Adaptation of a-Stratified Method in Variable Length Computerized Adaptive Testing

Jian-Bing Wen
The Chinese University of Hong Kong

Hua-Hua Chang
National Board of Medical Examiners

Kit-Tai Hau
The Chinese University of Hong Kong

Date of Submission: 28 March, 2001

Note: Paper presented at the American Educational Research Association Annual Meeting, Seattle, April 10-14, 2000. Send all correspondence to Kit-Tai Hau, Faculty of Education, The Chinese University of Hong Kong, Shatin, N.T., Hong Kong. KTHAU@CUHK.EDU.HK.

Abstract

Test security has often been a problem in computerized adaptive testing (CAT) because the traditional wisdom of item selection overly exposes high discrimination items. The a-stratified (STR) design as advocated by Chang and his collaborators (e.g., Chang & Ying, 1999; Hau & Chang, in press) of using less discrimination items in earlier stages of testing has demonstrated to be very successful in balancing and hence maximizing the usage of all items in the pool. However, under specific conditions such as the early stages of utilizing completely new item pools, it is possible that the STR strategy is slightly less efficient than the most widely used maximum information (Max-I) approach. In this series of simulation studies with variable-length CAT in which testing terminates at a targeted test information, we examined whether the use of more items in STR to attain similar accuracy as the Max-I in ability estimation would result in a greater exposure of all items. The simulations with self-generated items as well as an operational pool support the usefulness of the STR method in general. However, the results suggest that it is desirable to have fewer in number but less discriminating items at earlier stages of testing and have more in number of highly discriminating items at later stages. Limitations and implications for future studies are discussed.

Adaptation of a-Stratified Method in Variable Length Computerized Adaptive Testing

Computerized adaptive testing (CAT) in which items are selected from an item pool to fit the test-taker's ability, has become a popular mode of assessment in large-scale public examinations. Despite the attractiveness of high efficiency in accurately locating examinees' ability, test security has often been a problem because the traditional wisdom of item selection overly exposes high discrimination items. One strategy which has demonstrated to be quite effective in balancing the usage of both high and low discrimination items is the a-stratified (STR) design as advocated by Chang and his collaborators (e.g., Chang & Ying, 1999; Hau & Chang, in press).

However, as pointed out by Chang, his strategy of using less discrimination items in earlier stages of testing is a general philosophy that has to be further refined in actual operational implementation. The issue becomes more complicated with variable length CAT. On one hand, the STR may have the benefit of having a balanced item usage. But under some specific conditions such as in early stages of testing with completely new item pools (Hau & Chang, in press), more items are needed due to the longer test length required to attain the targeted accuracy in ability estimation. On the other hand, the traditional item selection method may result in an unbalanced item usage, but less items may be involved during the testing process. In this series of simulation studies, we compare how test efficiency, item usage and other psychometric properties may differ between the traditional maximum information (Max-I) item selection method and the STR strategy. We also examine the benefits and disadvantages when different proportions of high and low discrimination items are used at various stages of the testing process.

Item Selection and Exposure Rate Control

According to Lord's (1970) initial proposal, tailoring tests to test-takers' ability (or other traits) through selection of appropriate items would be desirable because an examinee is measured most effectively when the items are neither too difficult nor too easy. With the advancement of high speed computers, such a mode of testing has been realized in the early 1990s through various CAT designs. Item calibration, selection and other item pool maintenance of most CAT systems are generally conducted with the item response theory (IRT) which is particularly superior than conventional test theory (CTT) when examinees taking different set of items have to be compared on the same scale (Lord & Novick, 1968).

In CAT, each item is selected according to the examinee's ability currently estimated from previous items attempted (Lord,1980; Weiss, 1982). CAT is considered desirable because examinees' ability can be assessed more accurately using less items than corresponding paper and pencil tests. Another characteristic of CAT is that examinees are now tested with different subsets of items from the item pool at different testing sessions rather than with an identical set simultaneously. This however, leads to a test security problem because examinees at different sessions may share item content before their actual testing. The problem will become particularly serious when items are repeatedly used too many times before retiring and there is a substantial overlap of items between two examinees of similar ability. To tackle this test security problem in CAT, the STR method (Chang & Ying, 1999) has been proposed. In contrast to the widely used Max-I method which will result in the over exposure of the more discriminating items, the STR design proactively uses less discriminating items in early stages of testing (Hau & Chang, in press). However, much is still unknown about the maximization of such strategy in CAT.

Variable Length CAT

In an overly simplified division, there are two types of testing termination methods. In the fixed length method, test will terminate against a fixed number of items while in the variable length type, test will stop when the estimation of ability level has attained a certain accuracy level. In CTT, the measurement error can be estimated for each examinee only when the test has been administered to all students. In contrast, in CAT with items calibrated with IRT, test information can be estimated during any stage of the testing. Measurement error is the reciprocal of the square root of test information and it could be computed from the examinee's estimated ability and the parameters of the items already taken.

