Content Balancing in Stratified Computerized Adaptive Testing Designs

by

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 AERA Annual Meeting 2000, New Orleans.

Content Balancing in Stratified Computerized Adaptive Testing Designs

Chi-Keung Leung HKIEd
Hua-Hua Chang NBME
Kit-Tai Hau CUHK

Abstract

For computerized adaptive testing (CAT) programs that adopt three-parameter logistic model, high discriminative items generally provide more information for trait estimation. The information based item selection methods in administering computerized adaptive tests (CATs) tend to choose the item that provides maximum information at an examinee's estimated trait level. 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. On a different line of thought, Chang and Ying (1999) proposed the *a*-stratified design (ASTR) that attempts to equalize item exposure distribution by uplifting the usage of low *a* items. The method has been demonstrated to be effective in improving the utilization of the entire pool without sacrificing the efficiency in ability estimation when it is used with certain types of item pools. Nevertheless, the ASTR may result in a number of items being over-exposed in some pools where the correlation between the *a*- and *b*-parameters is significant. To remedy the over exposure problem, Chang, Qian and Ying (1999) developed the *a*-stratified with *b*-blocking method (BASTR) based on ASTR. These two stratified methods have not been tested under situations where content specifications are imposed. For addressing the issue of content balancing, an adaptation of the general ideas of the constrained CAT (CCAT; Kingsbury & Zara, 1989) to ASTR and BASTR was investigated in this study. In addition, the effects of incorporating Sympson-Hetter (SH; Sympson & Hetter, 1985) exposure control into ASTR and BASTR were also examined. The findings indicate that both ASTR and BASTR, with or without SH exposure control, can meet the content specifications, make better utilization of item pools, and yield lower test-overlap rates.

Introduction

The advances in modern computer technology and psychometrics have triggered the change of format of 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 (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). It is also better because it allows more flexibility in test schedule 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 skew 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. This would eventually damage test security and increase the cost in developing and maintaining item pools.

It is understandable, therefore, the control of item exposure and the enhancement of pool efficiency 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) proposed a multi-stage *a*-stratified design (ASTR) that partitions items into several strata in the ascending order of the item discrimination parameter. Each test then consists of 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 design 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 stratified design has been shown through simulation studies to be effective in the reduction of test-overlap rate and the enhancement of 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. 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 this situation, Chang, Qian and Ying (1999) have developed the *a*-stratified with *b*-blocking method (BASTR) based on ASTR. In BASTR, item pool is first divided into many small levels based on *b* parameters. Within each level, items are sorted in ascending order of *a* values. Then the items with smallest *a* values from each level are grouped into the first stratum, and the items with second smallest *a* values grouped into the second stratum,…, and eventually the last stratum contains those items with largest *a* values from each level.

In many situations, however, CAT design has to take into consideration additional constraints such as content balancing. For example, the school-heads participating a pilot study for introducing CAT at the end of the first learning stage (i.e. Primary 3) had expressed their wish to have individual reports for students stating their overall Mathematics abilities as well as their performances in each of the four major content areas. To ensure that each adaptive test measures all the four constructs, some mechanisms are needed to make sure that each test has the same mix of content areas. A method called constrained CAT (CCAT) was proposed by Kingsbury and Zara (1989). This content-balancing algorithm selects the most informative unadministered item from the content area that is farthest below its ideal administration percentage for each examinee.

As the performances of ASTR and BASTR have not been tested under situations where content

constraints are imposed, we propose to adapt the general ideas of CCAT into these stratified methods for such situations. This means that content specifications are imposed in the stratified CAT where items are selected from the areas with quota not fully used up. As a result, the CATs administered using these enhanced stratified methods would have a better face validity. In this study, the adaptation of the general ideas of CCAT into ASTR and BASTR were investigated and the performances of the integrated methods, with or without Sympson-Hetter exposure control, were examined.

Item Selection Method

Six item selection methods were compared in this study: three without SH exposure control and the other three with SH control. Content balancing was imposed in each method.

