a-Stratified CAT Design with Content-Blocking

Qing Yi

ACT, Inc.

Hua-Hua Chang

University of Texas at Austin

Qing Yi (62)
ACT, Inc.
2255 North Dubuque Road
PO Box 168
Iowa City, IA 52243
U.S.A

Yiq@act.org
Phone: (319) 341-2296
Fax: (319) 339-3020

Abstract

Content balancing is often required in the development and implementation of computerized adaptive tests (CATs). In the current study, we propose a modified *a*-stratified method, the *a*-stratified method with content-blocking. As a further refinement of the *a*-stratified CAT designs (e.g., Chang & Ying, 1999; Chang, Qian, & Ying, 2001), the new method incorporates content specifications into item pool stratification. Simulation studies were conducted to compare the new method with three previous item selection methods: (1) the *a*-stratified method; (2) the *a*-stratified with *b*-blocking method; and (3) the maximum Fisher information method with Sympson-Hetter exposure control (Hetter & Sympson, 1997; Sympson & Hetter, 1985). The results indicated that the refined *a*-stratified design performed well in reducing item overexposure rates, balancing item usage within the pool, and maintaining measurement precision, in a situation where all four procedures were forced to balance content.

Suggested running head: *a*-stratified with content-blocking

a-Stratified CAT Design with Content-Blocking

In computerized adaptive testing (CAT), items are selected sequentially to be administered to examinees based on certain selection algorithms. A traditional method is to select an item that maximizes Fisher information at a current estimated ability ($\hat{\theta}$) level (Lord, 1977). The theoretical basis for doing this is that it increases measurement precision, however, this often leads to unbalanced item usage within a pool. Highly discriminating items tend to be administered more frequently, which may cause these items to be overexposed. Less discriminating items, on the other hand, may rarely be selected and consequently these items may not be administered enough or even at all. Such uneven item usage poses test security and economic concerns for testing programs (e.g., Davey & Nering, 1998; Revuelta & Ponsoda, 1998; Stocking & Swanson, 1993; Weiss, 1985).

Ideally, all items in a pool should have similar exposure rates to meet the requirements of test security and efficiency of item usage as well (Hua & Chang, 2001; Mills & Stocking, 1996). Once items are developed, they have to satisfy particular specifications in order to form the required substantive and statistical distributions. It is relatively easy for item writers to control certain item characteristics such as content and difficulty level; but it is extremely challenging to produce only highly discriminating items. Once items are included in the pool, they have already undergone certain rigorous review processes and shown no problems. Any item that does not meet psychometric requirements should have been discarded; thus lower discriminating does not represent poor quality.

Chang and Ying (1999) have demonstrated that using lower *a* items at the beginning of CAT will not sacrifice measurement precision. Actually, the advantage of using a high *a* item may not be realized if the examinee's θ is not close to the *b* of that item. Therefore, instead of high *a*, items with lower *a*-values should be used when there is little known about θ (as in the early stages of a test). In practice, however, most popular item exposure control procedures have failed to yield more balanced item usage within a pool. Wainer (2000) examined the item usage within the GRE CAT pools and found that as few as 12% of the available items can account for as much as 50% of the functional pool (those items were actually administered). Obviously, one way to resolve this problem is to decrease higher *a* items' exposure rates while increase the usage of lower *a* items.

Item Exposure Control Method

Three types of item exposure control methods have been developed in CAT: (1) randomization procedures (McBride & Martin, 1983); (2) probabilistic approaches (Davey & Parshall, 1995; Hetter & Sympson, 1997; Stocking & Lewis, 1998; Sympson & Hetter, 1985; Thomason, 1995); and (3) *a*-stratified methods (Chang & Ying, 1999; Chang, Qian, & Ying, 2001). The first two approaches are designed to control high *a* items' exposure rates, while the focus of the third method is to obtain more balanced item usage within a pool.

Randomization Method

The randomization procedure first selects a number of "most informative" items and then randomly selects one item from them (McBride & Martin, 1983). By doing so, the most informative item at a current $\hat{\theta}$ may not always be administered. However, this method only provides limited protection to frequently selected items and will not increase the usage for those less frequently selected items (Davey & Nering, 1998).

Probabilistic Method

The Sympson-Hetter procedure (SH; Hetter & Sympson, 1997; Sympson & Hetter, 1985) uses item exposure control parameters to probabilistically control the frequencies with which the selected items are administered. This technique limits the maximum rate at which an item is administered to examinees. The idea underlying the SH method is that not every selected item has the same probability of being administered to an examinee. For frequently selected items, exposure control parameters can be set at a pre-specified maximum exposure rate; thus, upon selection those items cannot be freely administered to an examinee. The SH method is implemented in two stages: In the first stage, SH item exposure control parameters are computed through a series of simulated CATs administered to a target population. In the second stage, the resulting exposure control parameters are then used in a CAT to control the frequency with which items are administered. To accomplish this, the selected item's exposure control parameter is compared with a uniform random number. If the random number is less than or equal to the exposure control parameter, the selected item will be administered; otherwise, this item is set aside and the next optimal item is selected. A number of modifications of the SH method have been developed (e.g., Davey & Parshall, 1995; Nering, Davey, & Thompson, 1998; Stocking & Lewis, 1998). This probabilistic approach can effectively control the maximum exposure rate of an item. However, there are at least two limitations: First, items that are not selected cannot be administered; therefore, items that have low probabilities of being selected will still have low exposure rates. Second, exposure control parameters have to be updated through a large number of complicated simulations each time an item pool is changed or if the distribution of abilities of a targeted population changes.