Once a certain measurement accuracy has been achieved, the marginal return in further administration of items is low or unnecessary. That is, testing can be terminated when the measurement has become lower than a pre-defined threshold level or in general operational terms, when the Fisher information has reached a certain preset minimal level. Thus, one main advantage of variable length CAT is that while different examinees may be taking tests of different lengths, the measurement errors of all examinees will be approximately the same.

a-Stratified Design

Traditionally in adaptive testing, items which provide the maximum Fisher information will be used, which operationally means that high discrimination (a-parameter) items will be preferentially selected from the pool (Hau & Chang, in press; Owen,1975; Wainer, 1990; Weiss, 1982). Test security thus becomes a problem when these high discrimination items are repeatedly and overly used. Many methods have been proposed to control item exposure rates (e.g., Davey & Parshall, 1995; Stocking & Lewis , 1995, 1998; Sympson & Hetter, 1985). Chang and his colleagues (e.g., Chang & Ying, 1999) have proposed a multi-stage STR design to select items with lower a-parameters first. His argument is that the estimates of the examinee's ability would not be closed to the true value during the early stages of testing. Thus, items with a high a-parameter may not necessarily contribute more information than a low a one. Empirical results comparing the STR design with traditional Fisher information method using simulated and operational pool data showed promising results in terms of their reliability, average bias, mean squared error, number of over-exposed and under-utilized items, chi-squared statistic and test overlap rates (Chang, 2001; Hau & Chang, in press). As an extension to reflect the naturally positive and moderate correlation between item difficulty (b-parameter) and discrimination (a), Chang, Qian and Ying (in press) has also modified the STR design into the b-blocking a-stratified design.

But most of these researches using the STR method are in fixed-length CAT format. The present study compared the STR against the traditional Max-I design in variable-length CATs. While the traditional Max-I approach may lead to an uneven exposure of items in a pool, its high efficiency in ability estimation can result in shorter test length in variable length CATs. This may subsequently lead to a general decrease in item usage. It would thus be meaningful in this research to compare the STR method against the Max-I design to see how the potentially opposing trends act together under the variable length CAT condition.

In the implementation of STR in fixed length CAT, the item pool is partitioned into several stages according to the a item parameter. Items with smaller a are used in the earlier stages while the larger a ones are left towards the end of the test. In this study, the above procedures have to be adapted for the variable length CAT format because we do not know the test length in advance. Thus, instead of dividing the test length which varies from one examinee

to another, we partition the targeted test information into stages. Testing moves from one stage to the next one once a certain predetermined test information has been accumulated.

In this study, we also compare three strategies in partitioning the test information, namely, the increasing, the uniform, and the decreasing information approaches. Specifically, in the uniform approach, testing moves to the next stage when 1/4, 2/4, and 3/4 of the targeted test information (assuming 4 stages) are obtained. In the increasing information approach, relatively less test information (e.g., greater than 1/4) is obtained in the earlier stages while more information (e.g., more than 1/4) is obtained from the later stages. The converse applies to the decreasing approach.

Just for the sake of comparison, we also carry out a fixed length CAT. Other than the maximum information approach, three STR designs are also compared, which are namely, the increasing, uniform, and decreasing length approaches. In the uniform length approach, equal number of items are selected at each individual stage before moving on to the next one. In the increasing length one, less items (e.g., more than 1/4 of the total length) are selected from the earlier stages while more are from the later stages. The converse is true for the decreasing approach.

In general the STR design follows these steps:

1. partition the item pool into m (4 in this study) strata according to the a-parameter with lowest a items being put in the first stratum and the largest a ones in the last stratum;
2. partition the test into m stages as well;
3. In each stages of the test, select items from the *k*th stratum which is closest to the current ability;
4. repeat step 3 until some pre-set proportion of targeted test information or predetermined length has been reached, then goes onto the *(k+1)*th group; and
5. test will terminate when the total test information is greater than the targeted value or when the number of items administered is larger than a predefined number.

Simulation Study

Procedures

The present simulation study compared the traditional Max-I approach against several variations of the STR approaches under the variable length CAT design, which include the increasing, uniform, and decreasing information approaches (see details above). A group of 5,000 simulated examinees' ability was generated from the standard normal distribution N(0,1). Two pools of items were also generated. The first one followed the 2-parameter IRT model and contained 400 items. The items were divided into four equal strata, with each of the 100 item strata having a parameter of 0.5,1,1.5 and 2 respectively. Within each stratum of items having identical discrimination, the item difficulty b followed a standard normal distribution N(0,1). We also replicated our results using another 3-parameter independent item bank with parameters imitating those in a retired operational quantitative test. There were 360 items sorted by the discrimination index into 4 strata of 90 items each. In the max-I approach, items were selected from the pool with differentiation into strata.

In the variable length CAT, the targeted test information of each session to terminate testing was set at 15. Maximum likelihood estimation (MLE) was used to estimate examinees' ability (θ). In the Max-I approach, items were selected from the item pool which provided the maximum Fisher item information upon the currently estimated examinee's ability. In the three

versions of STR design, the item bank was partitioned into four strata in an ascending order of the discrimination parameter while the testing process was divided into four corresponding stages (see details above). The increasing, uniform, and decreasing test information strategies differed in the amount of test information being attained in moving from one stage to another. In the equal information strategy, testing proceeded to the next stage with items selected from the next stratum when test information had attained 25%, 50% and 75% of the total targeted test information. For example, testing progressed to the second stage when test information reached 3.75 (1/4 of 15).

