Integrating Stratification and Information Approaches for Multiple Constrained CAT

Chi-Keung LEUNG

The Hong Kong Institute of Education

Hua-Hua CHANG

National Board of Medical Examiners

Kit-Tai HAU

The Chinese University of Hong Kong

Paper for presentation at the NCME Annual Meeting, April 11-13, 2001

Integrating Stratification and Information Approaches for Multiple Constrained CAT

Chi-Keung LEUNG
The Hong Kong Institute of Education

Hua-Hua CHANG
National Board of Medical Examiners

Kit-Tai HAU
The Chinese University of Hong Kong

Abstract

It is widely believed that item selection methods using maximum information approach (MI) maintain high efficiency in trait estimation by repeatedly choosing high discriminating items. As a result, these methods can yield extremely skewed item exposure distribution in which items with high *a* values may be over-exposed while those with low *a* values may never be selected. The *a*-stratified design (ASTR) and its extension, the *a*-stratified with *b*-blocking method (BASTR), were proposed in an attempt to simultaneously control the exposure of high *a* items and improve the utilization of low *a* items. The latter was developed to take into consideration the correlation between the *a*- and *b*-parameters. Using low *a* items first is a new philosophy of item selection advocated in both stratification designs. The main objective of the study was to investigate how the integration of the traditional information based selection and the new philosophy of using low *a* items first would perform in a multiple constrained setting. Specifically, the performances of MI, BASTR, and their integration, MIBASTR, were examined and compared. Results indicate that BASTR was the best in utilizing the entire pool and tackling item security problems. On the other hand, MI and MIBASTR offered high and comparable measurement efficiency. But the latter outperformed the former in item exposure control and pool utilization.

Introduction

In the past decade, many large-scaled testing programs have been partly or completely converted into the form of computerized adaptive testing (CAT) that was first developed under the framework of item response theory (IRT; Lord, 1970). In CAT, examinees are presented with tailor-made tests. One item is selected at a time on the basis of the currently available estimate of the examinee's ability (Lord, 1980; Weiss, 1982). One of the main advantages of CAT over P&P is that it enables more efficient and precise trait estimation (Owen, 1975; Wainer, 1990; Weiss, 1982). Others include a larger flexibility in test scheduling and the incorporation of alternate item forms (Straetmans & Eggen, 1998). A key issue in CAT is how to adaptively select the best test items from the item pool. The traditional item selection algorithms rely on local item information. This means that an item is selected if it has the maximum Fisher information at the current ability estimate based on the responses to previously administered items. It has been noted that this information criterion would cause skewed item exposure (Davey & Parshall, 1995; McBride & Martin, 1983; Stocking & Lewis, 1995; Sympson & Hetter, 1985; Thomasson, 1995; van der Linden, 1998). In particular, items with large value of discrimination parameter may be overly exposed while some others are never used. Over-exposure would eventually cause threat in item security while under-utilization would reduce the cost effectiveness of developing and maintaining those inactive items in the pool.

It is understandable, therefore, the control of item exposure and the enhancement of pool efficiency are important issues in CAT designs (Mills & Stocking, 1996; Stocking & Swanson, 1998; Way, 1998). Methods that simultaneously control maximum item exposure rate to improve item security and uplift exposure of under-utilized items to enhance item pool efficiency have been proposed by Chang and Ying (1999) and Stocking and Swanson (1998), among others. In contrast to the traditional approach of looking for the most informative items at every stage of item selection, Chang and Ying (1999) proposed the multi-stage a-stratified design (ASTR) that partitions items into several strata in an ascending order of the item discrimination parameter. Each test then consists of matching numbers of stages and strata, with items of the first stage being selected from the first stratum, and so on. One major rationale for such a design is that at early stages, the gain in information by using the most informative item may not be realized because the ability estimation is still relatively inaccurate. Thus items with high discrimination values should be saved for later selection stages to pin-point the ability. The ASTR has been shown through simulation studies to be effective in both reducing item-overlap rate and the enhancing pool utilization, when it is used with certain type of item pools. The findings of Hau and Chang (in press) also support that the ASTR offers comparable efficiency as MI

but has the potential advantages of achieving a more balanced item usage and stable resultant pool structure after item replenishments.