Method 1 (MI; the maximum information approach):

i) Items were labeled according to the content areas identified.
ii) Content specifications for individual CATs were set (e.g. the number of test items from each of the identified content areas).
iii) At each item administration stage, the best unadministered item with maximum information at the current ability estimate was selected.
iv) The content area of the selected item was checked whether its quota was filled up. If not, the best item was administered. Otherwise, the process of selecting next best item and checking of its content area continued until an item was administered.
v) Steps (iii) and (iv) were repeated until the test was completed.

Method 2 (ASTR; the *a*-stratified design):

i) Items were labeled according to the content areas identified.
ii) Content specifications for individual CATs were set (e.g. the number of test items from each of the identified content areas).
iii) The item pool was partitioned into k strata in an ascending order of the a-parameter.
iv) Each test was divided into k stages in which a specified number of items were administered from the corresponding stratum.
iv) At each item administration stage, the best unadministered item of the corresponding stratum with difficulty closest to the current ability estimate was chosen.
v) This item was checked whether it belonged to a content area that had already used up its pre-specified quota. If not, it was administered. Otherwise, the next best item was identified and checked against the desired content areas before administration. The process continued until a suitable item was administered. In case that no suitable item was found at the current stratum, backward searching at previous stratum was imposed.
vi) The test moved to the next stage when a pre-specified number of items were administered as described in Step (v).
vii) The test stopped when all stages were completed.

Method 3 (BASTR; the *a*-stratified design with *b*-blocking):

i) Items were labeled according to the content areas identified.

ii) Content specifications for individual CATs were set (e.g. the number of test items from each of the identified content areas).

iii) The item pool was partitioned into *k* strata as follows. First the pool was divided into different levels based on the *b*-parameters so that items in each level had similar *b* values. Within each level, the items were sorted according to the ascending order of *a*-parameters. Items with the lowest *a* values from each level were assigned to the first stratum whilst items with highest *a* values were assigned to the last stratum. As a result, each stratum then had similar distribution of the *b*-parameters but the average value of the *a*-parameters increased across the strata.

iv) Each test was divided into *k* stages in which a specified number of items were administered from the corresponding stratum.

v) At each item administration stage, the best unadministered item of the corresponding stratum with difficulty closest to the current ability estimate was chosen.

vi) This item was checked whether it belonged to a content area that had already used up its pre-specified quota. If not, it was administered. Otherwise, the next best item was identified and checked against the desired content areas before administration. The process continued until a suitable item was administered. In case that no suitable item was found at the current stratum, backward searching at previous stratum was imposed.

vii) The test moved to the next stage when a pre-specified number of items were administered as described in Step (vi).

viii) The test stopped when all stages were completed.

Method 4 (MI_SH; integrated method of MI and SH):

i) The target maximum exposure rate, r_j, for the *j*th item was set at .2, a commonly used cut-off value.

ii) Prior to the actual testing, 25 cycles of SH adjustment iterations were run to obtain exposure control parameters for individual items. First, the initial exposure control parameters were set to 1. In each iteration, simulated CATs as described in Method 1 (MI) were administered to a large group of 5,000 simulees randomly sampled from the ability distribution of the expected real examinee population. The frequency for the *j*th item being selected and the frequency of that item being administered were tallied. The proportions, $P(S_j)$ and $P(A_j)$, of these frequencies as in the whole sample were computed. The value of $P(A_j)$ was then compared to r_j. If $P(A_j)$ was larger then r_j, the exposure parameter for that item would be adjusted downward to r_j / $P(S_j)$, otherwise the exposure parameter would be adjusted upward (to a maximum of 1) by multiplying a factor (e.g. 1.05). A set of exposure control parameters was obtained after the 25 adjustment iterations were completed.

iii) An item with maximum information from a content area with quota not fully used up was selected as described in Method 1.

iv) The exposure parameter of the selected item was checked against a random number generated from the uniform distribution U(0,1). If the parameter value was greater than the random number, the item would be administered. Otherwise, the next best item would be selected and checked against a new

random number. This step continued until an item was administered or the pool was exhausted.

v) Steps (iii) and (iv) were repeated until the test was completed.