In the increasing test information strategy, increasing more information was obtained from the latter stages of the testing. Operationally, 10%, 20%, 30% and 40% of the test information was achieved in the first, second, third and fourth stages respectively. That is, testing proceeded to the next stage when test information was 1.5 (10%), 4.5(30%), 9(60%) respectively. The decreasing test information strategy operated in exactly the reversed order with 40%, 30%, 20% and 10% of test information obtained from the first, second, third and fourth stages respectively.

For the sake of comparison, we also conducted a fixed length (40 items) CAT using the Max-I and three versions of the STR designs, the latter consisted of increasing, uniform, and decreasing length approaches. Operationally, the ratio of items selected from the four strata in the uniform approach was 2.5: 2.5: 2.5: 2.5, while that for the increasing length approach was 1: 2: 3 : 4 and that for decreasing length was 4: 3 : 2 : 1 respectively. As the total test length was 40, the numbers of items in each stage of the increasing length approach were 4 : 8 : 12 : 16 while those of the decreasing on were 16 : 12 : 8 : 4.

Performance Evaluation Criteria

Various performance indicators are used in the comparison of the different strategies, which include test efficiency, error of ability estimation, item exposure rate and test overlap rate. In the variable length design, one indicator of test efficiency is the total number of items administered or needed to achieve the targeted test information. Test efficiency can then be expressed as the average amount of test information contributing by each item, as follows,

$$\text{Efficiency} = \frac{\sum_{i=1}^{M} \text{inf}_i}{\sum_{i=1}^{M} L_i},$$

where M is the total number of examinees, L_i is the test length of the *i*th examinee, inf_i is the test information of the *i*th examinee.

The accuracy in ability estimation is indicated by the Bias and MSE (mean square error) as defined below:

$$\text{Bias} = \frac{1}{M}\sum_{i=1}^{M}(\hat{\theta}_i - \theta_i);$$

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

where θ_i and $\hat{\theta}_i$ are the true (simulated) and estimated ability of the *i*th examinee. A slightly modification of Chang & Ying (1999) χ^2 statistics is also used to measure the skewness

or unevenness of item exposure, as defined by:

$$\chi^2 = \sum_{i=1}^{N} \frac{\left[A_i - \left(\sum_{i=1}^{N} A_i \Big/ N \right) \right]^2}{\sum_{i=1}^{N} A_i \Big/ N}$$

where N is the total number of items in the bank, A_i is the item exposure rate of the ith item. The lower the χ^2 statistics is, the more uniformly the items are being selected and exposed. Statistically, when all item exposure rates are equal, then χ^2 statistics is 0.

The test overlap rate is another parameter to quantify the extent of a similar set of items being exposed to different examinees. It is defined as the expected number of common items encountered by two randomly selected examinees divided by the expected test length in the test. There are C_M^2 pairs of tests among M examinees, thus, the overlap rate is:

$$R_t = \frac{TO_{\times\ddot{U}} / C_M^2}{(\sum_{i=1}^{M} L_i) / M} = \frac{2TO_{\times\ddot{U}}}{\text{£}\ddot{}M - 1\text{£}\textcircled{9}\sum_{i=1}^{M} L_i}$$

Results and Discussion

The characteristics of the items in the self- and operational pool are tabulated in Table 1. It can be seen that in the former, the characteristics match exactly with the intended values (e.g., for a, mean = 1.25; for b, mean =0, SD = 1).

For the self-simulated item bank, the performance of the decreasing, uniform, and increasing information methods were compared against that of the maximum information method. It was found that the four methods were very similar in terms of their bias, MSE and correlation of ability estimates with true values (see Table 2, Figures 1 to 4). Their bias and MSE were consistently small while the correlations between the estimated and true ability was very high (>.96).

The efficiency of maximum information method was the highest as indicated by its shortest test length among the four methods. Despite this attractive characteristic, it should be noted that with the maximum information method when more and more highly discriminating items retire from the pool due to over exposure, its efficiency cannot be maintained (see Hau & Chang, in press). Similar to earlier findings, results showed that the maximum information method could not raise the usage of the under-utilized items. It had almost 100 (almost 50%) more items which were under-utilized than the stratified methods.

The comparison in the test overlap, chi-square, and number of over exposed items were very consistent in that the increasing information method was better than the maximum information, the uniform information and the decreasing information method. The maximum information method was the second best while the decreasing information method was the worst. Though it cannot be concluded that the increasing information method (10%, 20%, 30%, 40%) as implemented here is the best strategy, the results do point out that it is perhaps more desirable to use relatively fewer items of low discrimination items at the early stages and move quickly to subsequent stages with the use of more discriminating items.

The number of items used at each of the four stages are also summarized in Table 3. It

could be seen that the decreasing and uniform information methods had used relatively large number of items in the earlier stages of testing. As the ability estimation at these stages were generally far from the true ability, the information obtained was also low.

Results with the operational item bank with the variable-length CAT almost replicated that with the self-simulated items (see Table 4, Figures 5 - 8). Basically, the four methods were similar in their accuracy of ability estimation. While the maximum information was the best in terms of its efficiency, the increasing information method is better than the maximum information and other stratified methods in keeping the numbers of over- or under-exposed items to the minimum.