Nevertheless, the correlation of *a*- and *b*- parameters of the items in some pools may be significant. In such cases, there may not be sufficient items with low *b*s in the last stratum. Consequently, the ASTR would result in quite a number of items being over-exposed. To tackle such problem, Chang, Qian and Ying (in press) developed the *a*-stratified with *b*-blocking method (BASTR) by refining ASTR. In BASTR, item pool is first divided into many small levels based on *b* parameters. Within each level, items are grouped in ascending order of *a* values. Then the first group of items with the smallest *a* values from each level are merged to form the first stratum, and the second group of items with the second smallest *a* values merged into the second stratum,…, and eventually the last stratum contains those items with the largest *a* values from each level. As a result, the *b* distribution remains steady but the average *a* value increases across stratum.

In many situations, however, CAT design has to take into consideration additional constraints such as content balancing, item type and statistical specifications. Various models have been proposed to solve such complex and sometimes conflicting requirements in test assembly (Armstrong, Jones & Kunce, 1998; Luecht, 1998; Stocking & Swanson, 1993, 1998; van der Linden & Reese, 1998). One strategy, as advocated by Stocking and Swanson (1993) in the Weighted Deviation Model (WDM) is to relax test constraints as desired properties. It can deal with many kinds of constraints and provide acceptable solution to those systems that may not have an optimal combination satisfying all criteria. Leung, Chang, and Hau (2000) have shown that the two integrated item selection methods, BASTR-WDM and ASTR-WDM, can tackle the problem of multiple constraints to a certain extent. However, they have also pointed out that more thoughts on pool partition and item selection are needed for better face validity and measurement efficiency. One major problem in BASTR-WDM and ASTR-WDM is that it may not be easy to find suitable items in the last stratum to satisfy the unfulfilled constraints during the last stage of an adaptive test.

The paper proposes to combine the strengths of stratification design and information approach for multiple constrained CAT. Specifically, an item pool is first partitioned into two strata based on the pool characteristics using the general concept of BASTR. As a result, both strata cover a wide range of *b*s with the first stratum having a smaller average of *a*s. The test is then divided into two stages. As the trait estimation is generally inaccurate in the early stage of a test, the items of the first stage are selected from the first stratum by matching the *b*-parameter and the current ability estimate. And in the final stage, items are selected from the second stratum based on the maximum information approach to pin-point

the true theta. It is suggested that the size of the first stratum and the number of items for the first stage of a test are much smaller than the corresponding quantities in the second stratum and the final stage. The main reasons of doing this are: to uplift the utilization of low *a* items, to increase the availability of suitable items for unfulfilled constraints during the final stage, and to increase the precision in trait estimation. In the study, the performance of the new integrated method, MIBASTR, was examined and compared with the other two methods, BASTR and MI.

Simulation Studies

A series of simulation studies were conducted to investigate the performance of MI, BASTR, and MIBASTR, when multiple constraints were imposed. There were 19 non-statistical constraints arising from two intrinsic features: content area and cognitive level. The sample consisted of 5,000 examinees with abilities (referred as true ability θ hereafter) randomly generated from N(0,1). Each examinee received three simulated tests administered respectively by the three item selection methods incorporated with WDM: MI, BASTR, and MIBASTR.

Item pool

The item pool contained 700 upper primary mathematics items from four content areas crossed with three cognitive levels. Table 1 shows the distribution of the items. The mean estimated discrimination (*a*) value for the items was 1.02, with SD = .33, and a range of .29 to 2.63. The mean estimated difficulty (*b*) was .16, with SD = 1.06, and a range of –3.44 to 3.40; thus there was a close match between the *b* distribution and ability distribution. The mean estimated pseudo-guessing parameter (*c*) was .17, with SD = .008, and a range of .03 to .50. The correlation of *a*- and *b*-parameters was .45.

Test constraints

The test length was fixed at 42 items. Table 2 presents the desired ranges and relative weights for various combinations of content area and cognitive level.