Method 5 (ASTR_SH; integrated method of ASTR and SH):

i) The target maximum exposure rate, r_j, for the jth item was set at .2, a commonly used cut-off value.

ii) Prior to the actual testing, 25 cycles of SH adjustment iterations were run to obtain exposure control parameters for individual items. First, the initial exposure control parameters were set to 1. In each iteration, simulated CATs as described in Method 2 (ASTR) were administered to a large group of 5,000 simulees randomly sampled from the ability distribution of the expected real examinee population. The frequency for the jth item being selected and the frequency of that item being administered were tallied. The proportions, $P(S_j)$ and $P(A_j)$, of these frequencies as in the whole sample were computed. The value of $P(A_j)$ was then compared to r_j. If $P(A_j)$ was larger then r_j, the exposure parameter for that item would be adjusted downward to r_j / $P(S_j)$, otherwise the exposure parameter would be adjusted upward (to a maximum of 1) by multiplying a factor (e.g. 1.05). A set of exposure control parameters was obtained after the 25 adjustment iterations completed.

iii) An item with b closest to the current ability estimate from a content area with quota not fully used up was selected as described in Method 2.

iv) The exposure parameter of the selected item was checked against a random number generated from the uniform distribution U(0,1). If the parameter value was greater than the random number, the item would be administered. Otherwise, the next best item would be selected and checked against a new random number. This step continued until an item was administered or the pool was exhausted.

v) Steps (iii) and (iv) were repeated until the test was completed.

Method 6 (BASTR_SH; integrated method of BASTR and SH):

In this method, the procedures for generating exposure control parameters, item selection and item administration were similar to those in Method 5, except that the item pool was partitioned as described in Method 2 (BASTR).

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 six 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 ratio of the estimated scores on 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 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 bias was computed as

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

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, mean squared error was computed as

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

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 test security and validity problems. Since one of the main concerns of this paper was to compare the six 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. 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 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. Wightman (1998) explicitly argued that efficient usage of the available items in the pool would be an important evaluation criterion of CAT assemblies. 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:

$$\chi^2 = \sum_{j=1}^{N}\frac{(er_j - L/N)^2}{L/N} \quad (3)$$

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

Equation 3 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.

Test-overlap rate: The test-overlap rate is another important summary index in measuring item exposure control (Mills & Stocking, 1996; Way, 1998). Test-overlap rate is indicated by the proportion of items shared by pairs of examinees, averaged across all possible pairwise 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 test-overlap rate, the bigger the damage to the 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 test-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 test-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 test overlap.

Simulation Studies

Two simulation studies were conducted to investigate the performance of the ASTR and BASTR, with or without SH exposure control, when content specifications were imposed. Their performances were compared with the other two item selection and administration methods (MI and MI_SH). In each study, simulated tests were administered to a sample of 5,000 simulees with abilities (referred as true ability θ hereafter) randomly generated from N(0,1).

Study 1: P3 Mathematics Items

The item pool for the first study contained 316 items from four content areas each having 41, 58, 169, and 48 items respectively. The items were constructed in such a way that students' motivation was looked after. Thus, the items were relatively easy so that the students were willing to try all the items in paper form for the purpose of valid calibration of item parameters. The mean estimated discrimination (a) value for the items was .94, with SD = .32, and a range of .25 to 2.19. The mean estimated difficulty (b) was -.53, with SD = 1.08, and a range of –3.55 to 2.44; thus there was a mismatch between the b distribution and ability distribution. The mean estimated pseudo-guessing parameter (c) was .18, with SD = .006, and a range of .05 to .44. The correlation of a- and b-parameters was .363.

The test length was set at 32 with content specifications of 4, 6, 17, and 5 from the four areas respectively. Each examinee of the sample was administered with six tests: one from each item selection method.

In each test, information obtained by three artificial items in addition to the prescribed length was used to mimic prior information about the examinee. All the three items had the same values for a (= 1) and c (= .2). The value for b_1 of the first item was randomly selected from N(0, 1). If the first item was answered correctly, the second item was more difficult and the value for b_2 was set at $b_1 + 2$, otherwise $b_2 = b_1 - 2$. The procedure was repeated for the third item. The use of maximum likelihood estimation, as in all six methods, was possible with the administration of these three artificial items.

