An Enhanced Stratified Computerized Adaptive Testing Design

Chi-Keung Leung *Hua-Hua Chang* *Kit-Tai Hau*
Hong Kong Institute of Education *Chinese University of Hong Kong* *Chinese University of Hong Kong & NBME*

Abstract

Item security and cost effectiveness are two an important issues in computerized adaptive testing designs. Over-exposure of items may cause problems in item security and test validity. If an item is pre-exposed to some of the examinees before the test, it can no longer provide valid measurement on the trait that it is developed to measure. If only a small to moderate portion of items are used in tests, the development and maintenance of the item pool would not be cost effective. The traditional item selection method based on maximizing Fisher information has created several problems: high test-overlap rate, over-exposure of some items and highly skewed item exposure distribution. These problems can be partially resolved by incorporating the Sympson-Hetter (SH) probabilistic procedure into the maximizing information method. An a-stratified design is a new concept to address the issues of item security and pool utilization. It has been demonstrated to be effective in lowering the test overlap rate and improving the utilization of the entire pool when content constraints are not of main concerns. But unfortunately it cannot really solve the problem of high item exposure rates when the test length is moderately long. This paper proposes an enhanced stratified design by incorporating the SH procedure into the a-stratified design and compares its performance with the original a-stratified design and Fisher-SH method through simulation studies. The results indicate that the enhanced stratified design yielded much well-balanced item exposure distribution, further reduced the test overlap rate and the numbers of overly exposed items, and maintained high reliability as well as low average bias, mean squared error and scaled chi-square.

Introduction

With the availability of small powerful computers and advancement of psychometric measurement, the concept of tailoring tests for individuals becomes further realized. Some large-scaled periodic tests such as Graduate Management Admissions Test

(GMAT®) and the Test of English as a Foreign Language (TOEFL®) have been completely or partially converted into the form of computerized adaptive testing (CAT; Educational Testing Service, 1998). Most of the CAT designs are built on Item Response Theory (IRT; Lord, 1970). In computerized adaptive testing, 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 conventional paper-and-pencil test (P&P) is that it enables more efficient and precise trait estimation (Owen, 1975; Weiss, 1982; Wainer, 1990). 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. In particular, items with large value of discrimination parameter may be overly exposed while others never used (Chang & Ying, 1996). This would eventually damage item security and inflate the cost of administrating the test.

Control of item exposure is an important issue in computerized adaptive testing designs. Over-exposure of items may cause item security and test validity problems. If an item is pre-exposed to some of the examinees before the test, it can no longer provide valid measurement on the trait that it is developed to measure. Methods to curb high exposure rates have been proposed by Sympson & Hetter (1985), Davey & Parshall (1995), Chang & Ying (1997) and Stocking & Lewis (1995, 1998) among others. The idea of Sympson & Hetter (1985) is to put a filter between item selection and administration. Each item has an exposure control parameter that is determined through a series of adjustment simulations so that the probability of adminsitration is restricted to about the pre-specified maximum exposure rate. The extended method of Stocking & Lewis (1998) is to generate a matrix of item exposure parameters conditional on examinee's abilities while the method proposed by Davey & Parshall (1995) is to restrict the frequency of item administration, conditional on all items that have already been included in the test.

Chang & Ying (1996) argued that the information criterion could be inefficient during the early stages of an adaptive test as the estimated trait may not be close to its true value, and therefore items with high a-parameter values should be saved for later stages. They later proposed an a-stratified CAT design which partitions the item pool into *k* levels according to the values of *a*-parameter (Chang & Ying, 1997).