Item selection methods

The performances of the following three item selection methods were compared when WDM was incorporated to deal with multiple constraints. In the WDM, test constraints are treated as desired properties. Thus, when there is no feasible solution to the system of the constraints, the objective function could find an item yielding the minimal deviation from the constraints. The model used in the study was:

Minimize

$$\sum w_j d_{L_j} + \sum w_j d_{u_j} + w_\theta d_\theta \; w_j \qquad \text{(sum of weighted deviations)} \qquad (1)$$

Subject to

$$\sum_{i=1}^{N} a_{ij} x_i + d_{L_j} - e_{L_j} = L_j \quad , j = 1, \ldots, 19 \qquad \text{(for lower bounds of constraints)} \quad (2)$$

$$\sum_{i=1}^{N} a_{ij} x_i + d_{U_j} - e_{U_j} = U_j \quad , j = 1, \ldots, 19 \qquad \text{(for upper bounds of constraints)} \quad (3)$$

$$\sum_{i=1}^{N} I_i(\theta) x_i + d_\theta - e_\theta = \infty \qquad \text{(for max. information criterion)} \quad (4)$$

$$d_{L_j}, d_{U_j}, e_{L_j}, e_{U_j} \geq 0 \quad , j = 1, \ldots, 19 \qquad (5)$$

and

$$x_i \in \{0, 1\}, i = 1, \ldots, N, \qquad \text{(Stocking \& Swanson, 1993, p. 280-281)} \qquad (6)$$

where a_{ij} equaled to 1 if item *i* had property *j* and 0 otherwise; w_j was the weight assigned to constraint *j*; w_θ was the weight assigned to the information constraint; L_j and U_j were the lower bound and upper bound of constraint *j*; d_{L_j} and d_{U_j} represented deficit from the lower bound and surplus from upper bound respectively; e_{L_j} and e_{U_j} represented excess from lower bound and deficit from upper bound respectively; x_i equaled to 1 if *i*th item was included in the test, or 0 otherwise.

The heuristic for item selection using WDM was also adopted:

1. For every item not already in the test, compute the deviation for each of the constraints if the item were added to the test;
2. Sum the weighted deviations across all constraints;
3. Select the item with the smallest weighted sum of deviations (p. 281).

Method 1 (MI):

1. The content area and cognitive level of each item were indexed.

2. The optimal function of Equation 1 was used.
3. Test constraints and maximum information were treated as desired properties as stated in Equations 2 to 4.
4. The bounds and weights for the 19 non-statistical constraints were set as stated in Table 2. The lower bound for information was set at 100.0 which was practically unreachable in the study. The weight for information was assigned the value of 15.0 that was used by Stocking and Swanson (1993) and was found to have yielded very small mean squared error here.
5. Items were selected by following the three steps of the heuristic mentioned earlier.

Method 2 (BASTR):

1. The content area and cognitive level of each item were indexed.
2. Test constraints were mathematically formulated as Equations 2 and 3.
3. The item pool was partitioned into 4 strata as follows. First, the pool was divided into 20 different levels based on the *b*-parameters. Then the items within each level were sorted in ascending order of *a*-parameters and then divided into 4 groups. All the first groups of the 20 levels were merged to form the first stratum;…; and all the last groups merged to form the last stratum. As a result, each stratum then had similar distribution for *b*-parameters but the average value of *a*-parameters increased across the strata. Table 3 shows the parameter distributions across strata after stratification.
4. Each test was divided into 4 stages. In the first stage, 10 items were administered from the first stratum. In the second stage, next 10 items were administered from the second stratum. And then 11 items from each of the last two strata.
5. At each point of item selection, a hundred unadministered items with difficulty parameter closest to the currently estimated theta were chosen. For each of them, the absolute difference between the *b* value and the current ability estimate was computed. The weight for such absolute difference was set at 5.0. Then the total weighted deviations were summed up as if the item had been added to the test. The item with the smallest weighted sum of deviations was administered.

Method 3 (MIBASTR):

1. The constraints and information criterion were formulated as Equations 2 to 4.
2. The item pool was partitioned into two levels. The first level corresponded to the first stratum while the second level was formed by combining the last three strata of BASTR.