For the ASTR and ASTR_SH, the item pool was partitioned into four strata in ascending order of *a* parameters. Thus, the first stratum contained items with smallest *a*s and the last stratum contained those with largest *a*s. For the BASTR and BASTR_SH, the item pool was first divided into 79 levels (4 items each) in ascending order of *b* parameters. Then the items with lowest *a*s in each level were grouped into the first stratum,…, and the items with highest *a*s were grouped into the last stratum.

Study 2: Higher Mathematics Items

The item pool for this study contained 420 higher mathematics items from six content areas each having 97, 70, 63, 63, 99, and 28 respectively. The mean estimated discrimination (*a*) value for the items was .96, with SD = .29, and a range of .30 to 1.93. The mean estimated difficulty (*b*) was .18, with SD = .97, and a range of –3.10 to 2.58; thus there was a closer match between the *b* distribution and ability distribution. The mean estimated pseudo-guessing parameter (*c*) was .15, with SD = .005, and a range of .03 to .28. The correlation of *a*- and *b*-parameters was .367.

The test length was set at 40 with content specifications of 9, 7, 6, 6, 9, and 3 items from the six areas respectively.

The procedures for the administration of initial items and pool stratification were similar to those in Study 1 except that pool was first divided into 105 levels (instead of 79) for the BASTR and BASTR_SH.

Results

Study 1 P3 Mathematics Items

The results of the first simulation study are summarized in Table 1. When SH exposure control algorithm was not incorporated, the three methods were virtually unbiased as they all yielded small biases. The three methods had similar reliabilities. The MI method was relatively more efficient as it gave the smallest MSE of .0537. However, its efficiency in trait estimation was at the expense of item and test security. It over-exposed 72 items and caused the highest test-overlap rate. Besides, it yielded the highest scaled χ^2 and left 179 items under-utilized, meaning that MI could not make use of the entire item pool. In contrast, the ASTR and BASTR methods made a better use of all items as they provided much smaller values of scaled χ^2 and smaller numbers of under-utilized items. In addition, they yielded much lower test-overlap rates and over-exposed a smaller number of items, that in turn would cause less threat to item and test security. The BASTR and ASTR provided comparable MSEs and reliabilities, but BASTR performed relatively better by yielding a smaller number of under-utilized items, a smaller number of over-exposed items, a smaller scaled χ^2, and a smaller test-overlap rate.

When SH exposure control was incorporated, the reliabilites remained relatively stable but all the MSEs for the three methods increased, meaning that there was a decrease in estimation efficiency. This SH algorithm appeared working particularly well with the MI method as it substantially reduced the number of over-exposed items, the number of under-utilized items, the test-overlap rates and scaled χ^2 that were much larger when yielded by MI alone. Among the three methods with exposure control, BASTR_SH