In the fixed-length CAT tests with the operational item bank (see Table 5), all four methods tended to provide unbiased estimated ability. As indicated by the MSE and the correlation between estimated and true ability, the increasing length method was as good as the maximum information method, but both were better than the decreasing and uniform length methods. In terms of efficiency, the maximum information method was better than the increasing length method, both of which were better than the uniform and the decreasing length ones. However, in terms of Chi-Square, test-overlap, number of over- and under-utilized items, all the stratified methods were better than the maximum information method. The results with the self-simulated item bank were generally similar (see Table 6).

In the variable length CAT, all four methods are accurate good in ability estimation. This is understandable because all testing ends at the same targeted test information. However, the maximum information method has the shortest test length and provides the highest efficiency in ability estimation. Similar advantages have also observed in the fixed length CAT in that the maximum information method has the lowest MSE and the highest efficiency. Nevertheless these advantages have to be interpreted cautiously because as testing continues and when more and more highly discriminating items have retired, efficiency will be greatly threatened.

Furthermore, it can be noted that increasing information and increasing length methods show very promising results. As compared to the maximum information method, the increasing information or length methods are only slightly less efficient but clearly outperforms all other methods in having very low chi-square, test overlap, and number of over- or under-utilized items. These results taken together support the usefulness of the stratified method in general. However, the results also point out the necessity and the possible direction in which the stratified method should be fine-tuned. Specifically, the results suggest that it is desirable to have fewer in number but less discriminating items at earlier stages of testing and have more highly discriminating at later stages.

It is easy to understand that if a certain testing strategy, such as the stratified design, has to use more items to achieve the same accuracy in ability estimation, due to the greater number of items being used, the number of overly exposed items would be greater. However, results in the present study show that test efficiency does not necessarily always work against the balanced usage of items. For example, the increasing information or length designs can attain very satisfactory balanced usage of items without much sacrifice of efficiency. The optimum combinations for different testing situations, however, have yet to be determined. Future simulation studies with self-generated or operational item pools are needed to shed light on this and other related issues.

References

Chang, H. H., Qian, J. H. and Ying, Z. L. (in press) a-stratified computerized testing with b blocking. Applied Psychological Measurement.

Chang, H. H. & Ying, Z. L. (1999). A-stratified multistage computerized adaptive testing. *Applied Psychological Measurement, 23(3), 211-222*.

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

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.

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

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

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., & Lewis, C. (1998). Controlling item exposure conditional on ability in computerized adaptive testing. *Journal of Educational and Behavioral Statistics, 23(1), 57-75*.

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.

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

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

Figure 1 Simulated Item Pool using Ascending Information Approach: Exposure Rate of Items Ranked by Discrimination Index

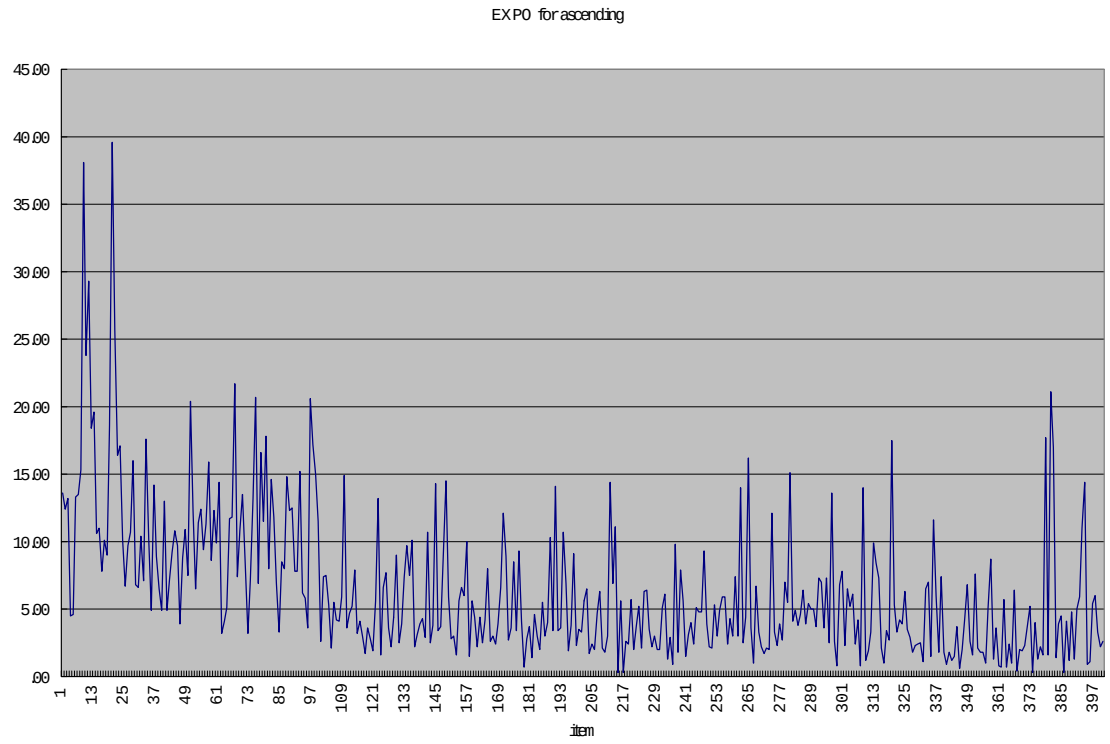


Figure 2 Simulated Item Pool using Uniform Information Approach: Exposure Rate of Items Ranked by Discrimination Index