From the findings of Chang & Ying (1997), one can see that the stratified design provides significant improvement in pool utilization and yields well-balanced item exposure distribution. However it could not guarantee that the exposure rate for every item is below certain threshold. In this paper, the authors suggest an integrated method that incorporates the Sympson & Hetter's idea into Chang & Ying's stratified design. This means that a filter is added in between item selection and administration in all stratified stages. The effectiveness of the enhanced stratified design is compared with that of (1) the method of Fisher Information Criterion with Sympson-Hetter Measure and (2) Chang & Ying Stratified Design. The comparison will be based on reliability, average bias, mean squared error, scaled chi-squared statistic, number of overly exposed items, number of under-utilized items and test overlap rates.

Methods

Method 1 (STR): The steps of multi-stage a-stratified design (Chang & Ying, 1997) are as follows:

(i) The item pool is partitioned into *k* subpools according to the ascending order of a-parameters of items. The first subpool contains items with lowest a-parameter values; the second subpool contains those with second lowest a-parameter values,..., and the last subpool contains those with largest a-parameter values.

(ii) Accordingly, an entire test is partitioned into *k* stages, with items administered from the *k*th subpool during the *k*th stage.

(iii) At each stage, two items are chosen from the unadministered items of the corresponding level by matching the difficulty parameters with the currently estimated trait of the examinee. The item with difficulty closest

to the trait estimate is administered if a random number from U(0, 1) is less than 0.5, otherwise the item with second closest difficulty parameter is administered. The test moves to the next stage when the specified number of items are administered.

(iv) The process continues until the last group of items has been administered from the last subpool.

Method 2 (SSH): The steps of an Enhanced Stratified Design (an integrated method of STR and SH measure) are as follows:

(i) The target maximum exposure rate is set at, say, 0.2.

(ii) Items are partitioned into *k* subpools as described in Method 1.

(iii) Before the simulation procedure, the exposure parameters of all items are set to one. Simulated adaptive tests as described in STR are then administered to a large group of simulees whose true abilities are randomly sampled from the ability distribution of the real examinee population. The frequency of an item administered is then compared to the target maximum exposure rate. The exposure parameters for those items with administration rate exceeding the target maximum exposure rate are successively adjusted downward by multiplying a factor (e.g. 0.95), whereas the exposure parameters for those items with less exposure rate than permitted are adjusted upward by multiplying another factor (e.g. 1.04). These adjustments continue through thousands of simulation, until the exposure parameters have stabilized and no single exposure rate exceeds the target value.

(iv) Each item carries an exposure control parameter pre-determined by the above step.

(v) At each stage, two items are selected as described in Method 1. The exposure control parameter of the best item is checked against a random number generated from the uniform distribution. If the exposure control parameter is greater than the random number, then the item is administered. Otherwise, the exposure control parameter of the second

best item is checked against with a new uniform random number.

(vi) If both items in (v) are not administered due to the control mechanism, two new items having difficulty parameter closest to the current ability estimate are selected from the rest of the unadministered items at the same level. These two newly selected items have to go through the same control mechanism as described in (v) before being administered. This step continues until an item is administered or no more items are left in that level.

(vii) In case that no item is administered and no more items are left, step (v) and (vi) are repeated until an item is administered.

(viii) Steps (v) to (vii) are repeated until the specified number of items are administered at that stage. Then the test moves to the next stage.

(ix) Steps (iv) to (viii) are repeated until the last group of items is administered at the last stage.

Method 3 (FSH): Maximum Fisher information method incorporated with Sympson & Hetter's exposure control measure. The steps are as follows:

(i) The target maximum exposure rate is set at, say, 0.2.

(ii) Each item is assigned an exposure control parameter pre-determined by series of simulations using maximum information selection method.

(iii) An item is selected if its Fisher information has the maximum value at the currently estimated ability trait.

(iv) The exposure parameter of the selected item is checked against a random number generated from the uniform distribution. If the parameter value is greater than the random number, the item will be administered. Otherwise, the next best item with maximum information will be selected and checked against a new random number. This step continues until an item is administered or the pool is exhausted.

(v) Steps (iii) and (iv) are repeated until the test terminates.