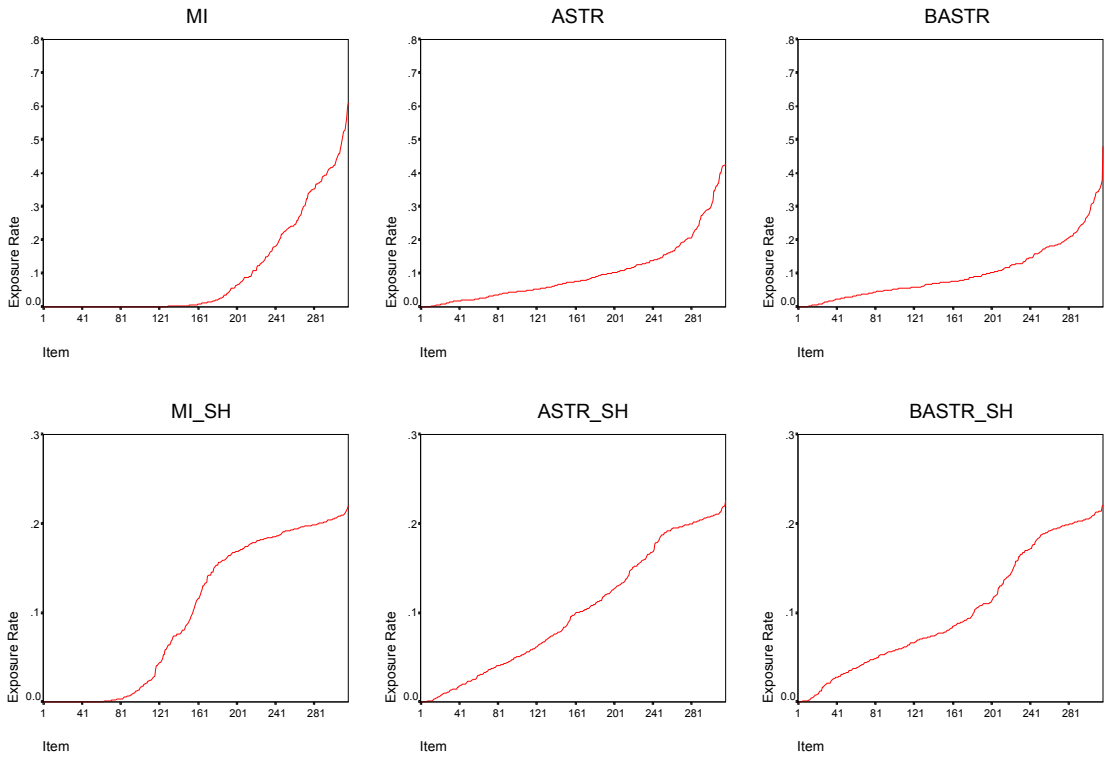
performed better in utilizing the entire pool by yielding the smallest scaled χ^2 and smallest number of under-utilized items. It also provided the smallest test-overlap rate. On the other hand, the MI_SH appeared to be more efficient in trait estimation as it yielded the smallest MSE. The performance of the ASTR_SH apparently was in between the other two methods in all aspects except the number of over-exposed items in which all the three methods yielded similar figures.

Table 1: Performance Summaries for Six CAT Methods in Study 1

Without exposure control	MI	ASTR	BASTR
Bias	.012	.015	-.011
MSE	.054	.082	.084
Reliability	.94	.93	.93
Scaled χ^2	71.2	27.1	22.1
N(exp<.02)	179	43	39
N(exp>.2)	72	42	39
Overlap Rate	.32	.17	.15
With exposure control	MI_SH	ASTR_SH	BASTR_SH
Bias	.011	.021	.010
MSE	.069	.087	.090
Reliability	.93	.92	.92
Scaled χ^2	21.7	15.2	13. 9
N(exp<.02)	105	47	30
N(exp>.2)	30	34	32
Overlap Rate	.16	.14	.13

Figure 1 shows the item usage for the six methods in Study 1. For each graph, items were sorted in ascending order of exposure rates. For MI, the graph lies close to zero exposure rate for about half of the pool and then rises rapidly towards a maximum rate above .6, meaning that many items were untouched while many others over-exposed. For ASTR and BASTR, the graphs rise steadily for most of the items and then go up quickly to maximum rates about .4, showing that item exposure distributions were less skewed and almost every item was utilized. For MI_SH, the graph is flat at the beginning, rises sharply in the middle, and then turns flat again. This means that exposure distribution was extremely skewed where many items were under-utilized or even untouched while many others having exposure rate close to the targeted maximum. In contrast, the graphs for ASTR_SH and BASTR_SH rise quite steadily, showing that exposure rates spread evenly and almost every item was used for testing.

Figure 1: Item Exposures for Six CAT Methods in Study 1



The results of the second study are summarized in Table 2. Compared to the results of Study 1, all the six item selection methods provided smaller MSE and higher reliability. All the methods remained virtually unbiased as the biases were close to zero.

Without SH exposure control, MI was relatively more efficient in trait estimation with smallest MSE. But it certainly caused an alarm in security at both item and test levels as it over-exposed more than 20% (84 items) of the item pool and yielded an unacceptably high test-overlap rate of .310. Besides, economic concern would be raised as half of the item pool was under-utilized. In contrast, at a little expense of estimation efficiency, both ASTR and BASTR made well use of item pool as they yielded much smaller values in the number of under-utilized items and scaled χ^2. In addition, they resulted in much smaller test-overlap rates.