***a*-Stratified Method**

The *a*-stratified methods (Chang & Ying, 1999; Chang et al., 2001) take a different approach. No simulations are needed to obtain item exposure control parameters. An item pool is stratified into several strata based on the values of item parameters, and the test is then divided into several corresponding stages. In the early stages of a test, items with lower *a*s are administered and during the later stages items with higher *a*s are administered; thus, this may result in more balanced item usage within a pool. The idea underlying the stratification approach is that early in a test little is known about an examinee's ability, therefore, it is more appropriate to use lower *a* items at the beginning of the test and save higher *a* items for later stages. As indicated by Chang and Ying (1996), the advantage of using a higher *a* item may not be realized if an examinee's true ability is deviate from his/her estimated ability, as is often the case early in a CAT. Less discriminating items, on the other hand, may be a better choice for the examinee.

STR_A

The original *a*-stratified method (Chang & Ying, 1999), denoted STR_A, stratifies an

item pool strictly based on an ascending order of the *a*-parameters. Research has shown that STR_A outperforms the maximum Fisher information method with SH exposure control (MSH) in that it increases the usage of items with lower *a* values while maintaining measurement precision (Chang & Ying, 1999; Hau & Chang, 2001; Leung, Chang, & Hau, 1999). As emphasized by Chang and Ying (1999), a crucial requirement for STR_A to perform well is that the *a*- and *b*-parameters are not correlated. However, for operational item pools, *a*- and *b*-parameters are often positively correlated (Lord, 1975). If the range of *b*-parameters is not wide enough to match examinees' ability distribution within each stratum, it is likely that some items are overly selected by STR_A (Ban, Wang, & Yi, 1999; Parshall, Hogarty, & Kromrey, 1999).

STR_B

The item overexposure problem observed in STR_A (e.g., Ban et al.; Parshall et al.) is mainly caused by the correlation between *a*- and *b*-parameters. When *a*s and *b*s are positively correlated, items with high *a* and low *b* values, for strata corresponding to the later stages, are scarce. The shortage of such items may lead them to become overexposed. This problem may be mitigated if an item pool is partitioned first based on *b*-parameters and then on *a*-parameters so that across strata the distribution of *b* closely matches that of the pool (Chang et al., 2001).

The *a*-stratified with *b*-blocking method, denoted STR_B, can be considered a combination of STR_A and the *b*-stratification approach proposed by Weiss (1973). An item pool is stratified twice: The first stratification is based on *b*-parameters, while the second is based on *a*-parameters. More specifically, an item pool is divided into blocks in an ascending order of the *b*-parameters. Within each block, items are sorted based on the *a*-parameters, from smallest to largest. Then, across all the blocks, items with the lowest *a*s are assigned to the first stratum, the second lowest *a* items to the second stratum, and eventually the highest *a* items to the last stratum. Finally, from each block the first stratum is combined to a single stratum one, and so forth. Now the resulting stratified pool has two properties: (1) the distribution of *b*-parameters in each stratum resembles that of the total pool and (2) the average value of *a*-parameters increases across strata. A simulation study has shown that STR_B outperforms STR_A in reducing item overexposure rates, increasing item usage within the pool, while improving measurement precision (Chang et al., 2001).

Content Balancing

In a real world CAT application, the relationship between statistical and content properties of items must be considered simultaneously (Stocking, 1998). Items in one content area may tend to be more difficult than items in another content area. Content balancing is required to obtain relatively comparable test scores among examinees in operational CATs (Stocking & Swanson, 1993). Clearly, the *a*-stratified methods should be fine-tuned to incorporate the content specifications of an item pool into the stratification and to make the idea of stratification more suitable for practical applications.

Content Specifications and Item Pool Stratification

In the current research, we propose the *a*-stratified method with content-blocking (STR_C) that takes both the content specifications and the relationship between the *a*- and *b*-

parameters into consideration during item pool stratification. Figure 1 illustrates an ideal situation in which an item pool can be stratified based on both content specifications and the values of item parameters. The item pool is first stratified into groups according to the content specifications, then the STR_B procedure is used to obtain all the strata within each content group. Finally, all items with the same stratum number are pooled across all the content groups to form the final strata. Specific steps of the STR_C procedure are described in the Appendix.

Insert Figure 1 About Here

Now, the resulting pool has three characteristics: (1) the content coverage of each stratum is similar to that of the full item pool; (2) the distribution of the *b*-parameters in each stratum is as similar as possible to that of the full item pool; and (3) the average value of the *a*-parameters increases across strata. Similar to its predecessors, STR_C administers items with relatively lower *a*s in the early stages of a CAT, and administers items with relatively higher *a*s during later stages. Within each stratum, items with *b* values closest to $\hat{\theta}$ are selected. Note that if there is only one content area, STR_C is equivalent to STR_B.

