. *Applied Psychological Measurement, 20*, 213-229.

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.*

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

ILOG, Inc. (1999). *CPLEX 6.5* [Computer progam and manual]. Incline Village, NV: Author.

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

Lueang, C.-K., Chang, H.-H., & Hay, K.-T. (1999, April). *An enhanced stratified computer adaptive testing design*. Paper presented at the Annual Meeting of the American Educational, Research Association, Montreal, Canada.

McBride, J. R., & Martin, J. T. (1983). Reliability and validity of adaptive ability tests in a military setting. In D.J. Weiss (Ed.), *New horizons in testing* (pp. 223-226). New York, Academic Press.

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

Parshall, C., Hogarty, K., and 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.

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.

Thomasson, G. L. (1995, June). *New item exposure control algorithms for* *computerized adaptive testing*. Paper presented at the Annual Meeting of Psychometric Society, Minneapolis, MN.

van der Linden, W. J. (1998a). Bayesian item selection criteria for adaptive testing. *Psychometrika, 63*, 201-216.

van der Linden, W.J. (Ed.) (1998b). Optimal test assembly os 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., & Chang, H.-H. (submitted). Alpha-stratified adaptive testing with large numbers of content constraints.

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

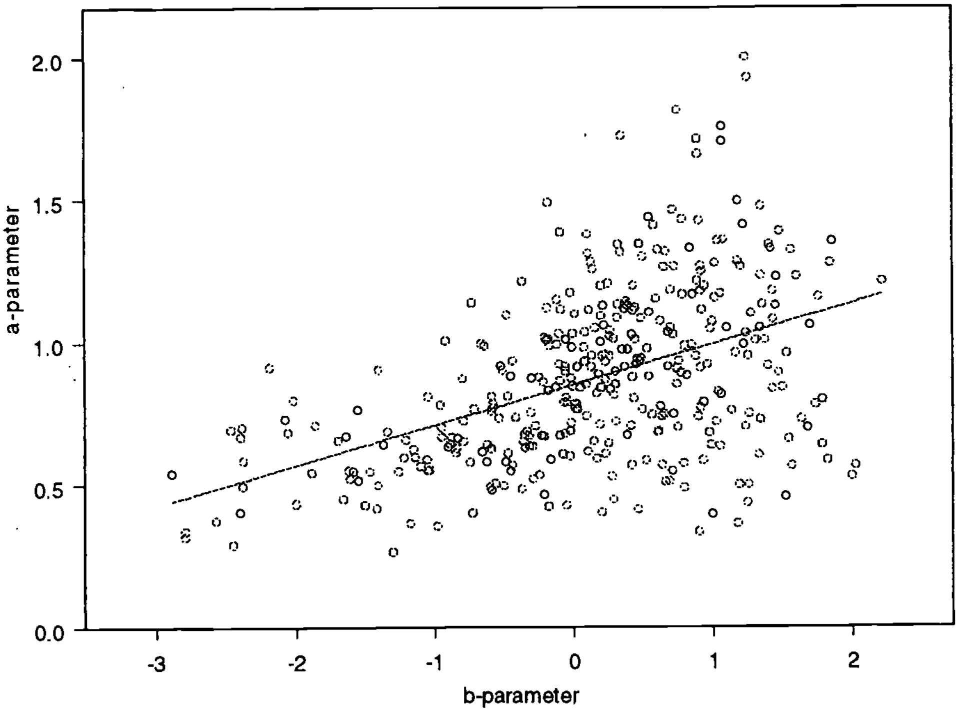
Wainer, H. (1990). *Computerized adaptive testing: A primer*. Hillsdale, NJ: Lawrence Erlbaum Association.

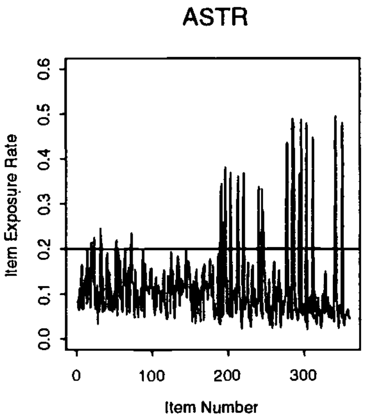
Weiss, D. J. (1976). Adaptive testing research in Minnesota: Overview, recent results, and future directions. In C. L. Clark (Ed.), *Proceedings of the first conference on computerized adaptive testing* (pp. 24-35). Washington, DC: United States Civil Service Commission.

Figure Captions

Figure 1. Relationship between thevalues of the *a* and *b* paameters of the GRE Quantitative item bank.

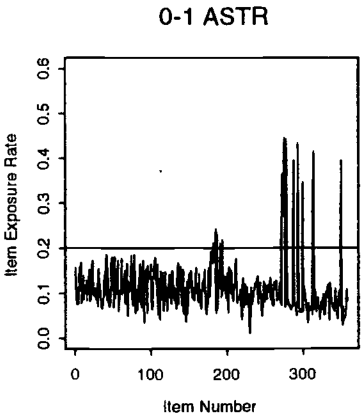
Figure 2. Item exposure rates for the to methods of item pool stratification.


2.0
1.5
1.0
0.5
0.0
a-parameter
-3
-2
-1
0
1
2
b-parameter



ASTR

Item Exposure Rate
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0
100
200
300
Item Number



0-1 ASTR

Item Exposure Rate
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0
100
200
300
Item Number

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

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*

RR-98-01 C.A.W. Glas, R.R. Meijer, E.M.L.A. van Krimpen-Stoop, *Statistical Tests for Person Misfit in Computerized Adaptive Testing*

RR-97-07 H.J. Vos, *A Minimax Sequential Procedure in the Context of Computerized Adaptive Mastery Testing*

RR-97-06 H.J. Vos, *Applications of Bayesian Decision Theory to Sequential Mastery Testing*

RR-97-05 W.J. van der Linden & Richard M. Luecht, *Observed-Score Equating as a Test Assembly Problem*

RR-97-04 W.J. van der Linden & J.J. Adema, *Simultaneous Assembly of Multiple Test Forms*

. . .

<u>Research Reports</u> 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