Comparing Three Item Selection Approaches for Computerized Adaptive Testing with Content Balancing Requirement

Chi-Keung Leung
Dept. of Mathematics
The Hong Kong Institute of Education
10 Lo Ping Road, Tai Po
Hong Kong

Hua-Hua Chang
Dept. of Educational Psychology
University of Texas at Austin
TX 78712-4155

Kit-Tai Hau
Dept. of Educational Psychology
The Chinese University of Hong Kong
Shatin, Hong Kong

Paper presented at NCME Annual Meeting on April 2, 2002

Comparing Three Item Selection Approaches for Computerized Adaptive Testing with Content Balancing Requirement

Abstract

This study examined and compared the performances of three item selection approaches for computerized adaptive testing with content balancing requirement. Results indicate that the traditional maximum information approach attained highest measurement efficiency but poorest control on item exposure and pool utilization. In contrast, the recently proposed stratification approach offered best control on item exposure and pool usage at the expense of efficiency. The integration of these two approaches provided a modest method that allowed the information approach a certain degree of trade-off of measurement efficiency in return of better item security and pool utilization.

Introduction

In the last two decades, the advancement in computing technology and psychometric theories have accelerated the change of format of large-scale testing from conventional paper-and-pencil (P&P) tests to the form of computerized adaptive testing (CAT) which was first developed under the item response theory models (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. 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). Other advantages include 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 distribution (Davey & Parshall, 1995; McBride & Martin, 1983; Stocking & Lewis, 1995; Sympson & Hetter, 1985; Thomasson, 1995; van der Linden, 1998). In particular, highly discriminating items may be overly exposed while less discriminating ones are never used; and this would eventually damage item security and reduce the cost-effectiveness of item pool management.

It is understandable, therefore, the control of item exposure and the enhancement of pool usage are important issues in computerized adaptive testing 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) have proposed a multi-stage *a*-stratified design (ASTR) that partitions items into several strata in an ascending order of the item discrimination (*a*) parameters. Each test then consists of a matching number 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 stratification approach is that at early stages, the gain in information by using the most informative items may not be realized because the ability estimation is still relatively inaccurate. Thus items with high discrimination values should be used at later stages. The ASTR has been shown through simulation studies to be effective in both reducing test-overlap rate and enhancing pool utilization, when it is used with certain types of item pools. Nevertheless, the correlation of *a*- and *b*- parameters in some pools may be significant such that there would 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 remedy this undesirable effect, Chang, Qian, and Ying (2001) have developed the *a*-stratified with *b*-blocking method (BASTR). On the other hand, Yi and Chang (2000) proposed another alternative called multiple stratification (denoted by CBASTR here) in which items are divided into groups according to their content areas, *b*-parameters, and then *a*-parameters.

Recently, Leung, Chang, and Hau (2001) have proposed a mixed approach for multiple-constraint CAT to capture the strengths of both the information approach and stratification approach. In this new approach, item pool is partitioned into two large strata following the stratification approach and the testing is divided into two stages accordingly. As the last stratum is larger, there will be sufficient items to satisfy unfulfilled non-statistical constraints at the last stage. Further, in the first stage when there is little information about the true ability, less discriminating items are selected by matching *b*-parameter with the current ability estimate; and in the second stage when more information are accumulated, highly discriminating items are selected based on information to accurately locate the true ability. The new mixed strategy, when integrated with the Weighted Deviation Model (Stocking & Swanson, 1993), has demonstrated its inherited strength of the information approach in maintaining high measurement efficiency and those of the stratification approach in reducing test-overlap rate and in enhancing pool usage under multiple-constraint setting.

In some situations, CAT design has to take into consideration practical requirements such as content balancing. To ensure that each adaptive test has the same mix of contents, some mechanisms are needed to make sure that strict content balancing is incorporated. A method called constrained CAT (CCAT) was proposed by Kingsbury and Zara (1989). Basically, this content-balancing algorithm selects the most optimal item from the content area having current exposure rate farthest below its ideal administration percentage for each examinee. The CCAT is widely adopted because of its effectiveness in addressing content balancing requirement and simplicity for implementation.

As the potential advantages of the mixed method have been demonstrated under a multiple-constraint setting, they cannot be automatically generalized to CAT with specific content balancing requirement. In this study, the three item selection approaches were compared under such a situation. The CCAT was incorporated so that each adaptive test met the restrictive content specifications. The performance of each individual item selection method was evaluated in terms of correlation, average bias, mean squared error,

number of over-exposed items, item overlap rate, and number of under-utilized items.

Method

Item Selection

Method 1 (MI-CCAT): At each step of testing using this information driven selection algorithm, the most informative unadministered items was selected from the content area having the current exposure rate farthest below its ideal administration percentage for each test.