Content Balancing

In order to cover approximately similar content areas in each adaptive test, certain content balancing control mechanisms need to be incorporated into CAT item selection algorithms. Kingsbury and Zara (1989) proposed a method called constrained CAT (C-CAT) in which an item is selected from the content area that has the largest discrepancy from a pre-specified proportion of items for each content area. C-CAT has been used in previous research (e.g., Chang, Ansley, & Lin, 2000). However, Chen and Ankenmann (in press) indicated that this method might yield a highly predictable sequence of content areas, especially in the early stage of a CAT. They then proposed using a modified multinomial model to select a target percentage of items from each content area. Recently, after comparing several content balancing control methods, Leung, Chang, and Hau (2001) found that the procedure proposed by Chen and Ankenmann (in press) yielded much more balanced item usage within a pool.

Method

We conducted simulation studies to compare STR_C with the following three methods: (1) STR_A, (2) STR_B, and (3) MSH. The performance of each method was evaluated in terms of item exposure control, item usage within the pool, and measurement precision in content balanced CATs. The effect of the number of strata was also examined. The procedure that was based on a modified multinomial model to balance content was used for all four procedures.

Item Pool

The item pool consists of 539 items from nine forms of a large-scale math achievement test. The item parameters were calibrated based on the data from about 3000 examinees for each form. The three-parameter logistic (3-PL) item response theory (IRT) model was assumed and the BILOG computer program (Mislevy & Bock, 1990) was used. A total of 540 items are

included on the test, however, one item was excluded because its *a*-parameter value was 0.27 and its *b*-parameter value was 5.54. The *a*- and *b*-parameter values of the 539 items are positively correlated with a correlation coefficient of 0.48. This math test covers six content areas. Four scores, a total score and three subscores, are reported in the test. The content areas that make up the three subscores, denoted Content Area One, Content Area Two, and Content Area Three, were used for content consideration in the current research.

Item Pool Stratification

The item pool was stratified into four and six strata for each of the stratified methods. For the four strata, each of the first three strata consists of 135 items, and the fourth stratum has 134 items. For the six strata, each of the first five strata has 90 items, and the last stratum consists of 89 items.

STR_A

The first stratum consists of items with the smallest *a*s, the second with the second smallest *a*s, the third with the third smallest *a*s, while the last stratum has items with the largest *a*-parameters.

STR_B

For the four strata, the item pool was divided into 135 blocks according to the *b*-values. Each block, except the last that consists of three items, has four items. The items were arranged in an ascending order so that the first block contains items with the smallest *b*-values and the last block includes the largest *b* items. Each of the 135 blocks was then partitioned into four strata according to the *a*-parameters. Thus, in each block, the first stratum contains an item with the smallest *a* value and the fourth stratum contains an item with the largest *a* value. The stratification procedure is essentially the same as that of STR_A except that it is performed within a *b*-block. Finally, across the 135 blocks all items in the k^{th} stratum were pooled to form a single stratum, thus obtaining four strata. Similarly, the item pool was stratified to six strata. It is clear that each stratum covers roughly the same range of *b*-values, and the average value of the *a*-parameters increases across the strata. As indicated by Chang et al (2001), when the *a*s and *b*s are not correlated, the characteristics of the resulting stratified pool of the STR_B method should be very similar to that of STR_A.

STR_C

As a further refinement of STR_B, the item pool was first divided into three groups based on the content areas. The first group contains 215 items from Content Area One, and the second and third groups have 162 items from Content Areas Two and Three, respectively. Within each content group, following the steps of STR_B as described previously, we obtained four and six strata, respectively. Next, we pooled all strata ones from each content group to form a single stratum one, all strata twos into a single stratum two, all strata threes into a single stratum three and so on. The resulting item pool is thus partitioned into four and six strata such that each stratum covers roughly the same content distribution as well as the similar range of *b*-values while the average value of *a* increases across the strata.

Descriptive statistics of item parameters

Tables 1 and 2 include the descriptive statistics for the *a*- and *b*-parameter values of the item pool, and the three stratified methods across the four and six strata, respectively. The STR_B and STR_C methods lead to similar item parameter values across the strata, as indicated by the descriptive statistics. Tables 3 and 4 contain the content coverage of the item pool, and the three stratified methods across strata. Clearly, for STR_C the content coverage in every stratum was similar to that of the full item pool, compared to the content coverage of both the STR_A and STR_B methods.

Insert Tables 1, 2, 3, and 4 About Here

CAT Simulation

Ten thousand θ values were generated from a standard normal distribution. For each simulee, a fixed length CAT of 40 items was simulated.

Content balancing control

A content balancing control procedure that uses a modified multinomial model was implemented in each of the four CAT methods. As implemented, the content balancing algorithm assured that each CAT consisted of about 0.40 Content Area One items (16 items) and 0.30 each of Content Areas Two and Three items (12 items). The targeted proportions sum up to one and are treated as the parameters of a multinomial distribution. Sampling from this distribution, a corresponding content area was located. Items were then selected from this content area based on the item selection criteria. The parameters of the multinomial distribution were modified when a target proportion was achieved for a given content area; the unmatched proportions assigned to the rest of the content areas were divided by their sum to form a new multinomial distribution. This process continued until the pre-specified proportions of items from each content area were selected.