3. Accordingly, each test was divided into 2 stages: the first 10 items from the first stage and the rest 32 items from the final stage.
4. The item selection rule in the first stage was the same as described in BASTR.
5. In the final stage, all unadministered items of the second stratum were considered. For each of them, the deviation for each of the desired properties, including information, was computed as if the item were added to the test. The weighted deviations across all constraints were summed and then the item with smallest weighted sum of deviations would be administered.

Measures of Performance

Reliability: Irrespective of the item selection design, CAT should always provide reasonable reliability. Otherwise, test results cannot be used for inference or decision. Therefore, reliability was an evaluation criterion for the performance of the three selection methods. In this study, 5,000 true abilities were generated and their respective estimates were obtained according to the selection methods employed. Thus reliability here was interpreted as the correlation between the estimated scores and the true scores (Lord, 1980, p. 52). The higher the reliability, the better the item selection method would be.

Bias: Accuracy is another important criterion, which in this study was measured by the estimated bias and mean squared error. Let θ_i, i = 1,…, 5000 be the true abilities of the 5000 examinees and $\hat{\theta}_i$ be the respective estimators from the CAT. Then the estimated bias was computed as

$$\text{Bias} = \frac{1}{5000}\sum_{i=1}^{5000}(\hat{\theta}_i - \theta_i) \qquad (7)$$

The smaller the bias, the better the item selection method would be.

Mean squared error: Using the same notations of true ability and the estimator, the estimated mean squared error was computed as

$$\text{MSE} = \frac{1}{5000}\sum_{i=1}^{5000}(\hat{\theta}_i - \theta_i)^2 \qquad (8)$$

The smaller the MSE, the better the item selection method would be.

Number of over-exposed items: The exposure rate of an item is defined as the ratio of the number of times the item is administered to examinees over the total number of examinees taking the test. If an item has a high exposure rate, then it has a greater risk of being known to prospective examinees, which in turn would cause item security and test validity problems. Since one of the main concerns of this paper was to compare the three item selection methods in item exposure control, the number of overly exposed items was certainly one of the key evaluation criteria. Here an item was considered as overly exposed if its exposure rate was greater than .2, a commonly used cut-off value. The smaller the number of over-exposed items, the better the item selection method would be.

Number of under-utilized items: Items with very low exposure rate implies that they are rarely used. If there are too many items with low exposure rates, then the item pool is not well utilized, which challenges directly the cost effectiveness of the item pool and the appropriateness of the item selection method. In this study, an item was considered as under-utilized if its exposure rate was below .02. The smaller the number of under-utilized items, the better the item selection method would be.

Scaled chi-squared statistic: Chang and Ying (1999) proposed that a uniform exposure rate distribution should be the most desirable in order to have a maximum item pool utilization. If the pool size is N and test length is L, then the optimum uniform exposure rate is L/N. They introduced a scaled chi-square to measure the overall item pool usage efficiency:

$$\chi^2 = \sum_{j=1}^{N} \frac{(er_j - L/N)^2}{L/N} \qquad (9)$$

where er_j represents the observed exposure rate for the jth item.

Equation 9 reflects the discrepancy between the observed and the ideal exposure rates. The smaller the χ^2, the better the pool utilization and hence the item selection method would be.

Item overlap rate: The item overlap rate (sometimes called the test overlap rate) is another important summary index in measuring item exposure control (Mills & Stocking, 1996; Way, 1998). Item overlap rate is indicated by the proportion of items shared by pairs of examinees, averaged across all possible pair-wise combinations. Way (1998) argued that such an index, not being calculated on an individual item basis, provides a global picture of

how often sets of items are administered. The higher the item overlap rate, the bigger the damage to test validity due to information sharing among examinees who take the test at different times. He also stressed that this index is critical in determining the size and composition of item pools that are needed for a particular CAT. If the test length is L and there are P examinees, the item overlap rate here was computed by: (i) first counting the number of common items for each of the $P(P\text{-}1)/2$ pairs of examinees, (ii) adding up all the counts in the $P(P\text{-}1)/2$ pairs, and (iii) dividing the total count by $LP(P\text{-}1)/2$. The smaller the overlap rate, the better the item selection method would be. Chang and Zhang (1999) and Chen, Ankenmann and Spray (1999) separately found that the lower bound for the expected item-overlap rate is L/N, the ratio of the test length to the pool size. This lower bound serves as a baseline for comparison among item selection methods in controlling item overlap.