Method 2 (CBASTR-CCAT): The steps for this stratification approach algorithm were as follows:

(i) The number for strata and testing stages was set as 4.

(ii) The item pool was divided into 4 groups based on the content specifications: one group per content specification.

(iii) Items in each group were sorted according to ascending order of *b* parameters.

(iv) Each group was partitioned into 20 blocks; items with the lowest *b* items go to the first block and highest *b* items go to the last block.

(v) Items within each block were sorted in an ascending order of *a* parameters.

(vi) The sorted items of each block were divided into 4 strata according to *a* parameters; the lowest *a* item to the first stratum and the highest *a* item to the last stratum.

(vii) In each stage of testing, items were selected from the corresponding stage. At each step of item selection, an optimal item was chosen from the content area having current exposure rate farthest below its ideal administration percentage by matching the *b*-parameter with the current ability estimate.

Method 3 (CBASTR-MI-CCAT): The steps for this mixed approach algorithm were as follows:

(i) The item pool was partitioned into 2 levels. The first level corresponded to the first stratum in Method 2 and the second level was formed by merging the last three strata.

(ii) Accordingly, each test was divided into 2 stages: The first one-quarter of test items were selected from the first stage and the last three-quarters from the final stage.

(iii) The item selection rule for the first stage followed the *b*-matching criterion as described in Step (vii) of Method 2 and the selection rule for the final stage followed the information strategy as described in Method 1.

Simulation Design

The representative sample consisted of 5000 simulated examinees with true abilities (θs) randomly generated from $N(0, 1)$. A pool of 700 calibrated mathematics items from four content areas was used. The test length was fixed at 40 items and each adaptive test consisted of 14, 9, 9, and 8 items from the four areas respectively.

Evaluation

The performance of each item selection method was evaluated in terms of (a) correlation between θ and $\hat{\theta}$, (b) average bias, (c) mean squared error, (d) number of over-exposed items with exposure rate $\geqq .2$, (e) number of under-utilized items with exposure rate $\leqq .02$, and (f) item overlap rate.

Results

Table 1 summarizes the results of the study. The three item selection methods appeared unbiased as their estimated biases were all close to zero. They offered high and comparable correlation coefficients for the true and estimated abilities. In terms of MSE, the MI (.037) and the MI-CBASTR (.041) provided better measurement efficiency while the CBASTR (.061) was less efficient.

Table 1

Performance Summaries for the Item Selection Methods

	MI	CBASTR	MI-CBASTR
Bias	.0015	-.0001	-.0039
MSE	.037	.061	.041
Correlation	.98	.97	.98
N(exp<.02)	488	123	405
N(exp>.2)	80	15	62
Min exp	.00	.00	.00
Max exp	.63	.38	.52
Overlap Rate	.29	.10	.21

Note: CCAT was incorporated in all selection methods

As regards item security, the CBASTR performed better than the other two methods. It overly exposed about 2% of the items while the corresponding figures associated with the MI and the MI-CBASTR were 11.4% and 8.8% respectively. The maximum exposure rate for the CBASTR was .38 that was much lower than those for the MI (.63) and the MI-CBASTR (.52). Furthermore, the item overlap rates for the CBASTR, the MI and the MI-CBASTR were .10, .29, and .21 respectively, meaning that in average there were 4 items in common among a random pair of examinees in the CBASTR but 12 items and 8 items in common for the MI and the MI-CBASTR respectively.

In regard to pool utilization, the CBASTR also outperformed the other two. There were about 17.6% of items under-utilized in the CBASTR while the corresponding figures for the MI and the MI-CBASTR were 69.7% and 57.9% respectively.