Item selection

For each of the three stratified methods, the first item was randomly selected from the first stratum. As to MSH, the first item was randomly selected from the pool. The rest of the 39 items were selected according to the item selection criteria endorsed by each of the four methods. For the three stratified methods, two items were selected each time based on the same criteria, and then one was randomly chosen. For MSH, the next item was selected based on the maximum Fisher information at the current $\hat{\theta}$.

Ability estimation method

The expected a posterior (EAP) method was used to estimate ability initially until one correct and one incorrect item response were obtained, and five items had been administered. Afterwards, maximum likelihood estimation (MLE) was used.

Item exposure control parameters

The maximum exposure rate for MSH was set at 0.20. The MSH item exposure control parameters were obtained through a series of simulated CATs administered to those 10,000 simulees.

Evaluation Criteria

The performance of each method was evaluated using criteria similar to those in Chang and Ying (1999). These criteria evaluate item exposure control, item pool usage, and measurement precision in the simulated CATs.

Item exposure control and pool usage

The frequency of items falling into various ranges of exposure rates and descriptive information of exposure rates were summarized. The χ^2 index, a measure to quantify the equalization of item exposure rates, is computed as

$$\chi^2 = \frac{\sum_{i=1}^{N}(r_i - L/N)^2}{L/N}, \tag{1}$$

where N represents the item pool size, L denotes the test length, m is the number of examinees, and

$$r_i = \frac{\text{number of times the } i^{th} \text{ item is used}}{m}. \tag{2}$$

Note that L/N denotes a desirable uniform rate for all items, and Equation (2) represents the observed item exposure rate.

Both overall and conditional test overlap rates were computed. For the overall test overlap rate, the total number of common items administered to each of the $m(m-1)/2$ pairs of examinees was counted, where m is the number of examinees. This total count was then divided by $L[\,m(m-1)/2]$, where L is the test length. As to the conditional test overlap rate, examinees were assigned to five homogenous ability groups, and the above procedure was performed within each group.

Measurement precision

Bias and mean squared error (MSE) estimates are:

$$\text{Bias} = \frac{1}{m}\sum_{m=1}^{m}(\hat{\theta}_m - \theta_m) \tag{3}$$

and

$$\text{MSE} = \frac{1}{m}\sum_{m=1}^{m}(\hat{\theta}_m - \theta_m)^2 , \quad (4)$$

where $\hat{\theta}_m$ is the m^{th} examinee's estimated ability, θ_m is his/her true ability. In addition to calculating an overall bias and MSE, conditional measurement precision was also obtained at 13 equally spaced θ points from –3 to 3 in increments of 0.5 with 3,000 replications at each of the θ points. The correlation coefficient between $\hat{\theta}_m$ and θ_m ($\rho_{\hat{\theta}_m,\theta_m}$) was also calculated.

Results

Item Exposure Rate and Pool Usage

Table 5 presents the frequency of items with various exposure rates, descriptive statistics of exposure rates, and item pool usage for the four CAT methods across the four strata. None of the methods had any items that were never used. However, MSH had 314 items with exposure rates between 0 and 0.05, in which 284 items had an exposure rate ≤ 0.02. STR_A had 260 items with exposure rates between 0 and 0.05, STR_B had 187, and STR_C had 164 such items (only 5 items with an exposure rate ≤ 0.02). For both the STR_B and STR_C methods, most of the items (251 items) had exposure rates between 0.05 and 0.10. STR_A had 165, and MSH only had 35 items with such exposure rates. STR_A had 30, STR_B had 18, and STR_C had 5 items with exposure rates > 0.20.

MSH had a minimum item exposure rate of 0.001 and a maximum of 0.211. The exposure rates for STR_C ranged from 0.009 to 0.214. The maximum exposure rates for STR_A and STR_B were 0.404 and 0.355, respectively. STR_C resulted in the smallest standard deviation of item exposure rates.

The χ^2 measure ranged from 12.882 (STR_C) to 55.157 (MSH). STR_C generated the lowest overall test overlap rate (9.802%), while MSH had the highest (17.646%). According to the lower bound calculation of overlap rate for fixed length CATs, L/N (Chang & Zhang, in press), the lower bound in this study is $40/539 = 7.4\%$. Thus, with an overall test overlap rate of 9.802%, STR_C performed the best due to it having an overlap rate closest to the theoretical lower bound. The results of the conditional test overlap rates indicated that STR_C generated the lowest while MSH had the highest rates. When the test overlap rates were conditioned on examinees' ability, individuals with extreme ability levels yielded higher rates than that of the group with the median ability range. As expected, the conditional test overlap rates were higher than the overall test overlap rates.

Table 6 shows the same kind of information as Table 5 but for the case with six strata. All the stratified methods performed similarly when the number of strata increased from four to six, except the latter resulted in a larger number of items with an exposure rate ≤ 0.02 and had a slightly higher overall test overlap rates than the former.