When SH exposure control was incorporated, the unbiasedness of the methods seemed unchanged. As the exposures of active items were suppressed, all the scaled χ^2, test-overlap rates and the numbers of under-utilized items and over-exposed items were lowered, meaning that the item pool was better utilized. The MI_SH over-exposed less items than the other two methods did. This finding is on contrary to the results of previous research (Leung, Chang, & Hau) in which MI_SH over-exposed more items than ASTR_SH did under the situation where there was no content constraints and the SH exposure control parameters were obtained when they were stabilized. In terms of measurement efficiency, MI_SH performed relatively better as it yielded the lowest MSE and highest reliability, though the figures were

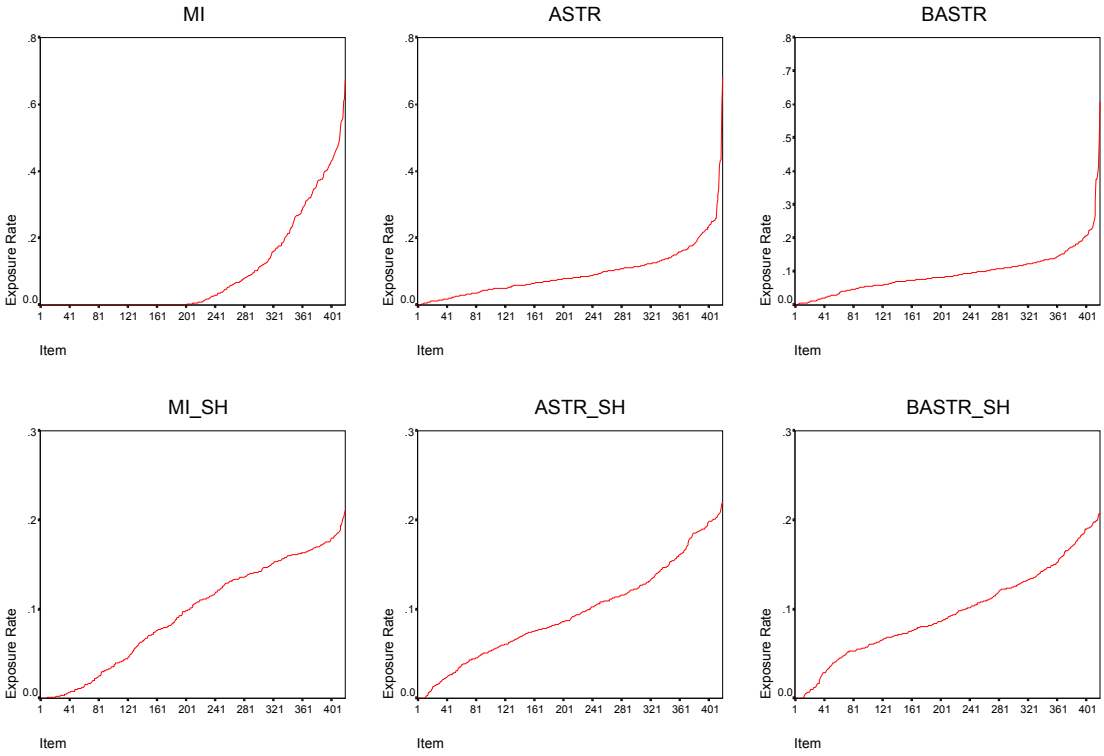
close to those by the other two methods. On the other hand, ASTR_SH and BASTR_SH made well use of the entire pool as they yielded smaller scaled χ^2 and much smaller numbers of under-utilized items. In addition, they provided smaller test-overlap rates.