Deviations from desired ranges: As the desired ranges are usually recommended by the experts of the relevant fields, it is anticipated that the CATs by all selection methods should have fallen within these ranges as far as possible. The less the deviations from the desired ranges of the non-statistical constraints, the higher the face validity and thus the better the item selection method would be.

Results

The results of the study are summarized in Table 4. The three item selection methods appeared virtually unbiased as their estimated biases all close zero. They offered high and comparable reliabilities. BASTR outperformed the other two in tackling the issues of item security and pool utilization. It greatly reduced the risk of item leakage by substantially lowering item overlap rate and the number of over-exposed items. In addition, it made well use of the item pool by yielding much smaller values in both the chi-square statistic and the number of under-utilized items.

But in terms of measurement efficiency, BASTR alone seemed inadequate. Its MSE was larger than that of the other two methods. Both MI and MIBASTR were more efficient and they yielded comparable MSE. Regarding item security, MIBASTR outperformed MI as it overly exposed less number of items and yielded a smaller overlap rate. Besides, the observed maximum exposure rate was also smaller. The method was also superior in addressing the issue of pool utilization by yielding smaller values of chi-square and the number of under-utilized items.

All the three methods conformed well to the multiple constraints as there was at most one observed deviation for all tests. BASTR had the smallest number of adaptive tests with deviation whilst MI had the largest.

Figure 1 shows the item exposure distributions for the three methods. The curve for BASTR indicates that the method utilized most items with a maximum exposure slightly over .3. In contrast, MI left about 60% of the pool items untouched, thus raising the issue of the cost effectiveness of developing and maintaining the pool. MIBASTR had similar problem but in a better position as it left about 45% of items never used. Besides, the item security problem in MIBASTR was less serious than in MI as MIBASTR over-exposed less number of items with a lower maximum exposure rate.

Figure 2 presents the individual exposure rates of items that were sorted in ascending order of *a*-parameter within each stratum. It is clear that MI repeatedly selected high discriminating items that were near the end of each stratum, leaving a lot of low *a* items around the front of stratum under-utilized. BASTR yielded a much even exposure distribution and utilized the pool more effectively. Figure 2(c) reflects that MIBASTR attempted to equalize the item exposures during the first stage that used items with smaller average *a* value and the *b*-matching selection criterion. But in the second stage that involved the items from all the last three strata and used information-based criterion, MIBASTR yielded similar exposure pattern as MI did. It appears that imposing the stratified design could lessen the problems of over-exposure and poor pool utilization associated with MI.

Discussion

There are two common problems associated with information-based item selection methods, namely poor pool utilization and item security threat. Both ASTR and BASTR were developed in an attempt to remedy these problems; the latter has been shown to be more appropriate for item pools in which *a*- and *b*- parameters are strongly correlated. These two stratified methods can work with WDM to meet multiple constraints to a larger extent (Leung, Chang & Hau, 2000). However, they on their own may not be as efficient as MI does. The main objective of the study was to investigate how the integration of MI and BASTR would perform in a multiple constrained setting.

Results indicate that the new philosophy advocated in the stratification designs, using low *a* items first, can be combined with the traditional wisdom of information-based criterion. It has been shown that both MI and MIBASTR offered comparable measurement efficiency and accuracy. But the latter outperformed the former in dealing with the practical issues of pool utilization and item security. MIBASTR demonstrated two important strengths of stratification. First is in enhancing pool utilization by reducing the number of under-utilized items and the value of scaled chi-square statistic. Second is in improving item security by lowering the maximum exposure rate, the number of over-exposed items, and the item overlap rate. The integrated method also inherited the efficiency of information approach

by yielding similar MSE and reliability as MI did.