Insert Tables 5 and 6 About Here

Figures 2 and 3 provide the item exposure rate distributions for MSH, and the three stratified methods across the four and six strata, respectively. Figures 2c and 3c show that STR_C yielded a relatively balanced item exposure rate distribution. In contrast, the exposure rate distributions for the other three methods were rather uneven (see Figures 2a and 3a, 2b and 3b, and 2d).

~~Insert Figures 2 and 3 About Here~~

Measurement Precision

Tables 5 and 6 also present the overall measurement precision for MSH and for the three stratified methods across strata. For the four strata, all procedures led to reasonably low overall bias that ranged from –0.0026 (MSH) to 0.0061 (STR_A). The MSH procedure led to the lowest overall MSE (0.0509), while STR_C resulted in a lower overall MSE (0.0661) than those of the other two stratified methods (STR_A, 0.0698 and STR_B, 0.0687). The overall bias and MSE obtained from the six strata were higher than those resulted in the four strata. The conditional bias and MSE of each method are presented in Figures 4 and 5 for MSH and for the stratified methods with four and six strata, respectively. The three stratified methods had similar conditional errors, especially the refined stratified methods (i.e., STR_B and STR_C). MSH had the largest conditional errors at the two ends of the θ scale. It is clear that content-blocking of the *a*-stratified method did not reduce the measurement precision, in fact, it actually improved measurement precision for STR_C as compared to STR_A and STR_B when content balancing control was required in the CA~~Ts.~~

Insert Figures 4 and 5 About Here

The three stratified methods resulted in similar correlation coefficient values between the estimated and true ability (See Tables 5 and 6). The correlation coefficient value ranged from 0.9664 (STR_A with six strata) to 0.9696 (STR_C with four strata). MSH, on the other hand, had a slightly higher $\rho_{\hat{\theta}_m,\theta_m} = 0.977$.

Discussion

In this paper, we have introduced a content-blocking method for item pool stratification. The original *a*-stratified method was proposed to make more even and efficient use of items in a pool while maintaining measurement precision. The STR_A method performs well under certain situations, however, it may lead to uneven item usage within a pool when content balancing is required. As a further refinement of the *a*-stratified methods, STR_C takes content specifications along with the characteristics of the *a*- and *b*-parameters into consideration in item pool stratification.

Simulation studies were conducted to examine the effectiveness of STR_C in comparison to STR_A, STR_B, and MSH, respectively. The results indicated that the proposed *a*-stratified method with content-blocking outperformed the other three methods in a situation where content balancing was required. It resulted in more balanced item usage within the pool, particularly

compared to the MSH procedure. There is a general belief that using a CAT necessitates a tradeoff between a balanced use of items and an efficient ability estimate. However, it is interesting to note that the measurement precision of STR_C was increased compared to that of STR_A and STR_B, and was relatively comparable to that of MSH. These results suggest that the proposed content-blocking method is a simple and effective approach to balance content in practical CAT designs.

Additionally, many aspects of the stratified methods need to be studied. First, the current stratified designs have been limited to dichotomous item responses, so a stratified approach to polytomous item responses should be investigated. Second, the effect of combining the 0-1 linear programming (van der Linden, 1998) with the *a*-stratified methods should be explored. This may simplify the procedures in dealing with more complex constraints such as cognitive item types. This may also help answer the questions of "How many strata are needed?" and "What is the effect of the parameters of the multinomial model used for content balancing control on the measurement precision, in the case that the *a*- and *b*-parameters are related to the strata?" Third, the general principle of item pool stratification may also be useful in assembling paper-and-pencil tests. More specifically, additional studies along the line of combining the stratified methods with 0-1 linear programming are needed to explore the possibility of developing automated procedures of creating parallel test forms under multiple constraints. Fourth, previous studies have not addressed the issue of "Whether there is a minimal value for the *a*-parameter of an item in order to be included in the pool?" This issue needs to be investigated in future. Finally, the number of strata in the empirical application in the current study was selected intuitively; it is interesting to note that the performance of the stratified methods did not improve as the number of strata increased from four to six. A more formal discussion in a future study about how to choose a target value for the number of strata would certainly be welcome.

References

Ban, J., Wang, T., & Yi, Q. (1999, June). *Comparison of the a -stratification method, the Sympson-Hetter method, and the matched difficulty method in CAT administration.* Paper presented at the Annual Meeting of the Psychometric Society, Lawrence, KS.

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

Chang, H., & Ying, Z. (1996). A global information approach to computerized adaptive testing. *Applied Psychological Measurement, 20,* 213-229.

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

Chang, H., & Zhang, J. (in press). Hypergeometric family and test overlap rates in computerized adaptive testing. *Psychometrika*.

Chang, S., Ansley, T., & Lin, S. (2000, April). *Performance of item exposure control methods in computerized adaptive testing: Further explorations.* Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans, LA.

Chen, S., & Ankenmann, R. D. (in press). Effects of practical constraints on item selection rules at the early stages of computerized adaptive testing. *Journal of Educational Measurement.*

Davey, T., & Nering, M. (1998, September). *Controlling item exposure and maintaining item security.* Paper presented at the colloquium Computer-Based Testing: Building the Foundation for Future Assessments, Philadelphia, PA.