Figure 1

Cumulative Exposure Distributions for the Item Selection Methods

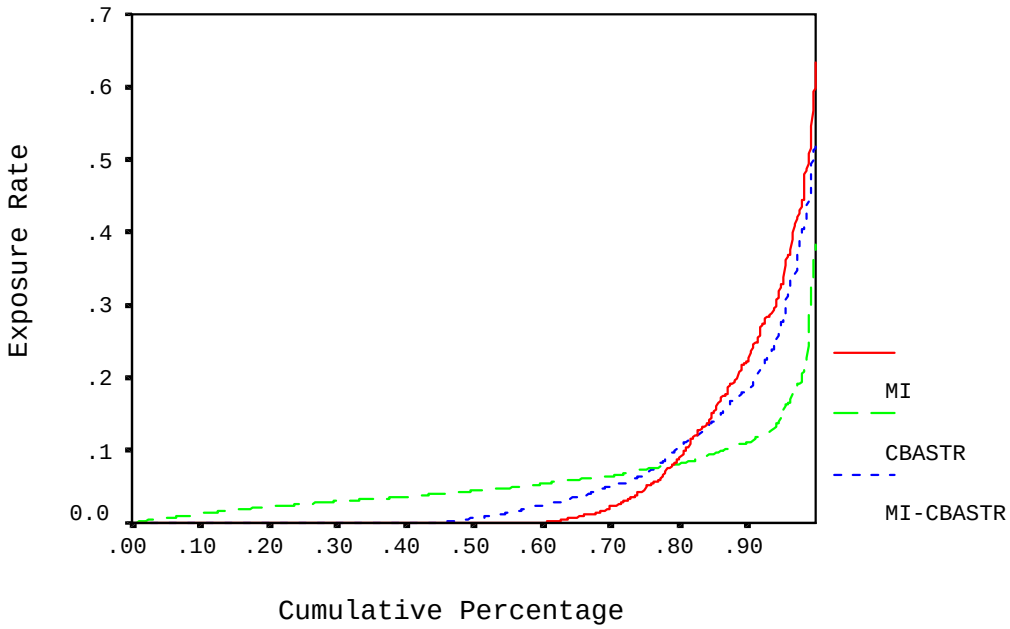


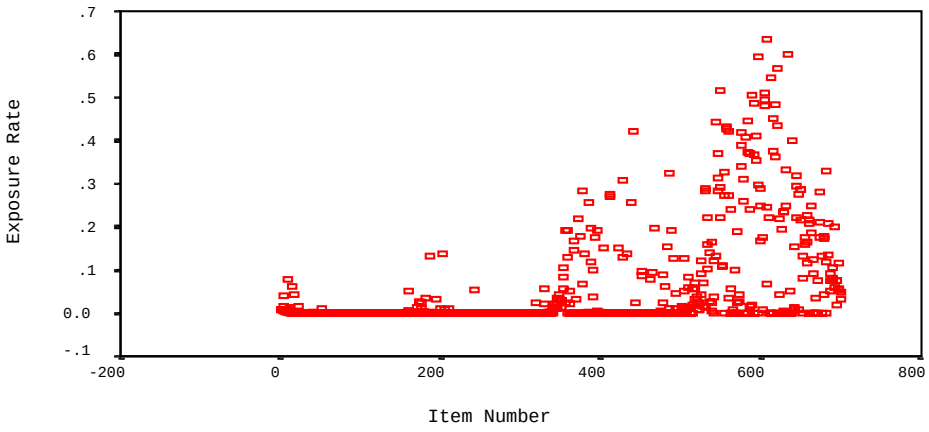
Figure 1 presents the cumulative exposure distributions for the three item selection methods. The curves for the MI and the MI-CBASTR are much skewed, when compared with that for the CBASTR. The bottoms of the curves show that the two information-related methods never made use of about 60% and 45% of the items, while the CBASTR utilized almost every single item of the pool. In addition, the problem of over-exposure was less serious in the CBASTR than that in the other two methods as indicated by the tops of the curves that represented the observed maximum exposure rates and the portions of the curves above the exposure rate of .2.

Figure 2 displays the individual exposure rates of items that were ordered according to their positions in the four strata as described in Method 2. The first 175 items belonged to the first stratum, the next 175 from the second stratum, and the last 175 items from the fourth one.