Measures of Performance

Reliability

No matter what kind of item selection design is incorporated into a CAT, it should provide reasonable reliability. Otherwise, the test results should not be used for decisions. Therefore reliability is a criterion for evaluating performance of the three selection methods. In this study, 3000 true abilities were generated and their respective estimates were obtained according to the selection methods employed. Thus reliability coefficient here is interpreted as the squared correlation of the true scores and the observed scores (Allen & Yen, 1979:72). The higher the reliability, the better is the item selection method.

Accuracy is another important criterion for evaluating the performance of a test. This criterion here is interpreted in terms of average bias and mean squared error.

Average bias

Let θ_i, i = 1,…, 3000 be the ability traits of the 3000 examinees and $\hat{\theta}_i$ be the respective estimators as a result of taking the test. Then the average bias is computed as

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

The smaller the average bias, the better is the item selection method.

Mean squared error

Using the same notations of true ability and the estimator, mean squared error is computed as

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

The smaller the mean squared error, the better is the item selection method.

Exposure rate of an item is defined as the ratio of number of times the item is administered to examinees over the total number of examinees. If an item has high

exposure rate, then it has larger risk of being known to prospective examinees, which in turn would cause test validity problem. An item having very small exposure rate means that it is seldom used to estimate the abilities of examinees. Too many items having small exposure rate would mean that the item pool is not well utilized, bringing up the issue of cost effectiveness.

Number of overly exposed items

Since one of the main concerns of this paper is to investigate which of the three methods described earlier perform better in addressing the issue of item exposure control, the number of overly exposed items is certainly one of the key criteria for evaluating the performance. Here if an item has an exposure rate over 0.2, it is classified as overly exposed.

Number of under-utilized items

If too many items are under-utilized, then the item selection procedure cannot make well use of the entire pool. It is also a public concern about resources utilization so that a good method that can curb high item exposure as well as uplift the usage of inactive items is desirable. Here an item is classified as under-utilized if it has an exposure rate below 0.02.

Scaled chi-squared statistic

Chang and Ying (1998) propose that the most desirable exposure rate distribution is uniform for better utilization of item pool. If the pool size is N and test length is L, then the desirable uniform exposure rate is L/N. They introduce a scaled chi-squared to measure the overall efficiency of item pool usage:

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

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

The smaller the χ^2, the better is the utilization of the pool.

Test overlap rate

Test overlap rate is another important summary index in measuring item exposure control. If the test length is N and there are m examinees, the test overlap rate here is computed by steps (1) counting the number of common items for each of the $m(m\text{-}1)/2$ pairs of examinees, (2) adding up all the $m(m\text{-}1)/2$ counts, and (3) dividing the total counts by $Nm(m\text{-}1)/2$. The smaller the overlap rate, the better is the item selection method.

Simulation Design

A number of simulation studies were conducted to investigate the performance of the proposed enhanced stratified design (SSH) and to compare it with the other two item selection and administration methods (STR and FSH). For all studies, simulated tests were administered to a sample of 3000 simulees with abilities randomly generated from N(0,1). These generated abilities are referred as true θ in later discussions.

Study 1

In the first study, a simplified situation was simulated in which the stratified designs had four strata each and all guessing parameters were set to zero and the a-parameters of items were assigned with 0.5, 1.0, 1.5 and 2.0 according to the stratum where they belonged to.

Item pool structure: The pool consisted of 400 items and was partitioned into 4 strata. The first stratum contained the first 100 items with the values of a-parameters all set to 0.5. The second stratum contained another 100 items with a-parameters set to 1.0. Similarly, the third and the fourth strata each contained 100 items with a- parameters set to 1.5 and 2.0 respectively. The difficulty parameters were randomly generated from N(0,1).

Ability distribution: As mentioned earlier, there were 3000 simulees with abilities randomly generated from N(0,1).

Test length (two cases): The test lengths for the first and second cases were set at 40 and 60 respectively.