Davey, T., & Parshall, C. (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, CA.

Hau, K., & Chang, H. (2001). Item selection in computerized adaptive testing: Should more discriminating items be used first? *Journal of Educational Measurement, 38,* 249-266.

Hetter, R. D., & Sympson, J. B. (1997). Item exposure control in CAT-ASVAB. In W. A. Sands, B. K. Waters, & J. R. McBride (Eds.), *Computerized adaptive testing: From inquiry to operation* (pp. 141-144). Washington, DC: American Psychological Association.

Kingsbury, G., & Zara, A. (1989). Procedures for selecting items for computerized adaptive tests. *Applied Measurement in Education, 2(4),* 359-375.

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

Leung, C., Chang, H., & Hau, K. (2001, April). *An examination of item selection rules by stratified CAT designs integrated with content balancing methods.* Paper presented at the Annual Meeting of the American Educational Researcher Association, Seattle, WA.

Lord, F. M. (1975). The 'ability' scale in item characteristic curve theory. *Psychometrika, 40,* 205-217.

Lord, F. M. (1977). A broad-range tailored test of verbal ability. *Applied Psychological Measurement, 1,* 95-100.

McBride, J., & Martin, J. (1983). Reliability and validity of adaptive ability tests in a military setting. In D. J. Weiss (Ed.), *New horizons in testing* (pp.223-226). 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.

Mislevy, R., & Bock, R. (1990). *BILOG 3: Item analysis and test scoring with binary logistic models.* [Computer program]. Chicago, IL: Scientific Software.

Parshall, C., Hogarty, K., & Kromrey, J. (1999, June). *Item exposure in adaptive tests: An empirical investigation of control strategies.* Paper presented at the Annual Meeting of the Psychometric Society, Lawrence, KS.

Nering, M., Davey, T., & Thompson, T. (1998). *A hybrid method for controlling item exposure in computerized adaptive testing.* Paper presented at the Annual Meeting of the Psychometric Society, Urbana, IL.

Revuelta, J., & Ponsoda, V. (1998). A comparison of item exposure control methods in computerized adaptive testing. *Journal of Educational Measurement, 35,* 311-327.

Stocking, M. (1998). *A framework for comparing adaptive test designs.* Unpublished manuscript.

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

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

Sympson, J., & Hetter, R. (1985). Controlling item-exposure rates in computerized adaptive testing. In *Proceeding of the 27th annual meeting of the Military Testing Association* (pp. 973-977). San Diego, CA: Navy Personnel Research and Development Center.

Thomasson, G. (1995, June). *New item exposure control algorithms for computerized adaptive testing.* Paper presented at the Annual Meeting of the Psychometric Society, Minneapolis, MN.

van der Linden, W. (1998). Optimal test assembly of psychological and educational tests. *Applied Psychological Measurement, 22*, 195-211.

Weiss, D. J. (1973). *The stratified adaptive computerized ability test*. Research Report (RR-73-3). Princeton, NJ: Educational Testing Service.

Weiss, D. J. (1985). Adaptive testing by computer. *Journal of Consulting and Clinical Psychology, 53,* 774-789.

Appendix

Specific steps of STR_C are:

1. Select *K*, the number of strata [see Chang and Ying (1999) for a discussion of how to decide on *K*];
2. Divide an item pool into *G* groups based on the content specifications (i.e., one group per content area);
3. Sort items into each of the *G* groups according to the *b*-parameter values, from smallest to largest;
4. Partition items in group *g* ($g = 1, 2, \ldots, G$) into P_g blocks based on the *b*-parameter values; items with the lowest *b* going to the first block and the highest *b* items going to the last block (see below for a detailed discussion of how to determine P_g);
5. Sort the items within each of the P_g blocks in an ascending order of the *a*-parameter values;
6. Assign the sorted items in the P_g blocks into the *K* strata based on *a*-parameter values: the first item in the block (item with the smallest *a*) is assigned to the first stratum, the second item to the second stratum, …, and the last item (with the largest *a*) to the last (K^{th}) stratum;
7. Pool all the strata ones from each group into a single stratum one, all the strata twos into a single stratum two, …, and all the strata *K*s into a single stratum *K*; and
8. Partition the test into *K* stages; in the k^{th} stage, adaptively select n_k items from the k^{th} stratum based on the similarity between *b* and $\hat{\theta}$, then administer the items (note that $n_1 + \ldots + n_k$ equals the test length; and $k = 1, 2, 3, \ldots, K$).

In Step 4, if n_g (i.e., the number of items in the g^{th} group) can be evenly divided by *K* then the number of blocks P_g equals n_g / K and the number of items in each block corresponds to the number of strata *K*. If n_g cannot be evenly divided by *K* then set the number of blocks P_g equal the integer part of $[(n_g / K)] + 1$ and the last block will have $n_{gp} < K$ items, where n_{gp} equals the integer remainder of n_g / K , while the other blocks each have *K* items. There may not be enough items in the last block to allocate an item to each stratum, therefore, effort needs to be made to balance the content coverage among strata when allotting those n_{gp} items. Note that Step 2 partitions an item pool into groups according to the content specifications, Steps 3 to 6 are similar to those of STR_B, and Step 8 is the same as that of STR_A.