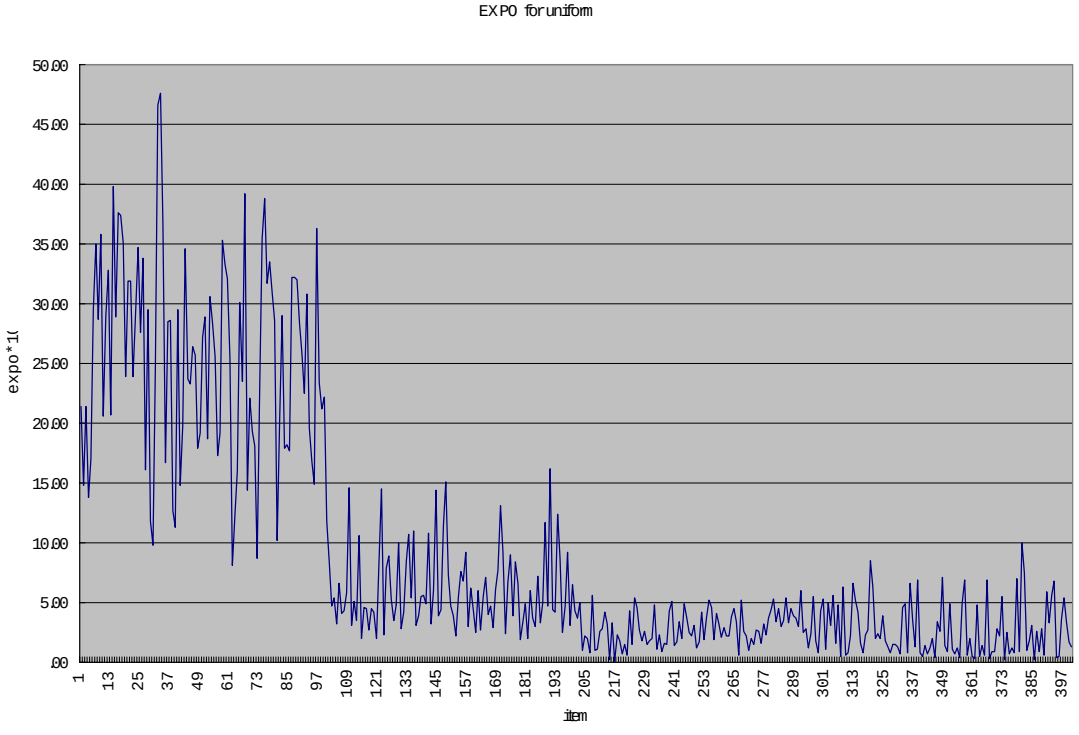


Figure 3 Simulated Item Pool using Descending Information Approach: Exposure Rate of Items Ranked by Discrimination Index

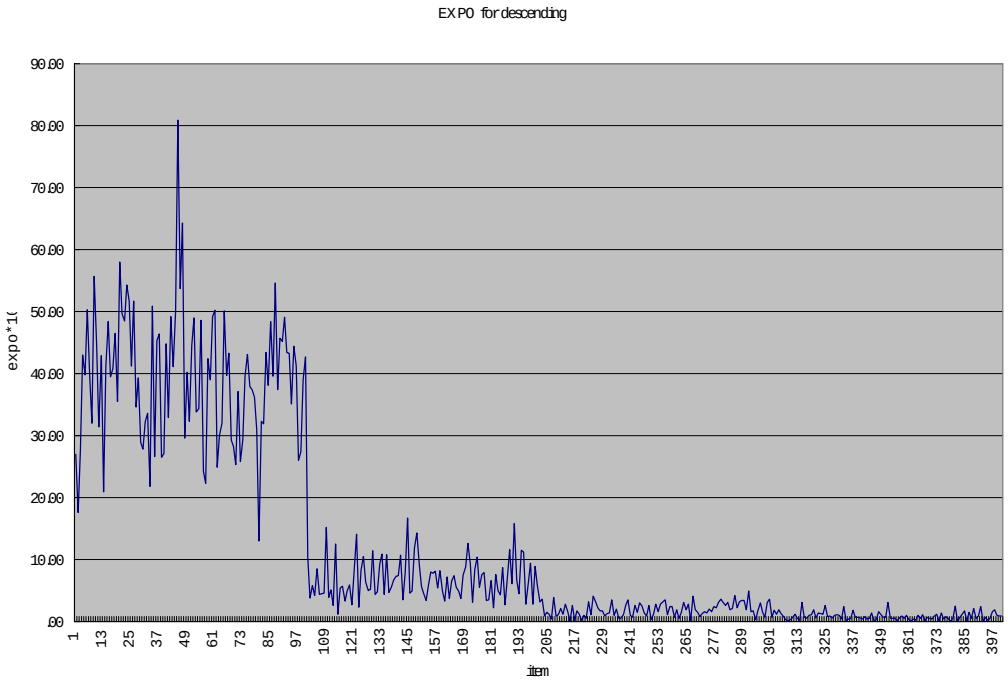


Figure 4 Simulated Item Pool using Maximum Information Approach: Exposure Rate of Items Ranked by Discrimination Index