The current findings have two implications for CAT design and pool management. First, there is no conflict between the new philosophy of stratification and the traditional wisdom of using information criterion. The two concepts can be combined to form an effective method for addressing many practical issues such as pool utilization, item security and measurement efficiency. Second, using low *a* items and *b*-matching criterion at the beginning of a test would not affect the measurement efficiency due to a considerable discrepancy between the true theta and its estimate during the initial stages. In fact, this strategy tends to equalize the exposure rates of those items of the first stratum. Thus, if there are some inactive items in a testing program, their usage can be uplifted by streaming them into the first stratum. This would improve the cost effectiveness of the development and maintenance of these items, without sacrificing the measurement efficiency.

The integration of MI and BASTR has been shown to be more promising than either one alone. It should be noted that though the over-exposure problem of MI has been lessened in MIBASTR, the number of over-exposed items is still high. Thus, future research may investigate how the integration of stratification design and information approach would perform when some measure of exposure control is imposed. Other factors that may affect the performance include the size and the number of strata, the number of constraints and the corresponding ratios, the matching of *b* distribution and the ability population, and the test termination rule.

References

Armstrong, R.D., Jones, D.H., & Kunce, C.S. (1998). IRT test assembly using network-flow programming. *Applied Psychological Measurement*, *22*, 237-247.

Chang, H.H., Qian, J., & Ying, Z. (in press). a-Stratified Multisage CAT with b-Blocking. *Applied Psychological Measurement*.

Chang, H.H., & Ying, Z. (1996). A Global Information Approach to Computerized Adaptive Testing. *Applied Psychological Measurement, 20*, 213-229.

Chang, H.H., & Ying, Z. (1999). A-stratified Multistage Computerized Adaptive Testing. *Applied Psychological Measurement, 20,* 213-229.

Chang, H.H., & Zhang, J. (1999, June). *Hypergeometric family and test overlap rates in computerized adaptive testing*. Paper presented at the annual meeting of the Psychometric Society, Lawrence, KS.

Chen, S., Ankenmann, R.D., & Spray, J.A. (1999, April). *Exploring the relationship between item exposure rate and test overlap rate in computerized adaptive testing*. Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal, Canada.

Davey, T., & Parshall, C.G. (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, USA.

Hau, K.T., & Chang, H.H. (in press). Item Selection in Computerized Adaptive Testing: Should More Discriminating Items be Used First?. *Journal of Educational Measurement*.

Leung, C.K., Chang, H.H., & Hau, K.T. (2000). *Solving Complex Constraints in a-Stratified Computerized Adaptive Testing Designs*. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, USA.

Lord, M.F. (1970). Some test theory for tailored testing. In W.H. Holzman (Ed.), *Computer Assisted Instruction, Testing, and Guidance*. New York: Harper and Row.

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

Luecht, R.M. (1998). Computer-assisted test assembly using optimization heuristics. *Applied Psychological Measurement*, *22*, 224-236.

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: Latent trait test theory and computerized adaptive testing*. New York: Academic Press.

Mills, C.N., & Stocking, M.L. (1996). Practical issues in large-scale computerized adaptive testing. *Applied Measurement in Education*, *9*, 287-304.

Owen, R.J. (1975). A Bayesian sequential procedure for quantal response in the context of adaptive mental testing. *Journal of the American Statistical Association*, *70*, 351-356.

Stocking, M.L., & Lewis, C. (1995). *A new method of controlling item exposure in Computerized Adaptive Testing*. Research Report 95-25. Princeton, NJ: Educational Testing Service.

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

Stocking, M.L., & Swanson, L. (1998). Optimal design of item banks for computerized adaptive tests. *Applied Psychological Measurement*, *22*, 271-279.

Straetmans, G.J., & Eggen, T.J. (1998). Computerized adaptive testing: what it is and how it works. *Educational Technology*, *38*, 45-52.

Sympson, J.B., & Hetter, R.D. (1985). 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.

Thomason, 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. (1998). Bayesian item selection criteria for adaptive testing. *Psychometrika*, *63*, 201-216.

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

Wainer, H., Dorans, N.J., Flaugher, R., Green, B.F., Mislevy, R.J., Steinberg, L., & Thissen, D. (1990). *Computerized adaptive testing: A primer*. Hillsdale, NJ: Lawrence Erlbaum.

Way, W.D. (1998). Protecting the integrity of computerized testing item pools. *Educational Measurement: Issues and Practice*, *17*, 17-27.