TABLE 1

Descriptive Statistics for a- and b-Parameters of the Total Item Pool and the Three Stratified Methods across Four Strata

Parameter		N	Mean	SD	Minimum	Maximum
Item Pool						
	a	539	1.030	0.325	0.276	2.317
	b	539	0.151	1.023	-3.106	3.277
First Stratum						
STR_A	*a*	135	0.653	0.109	0.276	0.781
STR_B	*a*	135	0.751	0.209	0.276	1.441
STR_C	*a*	135	0.764	0.206	0.276	1.358
STR_A	*b*	135	-0.371	1.139	-3.106	3.277
STR_B	*b*	135	0.153	1.027	-2.731	2.428
STR_C	*b*	135	0.145	1.022	-2.731	2.428
Second Stratum						
STR_A	*a*	135	0.892	0.057	0.786	0.996
STR_B	*a*	135	0.933	0.220	0.315	1.467
STR_C	*a*	135	0.941	0.221	0.315	1.480
STR_A	*b*	135	-0.152	0.887	-2.171	2.272
STR_B	*b*	135	0.157	1.044	-2.890	3.277
STR_C	*b*	135	0.158	1.014	-2.890	2.272
Third Stratum						
STR_A	*a*	135	1.100	0.069	0.997	1.230
STR_B	*a*	135	1.107	0.244	0.518	1.783
STR_C	*a*	135	1.100	0.262	0.518	1.783
STR_A	*b*	135	0.288	0.874	-2.038	2.183
STR_B	*b*	135	0.157	1.022	-2.587	2.807
STR_C	*b*	135	0.155	1.031	-2.587	2.807
Fourth Stratum						
STR_A	*a*	134	1.477	0.194	1.230	2.317
STR_B	*a*	134	1.329	0.300	0.742	2.317
STR_C	*a*	134	1.316	0.316	0.663	2.317
STR_A	*b*	134	0.844	0.703	-0.739	2.807
STR_B	*b*	134	0.135	1.010	-3.106	2.264
STR_C	*b*	134	0.146	1.037	-3.106	3.277

TABLE 2

Descriptive Statistics for a- and b-Parameters of the Three Stratified Methods across Six Strata

Parameter		N	Mean	SD	Minimum	Maximum
First Stratum						
STR_A	*a*	90	0.603	0.101	0.276	0.722
STR_B	*a*	90	0.683	0.176	0.276	1.293
STR_C	*a*	90	0.704	0.193	0.276	1.263
STR_A	*b*	90	-0.452	1.179	-2.890	3.277
STR_B	*b*	90	0.153	1.032	-2.731	2.428
STR_C	*b*	90	0.159	1.083	-2.731	3.277
Second Stratum						
STR_A	*a*	90	0.790	0.042	0.724	0.859
STR_B	*a*	90	0.851	0.190	0.315	1.354
STR_C	*a*	90	0.850	0.189	0.315	1.354
STR_A	*b*	90	-0.203	1.001	-3.106	2.271
STR_B	*b*	90	0.162	1.062	-2.890	3.277
STR_C	*b*	90	0.142	1.013	-2.890	2.271
Third Stratum						
STR_A	*a*	90	0.925	0.037	0.862	0.996
STR_B	*a*	90	0.961	0.192	0.421	1.438
STR_C	*a*	90	0.963	0.203	0.421	1.480
STR_A	*b*	90	-0.130	0.850	-2.171	2.032
STR_B	*b*	90	0.153	1.011	-2.567	2.271
STR_C	*b*	90	0.153	1.007	-2.567	2.183
Fourth Stratum						
STR_A	*a*	90	1.059	0.042	0.997	1.129
STR_B	*a*	90	1.081	0.216	0.465	1.530
STR_C	*a*	90	1.074	0.238	0.518	1.709
STR_A	*b*	90	0.205	0.860	-1.801	2.183
STR_B	*b*	90	0.162	1.027	-2.362	2.807
STR_C	*b*	90	0.155	1.043	-2.587	2.807
Fifth Stratum						
STR_A	*a*	90	1.238	0.067	1.136	1.371
STR_B	*a*	90	1.211	0.259	0.518	2.031
STR_C	*a*	90	1.206	0.268	0.566	1.815
STR_A	*b*	90	0.583	0.818	-2.038	2.807
STR_B	*b*	90	0.149	1.022	-2.587	2.264
STR_C	*b*	90	0.158	0.988	-2.114	2.198
Sixth Stratum						
STR_A	*a*	89	1.570	0.174	1.379	2.317
STR_B	*a*	89	1.395	0.307	0.742	2.317
STR_C	*a*	89	1.385	0.306	0.742	2.317
STR_A	*b*	89	0.910	0.684	-0.436	2.264
STR_B	*b*	89	0.126	1.013	-3.106	2.198
STR_C	*b*	89	0.138	1.031	-3.106	2.264

TABLE 3

Percentage of Items in Each Content Area in the Item Pool and the Three Stratified Methods across Four Strata