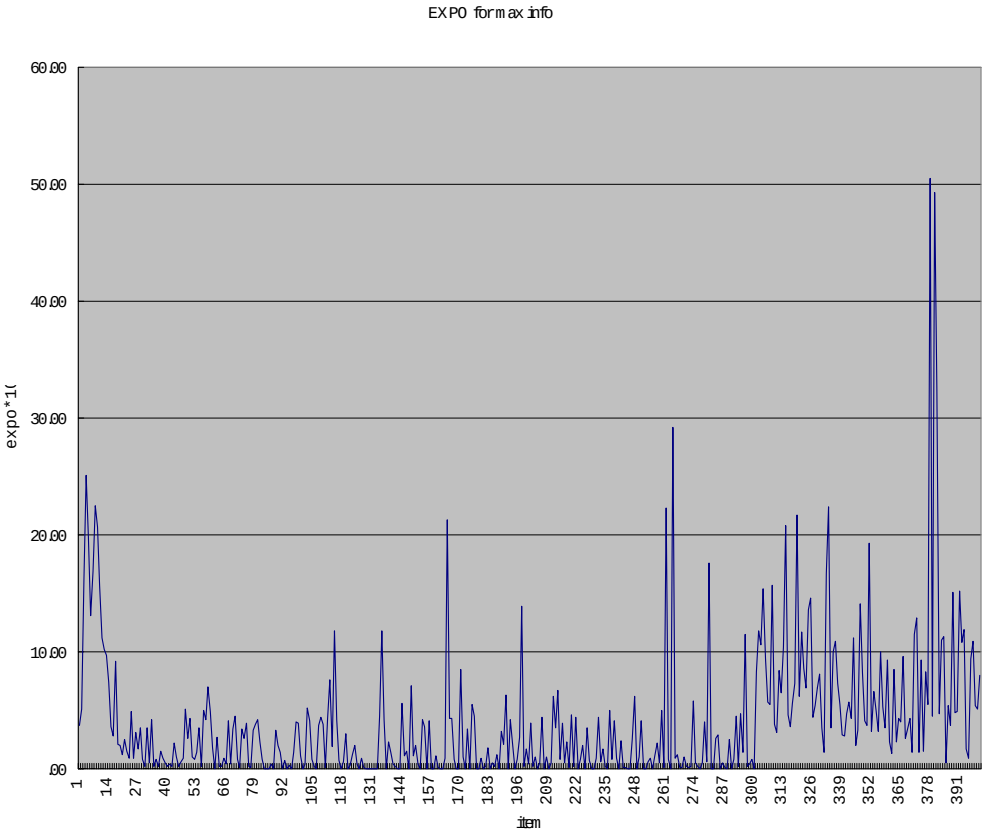


Figure 5 Operational Item Pool Using Uniform Information Approach: Exposure Rate of Items Ranked by Discrimination Index

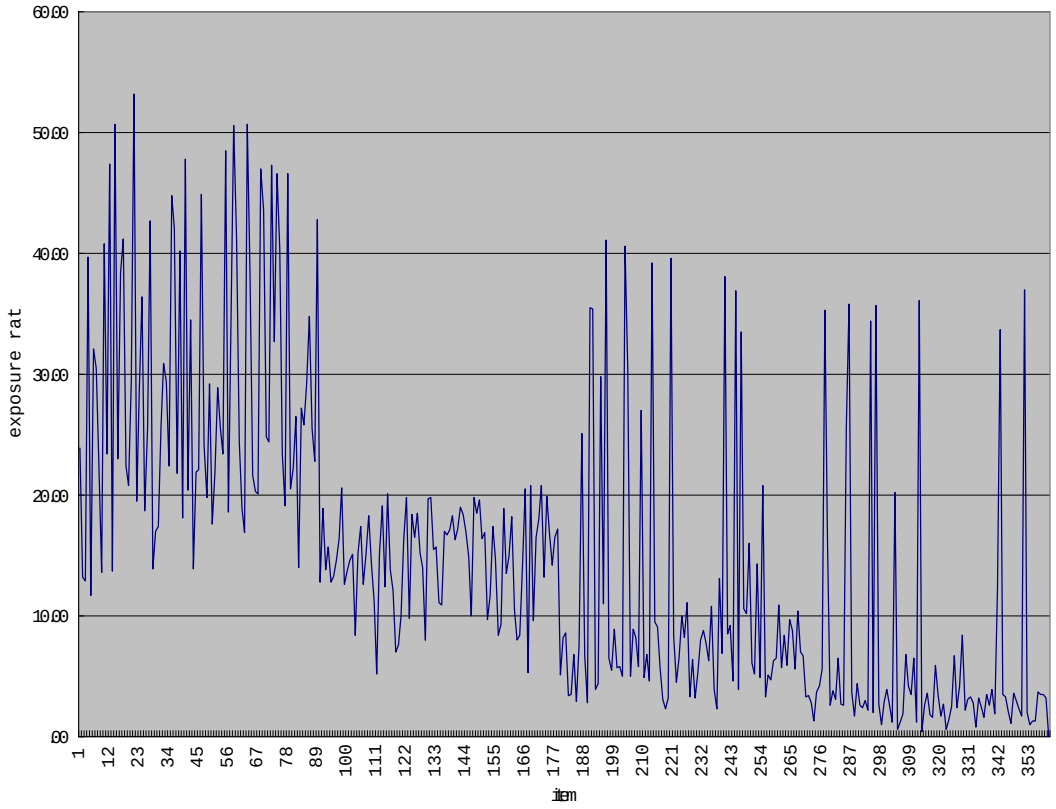


Figure 6 Operational Item Pool Using Ascending Information Approach: Exposure Rate of Items Ranked by Discrimination Index