Item selection method: Three item selection methods were tested, namely, (1) a-stratified method (STR), (2) a-stratified method combined with Sympson-Hetter probabilistic procedure (SSH), and (3) the Fisher information method combined with Sympson-Hetter probabilistic procedure (FSH).

Ability estimation method: The maximum likelihood method was used to estimate θ.

Study 2

Only the item pool and the test length of Study 2 differed from that of Study 1. Instead of using simulated item parameters, this study used operational parameters of 252 items from a 1992 NAEP main assessment sample (Johnson & Carlson, 1994). The test length was set at 24. Item selection methods and θ estimation procedures were the same as those in Study 1.

Findings

Study 1

The results of the first simulation study are summarized in Table 1 and Table 2. In both cases, the reliabilities for the three selection methods are comparable and quite high. Since a test length of 40 is probably long enough, there is not much gain in reliability when the test length increases to 60. The average biases and mean squared errors are all small and comparable. The enhanced stratified design (SSH) performs better than the simple stratified design in the way that it can further cut down the number of overly exposed items from 10 to 0 when test length equals to 40 and from 67 to 11 when test length equals 60. The SSH appears to be the best among the three methods in controlling item exposure. On the other hand, the FSH appears to be very inadequate whilst STR and SSH are promising in utilizing the entire pool in terms of scaled chi-square and the number of under-utilized items. Furthermore, the high test overlap rates of FSH found are consistent with the findings of Parshall, Davey and Nering (1998). On the contrary, the SSH yields the smallest overlap rates in both cases.

Table 1: Summary for Case 1 of Study 1; Test length = 40; Pool size = 400; Number

of Examinees = 3000.

	Stratified	*Stratified with SH*	*Fisher with SH*
Reliability	0.978	0.980	0.984
Average Bias	-0.0023	-0.00151	-0.00433
Mean Squared Error	0.0227	0.02172	0.0193
Scaled Chi-square	6.589	6.084	34.69
# of items with exposure rate ≦2%	1	1	166
# of items with exposure rate >20%	10	0	14
Overlap rate	0.116	0.114	0.178

Table 2: Summary for Case 2 of Study 1; Test length = 60; Pool size = 400; Number of Examinees = 3000.

	Stratified	*Stratified with SH*	*Fisher with SH*
Reliability	0.986	0.984	0.984
Average Bias	-0.00069	0.0022	-0.0023
Mean Squared Error	0.0151	0.0162	0.0197
Scaled Chi-square	6.337	4.284	15.747
# of items with exposure rate ≦2%	0	0	65
# of items with exposure rate >20%	67	11	27
Overlap rate	0.168	0.160	0.181

The scatterplots in Figures 1 and 2 show the relationship between the 3000 estimated abilities and their corresponding true values. Visually, all three methods provide reasonably good estimation for the true ability, and this is consistent with the reliabilities found in both cases.

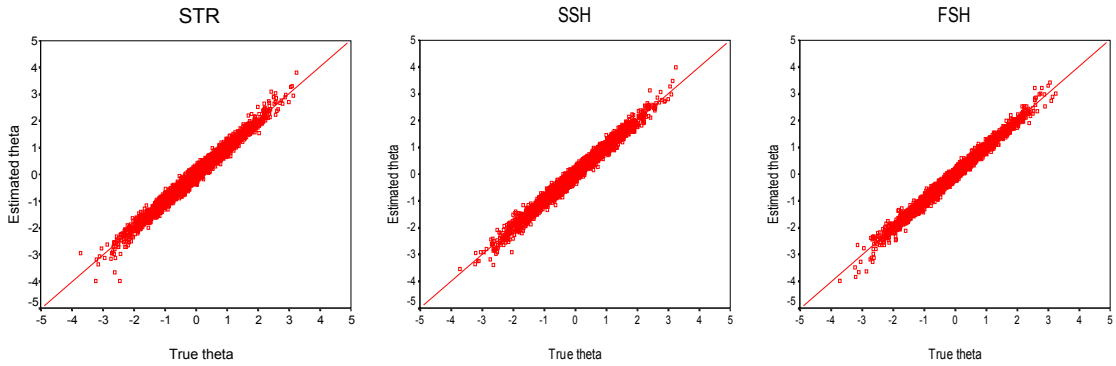

Figure 1. Plots of True Theta vs Estimated Theta for the three methods. Simulated Pool Size = 400, Test Length = 40, # of Simulated Examinees = 3000.