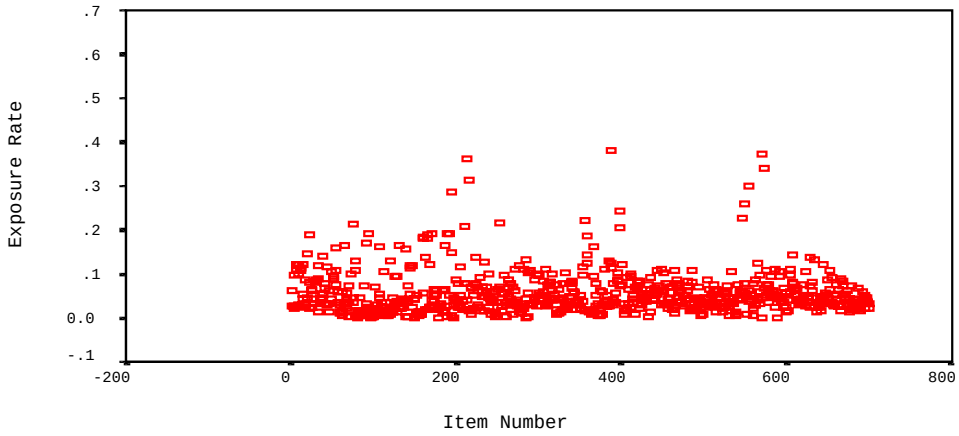
Figure 2

Individual Exposure Rates for the Item Selection Methods

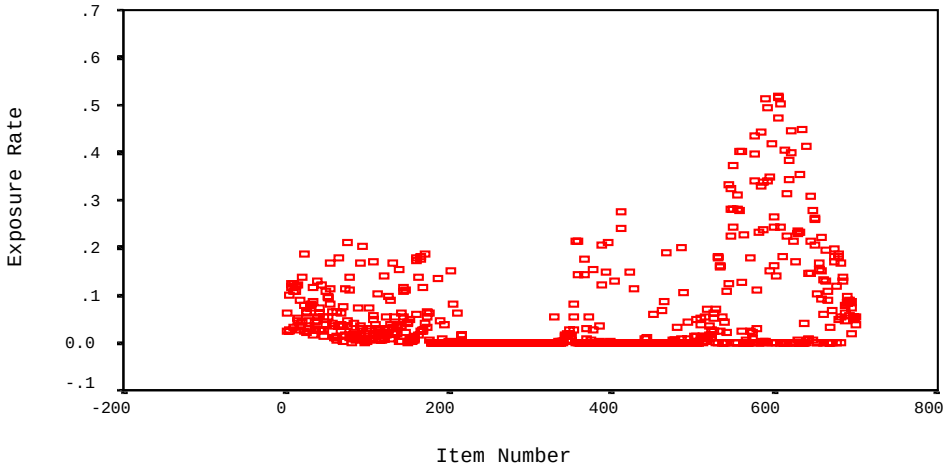
(a) MI



(b) CBASTR



(c) MI-CBASTR



It was observed that MI repeatedly administered high discriminating items clustered around the last two strata and left those less discriminating items in the first two strata never used or heavily under-utilized. In contrast, the CBASTR well utilized the entire item pool and yielded an even exposure distribution. Figure 2(c) indicates that the MI-CBASTR tended to equalize the item exposures for the first stratum and then yielded similar but less dense exposure pattern as the MI did for the rest of the items.

Discussion

As CAT has many promising advantages over traditional practice of P&P, it is anticipated that more and more large-scale educational testing programs will be available in this format. An item selection rules that can simultaneously (a) maintain high measurement efficiency, (b) improve test security by controlling both item exposure rates and test-overlap rate, (c) enhance pool utilization, and (d) satisfy content balancing requirement, are likely to receive much attention.

The present study demonstrates that all the three item selection approaches, when working with the CCAT, can satisfy the content balancing requirement. The results reflect that there are some degrees of trade-off between efficiency and the other two practical concerns of item security and pool utilization. It was observed that no individual method outperformed the others in all evaluation criteria. In general, the information based method attained better measurement efficiency while the stratified design offered better item security control and pool utilization.

When the MI and the CBASTR are considered to be two opposite extremes, their integration of MI-CBASTR appeared to be a modest method that allows the MI to acquire more control on item security and pool utilization at a little expense of efficiency. The contrast of Figures 2(a) and 2(c) leads to a conclusion that the *b*-matching selection criterion in the first stage of MI-CBASTR tends to equalize the item exposures. Furthermore, the use of less discriminating items in the first stage does not affect much on the overall efficiency (refer to Table 1). This outcome has some implications for CAT design. In particular, a testing program can uplift some of its inactive items by adopting the integrated approach and deliberately placing these items in the first stratum.

This study examined the performances of the three item selection methods in an environment where content balancing requirement was imposed. In some situations, stringent exposure control is necessary for the sake of item security. Under such a control, the utilization of active items that are usually more informative may be suppressed. As such, stringent exposure control is anticipated to exert greater impact on the efficiency of the information based selection methods than the stratified ones. Therefore, it would be of both research and practical interest to investigate and compare the performances of the three item selection approaches under exposure control in addition to content requirements. Furthermore, varying the pool size and test length may also affect the performance of individual item selection methods. Future research may explore how to optimize individual methods under various combinations of the potential factors.

References

Chang, H.H., Qian, J., & Ying, Z. (2001). A-stratified multistage CAT with b-blocking. *Applied Psychological Measurement, 25,* 333-341.

Chang, H.H., & Ying, Z. (1999). A-stratified Multistage Computerized Adaptive Testing. *Applied Psychological Measurement, 23*(3), 211-222.

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.

Leung, C.K., Chang, H.H., & Hau, K.T. (2001, April). *Integrating stratification and information approaches for multiple constrained CAT*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Seattle, 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.

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.

Yi, Q., & Chang, H. (2000). *Multiple stratification CAT designs with content control*. Unpublished manuscript.