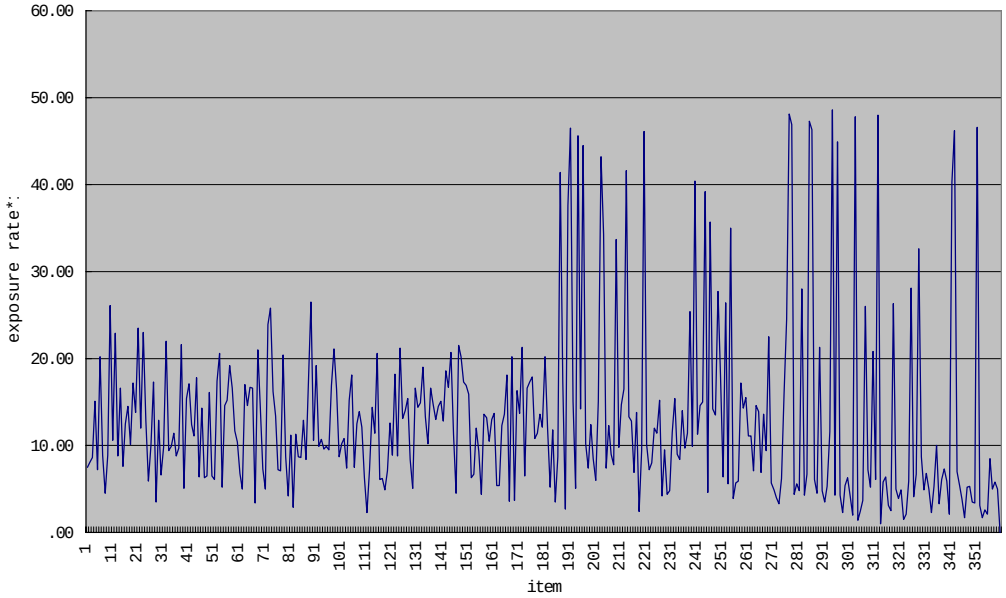


Figure 7 Operational Item Pool Using Descending Information Approach: Exposure Rate of Items Ranked by Discrimination Index

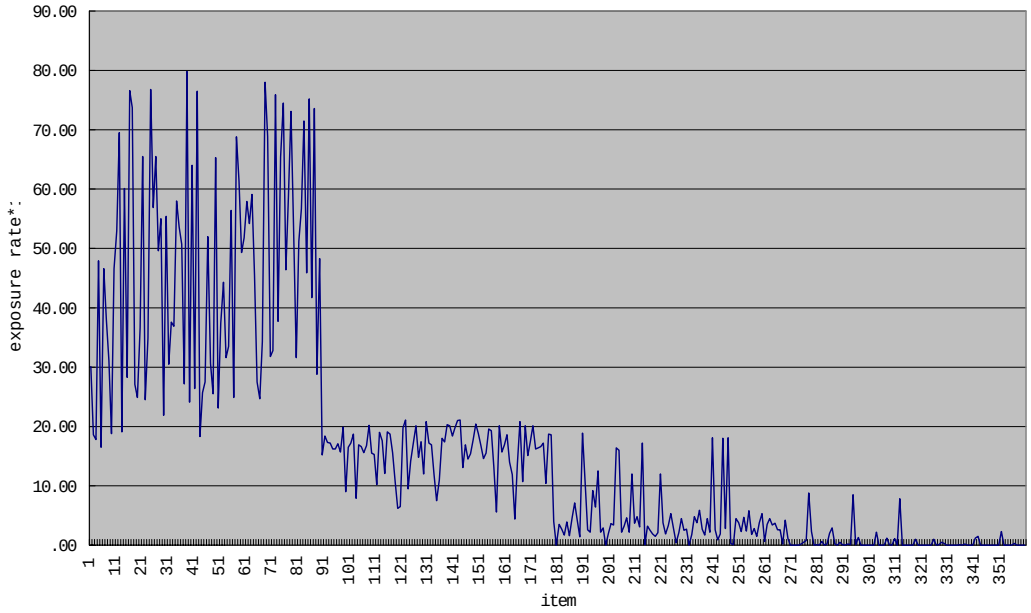
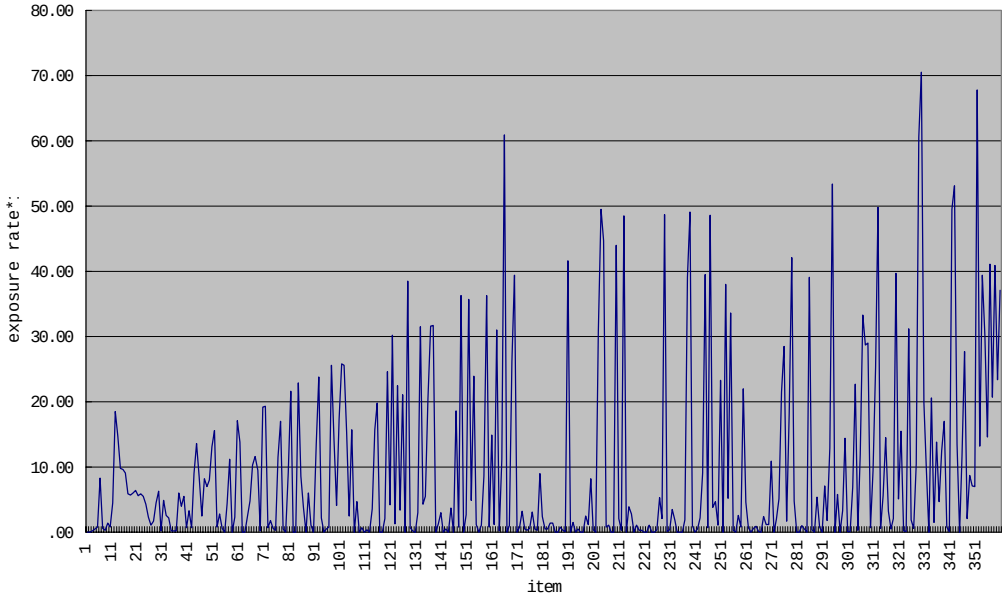