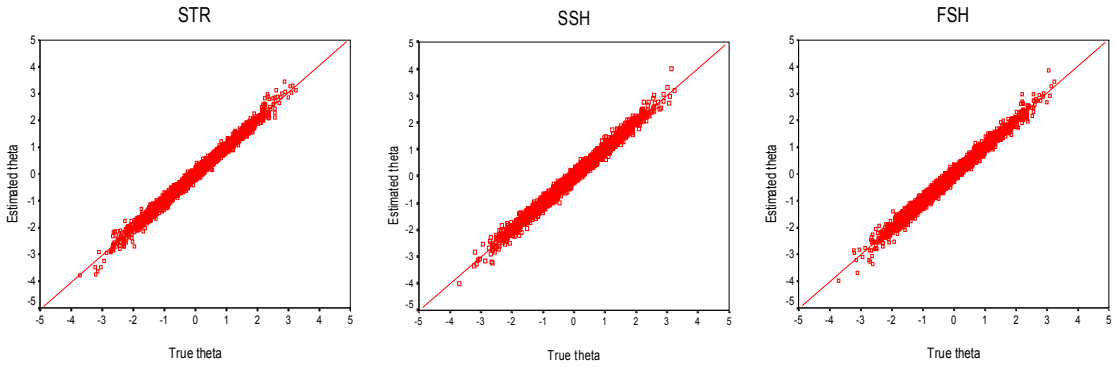

Figure 2. Plots of True Theta vs Estimated Theta for the three methods. Simulated Pool Size = 400, Test Length = 60, # of Simulated Examinees = 3000.

Figure 3 and Figure 4 are the plots of item exposure rates of individual items numbered in the ascending order of a-parameter. It is very clear that FSH method administers items with high a-parameters more frequently but under-utilizes those with low a-parameters, yielding a very uneven item exposure distributions. On the contrary, the STR and SSH methods provide a well-balanced item exposure distributions. Figure 4 also shows that simple STR overly exposes a lot of items when the test length increases from 10% to 15% of the pool size, indicating that control mechanism is needed. In contrast, SSH method performs very well in both item exposure control and pool utilization.

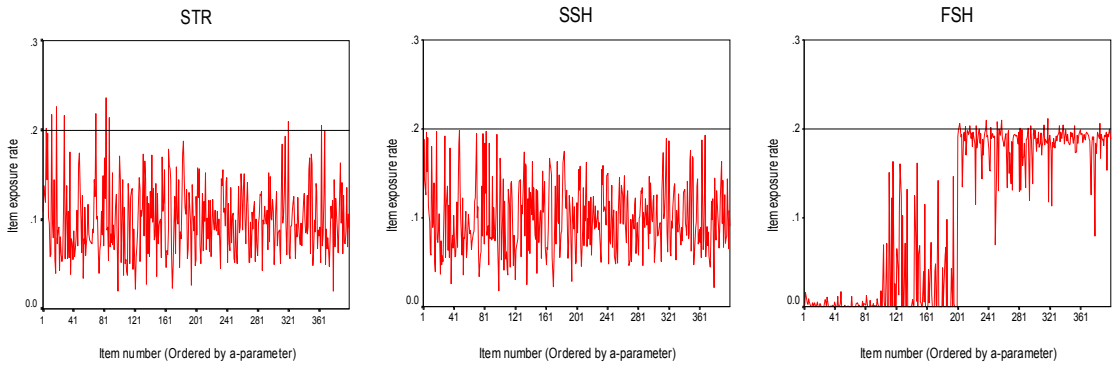

Figure 3. Item exposure rates for 400 simulated items for the three methods. Test Length = 40, # of Simulated Examinees = 3000.