Weiss, D.J. (1982). Improving measurement quality and efficiency with adaptive testing. *Applied Psychological Measurement*, *6*, 473-492.

Table 1: Item Distribution across Content Areas and Cognitive Levels

	Cog1	Cog2	Cog3
Cnt1	52	142	40
Cnt2	33	104	29
Cnt3	27	93	30
Cnt4	28	81	41

Table 2: Ranges and Weights for 19 Non-statistical Constraints

	Lower Bound	Upper Bound	Weight
Cnt1	10	17	40
Cnt2	8	12	40
Cnt3	7	11	40
Cnt4	7	11	40
Cog1	7	11	40
Cog2	19	31	40
Cog3	7	11	40
Cnt1, Cog1	2	5	7
Cnt1, Cog2	6	11	7
Cnt1, Cog3	2	4	7
Cnt2, Cog1	1	3	7
Cnt2, Cog2	4	8	7
Cnt2, Cog3	0	3	7
Cnt3, Cog1	0	3	7
Cnt3, Cog2	3	7	7
Cnt3, Cog3	0	3	7
Cnt4, Cog1	0	3	7
Cnt4, Cog2	4	8	7
Cnt4, Cog3	2	4	7

Table 3: Parameter Distributions across Strata of BASTR

Stratum		*a*-	*b*-	*c*-
1	Mean	.70	.16	.18
	S.D.	.17	1.11	.008
2	Mean	.92	.16	.17
	S.D.	.17	1.04	.008
3	Mean	1.10	.17	.18
	S.D.	.20	1.07	.008
4	Mean	1.38	.17	.17
	S.D.	.31	1.04	.008

Table 4: Summary Statistics for Three Selection Methods

	MI	BASTR	MIBASTR
Bias	-.0051	-.0035	-.0036
MSE	.034	.059	.036
Reliability	.98	.97	.98
Scaled χ^2	195.8	16.6	122.0
N(exp<.02)	472	60	362
N(exp>.2)	86	7	63
Min exp	.000	.000	.000
Max exp	.630	.317	.540
Overlap Rate	.274	.079	.225
Deviations from Non-statistical Constraints			
N(1 dev)	201	127	169
N(2 or above)	0	0	0

Figure 1: Item Exposure Distributions for Three Selection Methods

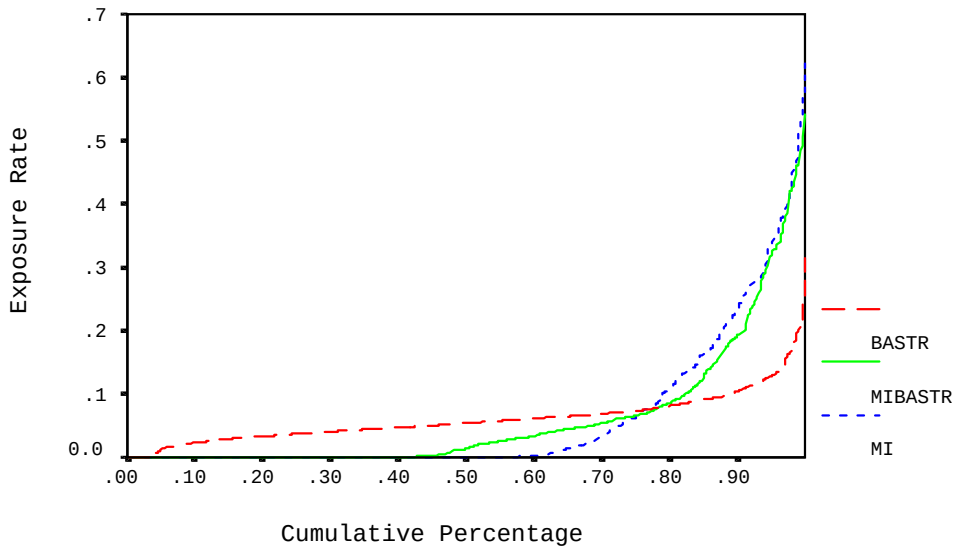

Figure 2: Individual Item Exposure Rates for Three Selection Methods

(a) MI (b) BASTR (c) MIBASTR