Figure 8 Operational Item Pool Using Maximum Information Approach: Exposure Rate of Items Ranked by Discrimination Index



	Operational Pool			Self-Simulated	
	Discrimination (a)	Difficulty (b)	Guessing (c)	Discrimination (a)	Difficulty (b)
Mean	0.87	0.14	0.16	1.25	-0.01
Std Dev	0.31	0.99	0.11	0.56	1.00
Max	2.00	2.21	0.50	2.00	2.76
Min	0.26	-2.89	0	0.50	-2.72

Table 1 Item Characteristics of the Operational and Self-Simulated Pools

	Decreasing Information	Uniform Information	Increasing Information	Maximum Information
Bias	-0.006	-0.005	-0.003	-0.001
MSE	0.067	0.067	0.070	0.075
Correlation	0.968	0.967	0.966	0.963
Average test length	47.09	33.81	21.31	9.43
Efficiency	0.349	0.476	0.763	1.741
Chi square	83.26	44.80	12.17	24.63
Test overlap	32.57%	19.64%	8.35%	8.50%
# of item over exposed(r>0.20)	98	63	1	4
# of item under utilized(r<0.05)	236	237	241	340

Table 2
Simulation Results with Self-Simulated Item Bank in Variable-Length CAT

	Self-Simulated Item Bank				Operational Item Bank			
	Stage 1	Stage 2	Stage 3	Stage 4	Stage 1	Stage 2	Stage 3	Stage 4
Decreasing Information	37.28	6.86	2.02	0.93	41.01	14.19	4.11	0.56
Uniform Information	23.35	6.11	2.83	1.53	26.05	13.21	9.78	5.64
Increasing Information	9.79	5.29	3.48	2.75	11.19	11.17	14.07	10.51

Table 3
Number of Items Used in Each Stages in the Operational and Self-Simulated Item Banks in the Variable-Length CAT

	Decreasing Information	Uniform	Increasing Information	Max-information
Bias	0.011	0.013	0.014	0.020
MSE	0.095	0.084	0.082	0.086
Correlation	0.958	0.961	0.961	0.959
Average test length	59.9	54.7	46.9	31.4
Efficiency	0.205	0.263	0.303	0.480
Chi square	88.27	37.17	27.05	77.37
Test overlap	41.14%	25.50%	20.54%	30.21%
# of item over exposed(r>0.20)	90	103	55	62
# of item under utilized(r<0.05)	156	98	63	232

Table 4
Simulation Results with Operational Item Bank in Variable-Length CAT

	Decreasing Length	Uniform Length	Increasing Length	Max-info
Bias	-0.004	-0.002	-0.001	-0.002
MSE	0.033	0.023	0.018	0.019
Correlation	0.984	0.988	0.991	0.991
Efficiency	0.820	1.187	1.499	1.940
Chi square	16.16	7.43	15.48	70.68
Test overlap	0.140	0.118	0.139	0.277
# of item over exposed(r>0.20)	37	10	24	98
# of item under utilized(r<0.05)	92	42	119	244

Table 5
Simulation Results with Operational Item Bank in Fixed-Length CAT

	Decreasing Length	Uniform Length	Increasing Length	Max-info
Bias	0.012	-0.002	0.000	-0.005
MSE	0.098	0.091	0.088	0.056
Correlation	0.954	0.958	0.959	0.973
Efficiency	0.280	0.335	0.384	0.546
Chi square	25.13	29.58	49.12	78.64
Test overlap	0.181	0.193	0.247	0.329
# of item over exposed(r>0.20)	53	30	29	85
# of item under utilized(r<0.05)	112	61	100	210

Table 6
Simulation Results with Self-Simulated Item Bank in Fixed-Length CAT