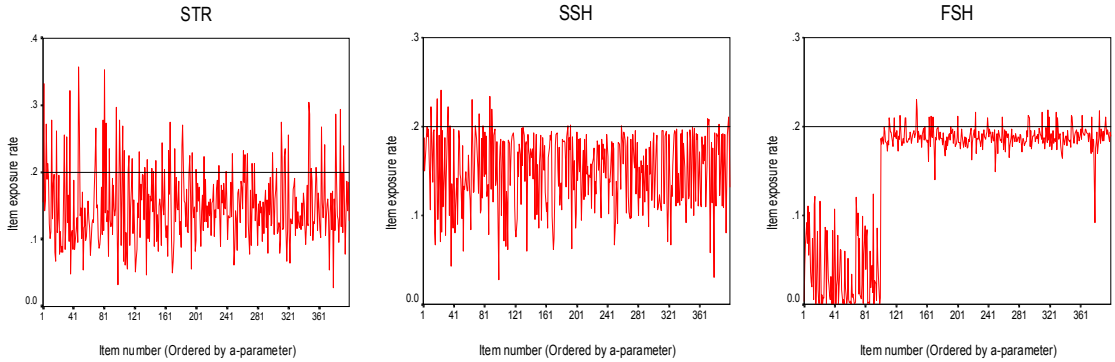

Figure 4. Item exposure rates for 400 simulated items for the three methods. Test Length = 60, # of Simulated Examinees = 3000.

Study 2

The results of the second simulation study are summarized in Table 3. In Study 2, all the reliabilities of the three methods have dropped slightly but the values are still high and comparable. One of the reasons may be the pool size is smaller so that the matching between item difficulties and true abilities is not as good as a larger pool does. The average biases and the mean squared errors are comparable. Both STR and SSH are prominent in utilizing the entire item pool whilst almost two-fifth of the pool is under-utilized with FSH. Moreover, the test overlap rates of STR and SSH are much smaller than that of FSH. The SSH remains the best in item exposure control with no item overly exposed and yields the smallest test overlap rate. Its smallest scaled chi-squared value shows that the item exposure distribution is more even.

Table 3: Summary for Study 2; Test length = 24; Pool size = 252 (operational items); Number of Examinees = 3000.

	Stratified	*Stratified with SH*	*Fisher with SH*
Reliability	0.941	0.939	0.951
Average Bias	0.00813	0.00837	0.00750
Mean Squared Error	0.0672	0.0675	0.0614
Scaled Chi-square	5.362	4.550	24.7
# of items with exposure rate ≦2%	1	1	101
# of items with exposure rate >20%	9	0	16
Overlap rate	0.116	0.113	0.177

Figure 5 shows the scatterplots of the estimates versus their true values. Though the band widths are larger than those of Study 1, the points lie reasonably close to the diagonal of coordinate plane except those near the lower tail. One possible explanation is that there are not enough operational items with difficulties close to the abilities of very weak examinees.

Figure 5. Plots of True Theta vs Estimated Theta for the three methods. Pool Size = 252, Test Length = 24, # of Simulated Examinees = 3000.

Figure 6 reflects the item exposure distributions for the three methods, showing that the item pool is well utilized by SSH and STR. SSH out-performs the STR by reducing the number of overly exposed items to zero.