Table 2: Performance Summaries for Six CAT Methods in Study 2

Without exposure control	MI	ASTR	BASTR
Bias	-.002	.005	-.003
MSE	.044	.067	.070
Reliability	.96	.94	.94
Scaled χ^2	95.7	26.5	20.2
N(exp<.02)	232	45	40
N(exp>.2)	84	33	24
Overlap Rate	.31	.13	.11
With exposure control	MI_SH	ASTR_SH	BASTR_SH
Bias	-.003	.003	-.010
MSE	.061	.069	.077
Reliability	.94	.93	.93
Scaled χ^2	15.9	13.1	11.8
N(exp<.02)	74	33	33
N(exp>.2)	4	17	10
Overlap Rate	.13	.12	.11

Figure 2 shows the item usage for the six methods in Study 2. For each method, items were sorted in ascending order of exposure rates. Similar to Figure 1, the item usage in MI was very uneven, leaving about half of the pool untouched and many others over-exposed. In contrast, the exposure rates for ASTR and BASTR spread evenly with only a few items having exposures high above the targeted value. But unlike Figure 1, the graph for MI_SH rises steadily. Although the three graphs for methods with exposure control look alike, the graph for MI_SH goes up slowly while the other two go up more rapidly at the start. This means that almost every item in ASTR_SH and BASTR_SH was well utilized but some in MI_SH were not touched. The graphs also show that the over-exposed items had exposures quite close to .2, the allowed maximum rate.

Figure 2: Item Exposures for Six CAT Methods in Study 2



Discussion

The ASTR was developed in an attempt to remedy the problems associated with the information-based item selection methods: extremely skewed item exposure distribution and high test-overlap rate. Based on this method, BASTR has been proposed in order to tackle the situations in which the correlation between *a*- and *b*- parameters is significant. The main objective of this study is to investigate whether these two stratified methods could meet content specifications by adapting the general ideas of CCAT and imposing a mechanism that allows backward searching of a suitable item in the previous stratum.

Results indicate that both ASTR and BASTR, with or without SH exposure control, can meet strict content specifications. Even without SH, they made better utilization of the entire pool, offered low test-overlap rates, over-exposed less items, and maintained high reliabilities. When SH method was incorporated, the numbers of over-exposed items were further reduced. The BASTR_SH provided the smallest scaled χ^2 in both simulation studies, yielding the highest reduction in the skewness of item exposure distribution. In addition, the BASTR_SH offered the lowest test-overlap rates, meaning that it reduced the potential damage due to the sharing of information among examinees taking the test at different times. From Figure 2, it seems that MI_SH can improve the pool utilization. But Figure 1 indicates that MI_SH left a large number of items untouched when there was a mismatch of *b* and ability distributions. Thus the robustness of MI_SH in pool utilization needs further investigations. All in all, the results also confirm that there is always some degree of trade-off between estimation efficiency and test security.

There are limitations on the generalization of the findings. Firstly, the two simulation studies

assumed that content specifications are the only constraints of item administration. It is worth to investigate whether the ASTR and BASTR also work well in other situations where complex constraints are imposed. Secondly, the numbers of prior iterations for generating SH exposure control parameters were fixed. The effect of incorporating SH would be different if resources are available to allow the exposure control parameters to be obtained when they are stabilized. Thirdly, as reflected from the results, the performance of individual methods varies with many factors such as the test length and item pool characteristics. Thus, the choice on item selection method really depends on the needs of individual testing programs. Future research also includes the investigation of the effect of varying the number of strata and the changing of stopping rule from fixed length to variable length.

References

Chang, H.H., Qian, J., & Ying, Z. (1999). *a*-Stratified Multisage CAT with *b*-Blocking. Unpublished manuscript.

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.

Educational Testing Service (1998, July). Computer-based GMAT and TOEFL introduced as computer power continues to improve testing. http://www.ets.org/aboutets/zgmattfl.html.

Hetter, R.D., & Sympson, J.B. (1997). Item Exposure Control in CAT-ASVAB. In W.A. Sands, B.K. Waters, & J.R. McBride (Ed.), CAT: from Inquiry to Operation. Washington, DC: American Psychological Association.

Kingsbury, G.G., & Zara, A.R. (1989). Procedures for selecting items for computerized adaptive tests. Applied Measurement in Education, 2, 359-375.

Leung, C.K., Chang, H.H., & Hau, K.T. (1999, April). *An enhanced a-stratified computerized adaptive testing design*. Paper presented at the American Educational Research Association Annual Meeting, Montreal.

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.

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., & 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., & 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.

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., & 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.

Wightman, L.F. (1998). Practical issues in computerized test assembly. Applied Psychological Measurement, 22, 292-302.