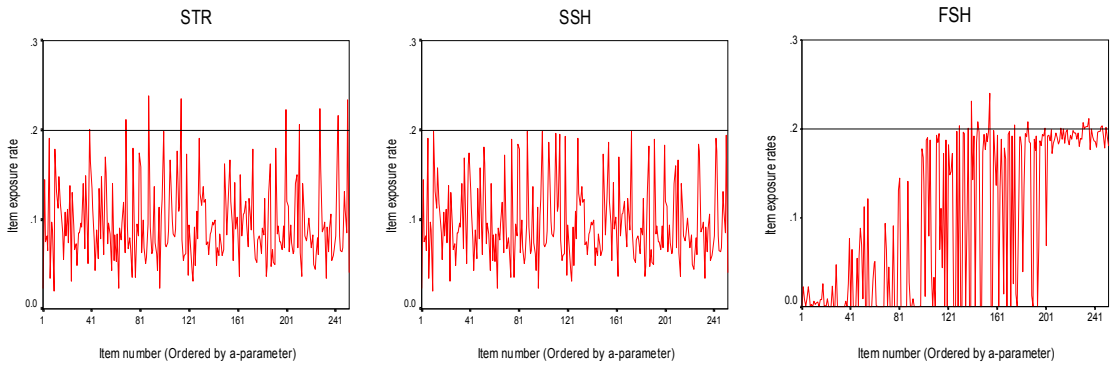

Figure 6. Item exposure rates for 252 operational items for the three methods. Test Length = 24, # of Simulated Examinees = 3000.

Discussion

The enhanced stratified design for computerized adaptive design further improves the capacity of its simple form in controlling item exposure rates. In the simulation studies with either simulated items or operational items, the enhanced stratified design allows no items being overly exposed when the test length is not more than 10% of the pool size. It also yields the lowest test overlap rates in all situations here when compared to the other two methods. Thus the stratified design combined with Sympson-Hetter procedure is a promising method to reduce the degree of damage arising from sharing items among examinees before taking the test.

Both the simple and the enhanced stratified design make better utility of the entire item pool. From the results, none or only one item has exposure rate below or equal to 2% by these two designs. On the contrary, as high as two-fifth of the entire item pool are under-utilized by the Fisher-Sympson-Hetter method. In addition, SSH provides the smallest scaled chi-square statistic in all comparison simulations and the ratio of this statistic for FSH and SSH could be as high as 5.7, indicating that the enhanced stratified design yields more even item exposure distributions.

From the results, all the three methods maintain very high and comparable efficiency in terms of reliability, average bias and mean squared error.

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

studies assumed content constraints are not of main concerns. It is worth to test whether the enhanced stratified design also works well in the situations where content balancing is an important issue. Secondly, the simulation tests terminated after a fixed length of items administered. When accuracy of estimation is the criterion for terminating a test, the stratified rules have to be modified. Thirdly, the numbers of strata were set at 4 and the sizes of strata are the same for each study. The effects of varying these quantities need to be investigated in order to search for optimal values for these variables.

References

Allen, M. & Yen, W. (1979). *Introduction to Measurement Theory*. CA: Brooks/Cole Publishing Company.

Chang, H., & Ying, Z. (1996). A Global Information Approach to Computerized Adaptive Testing. *Applied Psychological Measurement*, *20*(3), 213-229.

Chang, H., & Ying, Z. (1997, June). *Multi-stage CAT with stratified design*. Paper presented at the Annual Meeting of Psychometric Society, Goltlinsberg, NT.

Chang, H.H., & Ying, Z. (1998). *A-stratified Multistage Computerized Adaptive Testing*. Paper submitted to the Special CAT Issue of Applied Psychological Measurement.

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.

Johnson, E.G., & Carlson, J.E. (1994). *The NAEP 1992 technical report*. Washington DC: National Center of Education Statistics.

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.

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.

Parshall, C., Davey, T., and Nering, M. (1998). *Test dependent exposure control for adaptive test*. Paper presented at the 1998 Annual Meeting of the National Council on Measurement in Education, San Diego, CA.

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*, as described in Wainer, et al. (1990).

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.