	Content Area One	Content Area Two	Content Area Three
Item Pool			
	40	30	30
First Stratum			
STR_A	59.3	18.5	22.2
STR_B	45.9	24.4	29.6
STR_C	40.0	29.6	30.4
Second Stratum			
STR_A	44.4	27.4	28.1
STR_B	39.3	33.3	27.4
STR_C	40.0	29.6	30.4
Third Stratum			
STR_A	33.3	34.8	31.9
STR_B	43.0	27.4	29.6
STR_C	40.0	30.4	29.6
Fourth Stratum			
STR_A	22.4	39.6	38.1
STR_B	31.3	35.1	33.6
STR_C	39.6	30.6	29.9

TABLE 4

Percentage of Items in Each Content Area in the Three Stratified Methods across Six Strata

	Content Area One	Content Area Two	Content Area Three
First Stratum			
STR_A	60.0	17.8	22.2
STR_B	57.8	20.0	22.2
STR_C	40.0	30.0	30.0
Second Stratum			
STR_A	52.2	23.3	24.4
STR_B	32.2	33.3	34.4
STR_C	40.0	30.0	30.0
Third Stratum			
STR_A	43.3	27.8	28.9
STR_B	40.0	40.0	20.0
STR_C	40.0	30.0	30.0
Fourth Stratum			
STR_A	37.8	32.2	30.0
STR_B	38.9	31.1	30.0
STR_C	40.0	30.0	30.0
Fifth Stratum			
STR_A	27.8	38.9	33.3
STR_B	33.3	27.8	38.9
STR_C	40.0	30.0	30.0
Sixth Stratum			
STR_A	18.0	40.4	41.6
STR_B	37.1	28.1	34.8
STR_C	39.3	30.3	30.3

TABLE 5

Overall Measurement Precision, Frequency of Items with Various Exposure Rates, Descriptive Information of Exposure Rates, and Item Pool Usage for MSH and the Three Stratified Methods with Four Strata

	Methods			
	STR_A	STR_B	STR_C	MSH
Bias	0.0061	-0.0003	-0.0037	-0.0026
MSE	0.0698	0.0687	0.0661	0.0509
$\rho_{\hat{\theta}_m,\theta_m}$	0.9683	0.9684	0.9696	0.9766
Exposure Rate (EXP)				
EXP = 0.00	0	0	0	0
EXP $\leq$ 0.02	54	15	5	284
0.00 < EXP $\leq$ 0.05	260	187	164	314
0.05 < EXP $\leq$ 0.10	165	251	251	35
0.10 < EXP $\leq$ 0.15	51	52	86	24
0.15 < EXP $\leq$ 0.20	33	31	33	81
EXP > 0.20	30	18	5	85
Mean	0.074	0.074	0.074	0.074
Minimum	0.006	0.015	0.010	0.001
Maximum	0.404	0.355	0.214	0.211
Standard Deviation	0.066	0.051	0.042	0.087
χ^2	31.367	18.698	12.882	55.157
Overall Overlap	13.232%	10.881%	9.802%	17.646%
Overlap ($-3.880 \leq \theta < -0.874$)	40.191%	32.739%	32.792%	44.810%
Overlap ($-0.874 \leq \theta < -0.261$)	25.349%	19.958%	19.062%	35.062%
Overlap ($-0.261 \leq \theta < 0.258$)	18.020%	17.008%	15.800%	33.952%
Overlap ($0.258 \leq \theta < 0.855$)	19.610%	19.321%	17.337%	36.747%
Overlap ($0.855 \leq \theta \leq 4.644$)	26.423%	25.774%	24.047%	47.890%

TABLE 6

Overall Measurement Precision, Frequency of Items with Various Exposure Rates, Descriptive Information of Exposure Rates, and Item Pool Usage for the Three Stratified Methods with Six Strata.

	Methods		
	STR_A	STR_B	STR_C
Bias	0.0022	0.0017	0.0014
MSE	0.0747	0.0694	0.0688
$\rho_{\hat{\theta}_m,\theta_m}$	0.9664	0.9685	0.9681
Exposure Rate (EXP)			
EXP = 0.00	1	1	1
EXP $\leq$ 0.02	74	28	25
0.00 < EXP $\leq$ 0.05	257	213	151
0.05 < EXP $\leq$ 0.10	160	211	266
0.10 < EXP $\leq$ 0.15	59	52	77
0.15 < EXP $\leq$ 0.20	25	39	42
EXP > 0.20	37	23	2
Mean	0.074	0.074	0.074
Minimum	0	0	0
Maximum	0.427	0.327	0.227
Standard Deviation	0.070	0.054	0.043
χ^2	35.464	21.130	13.662
Overall Overlap	13.992%	11.333%	9.947%
Overlap ($-3.880 \leq \theta < -0.874$)	38.302%	31.260%	31.181%
Overlap ($-0.874 \leq \theta < -0.261$)	25.656%	20.473%	18.830%
Overlap ($-0.261 \leq \theta < 0.258$)	19.144%	18.007%	15.360%
Overlap ($0.258 \leq \theta < 0.855$)	20.142%	19.713%	17.302%
Overlap ($0.855 \leq \theta \leq 4.644$)	26.731%	26.727%	24.900%

Figure Captions

FIGURE 1. *Relationships between item parameter values and content specifications*

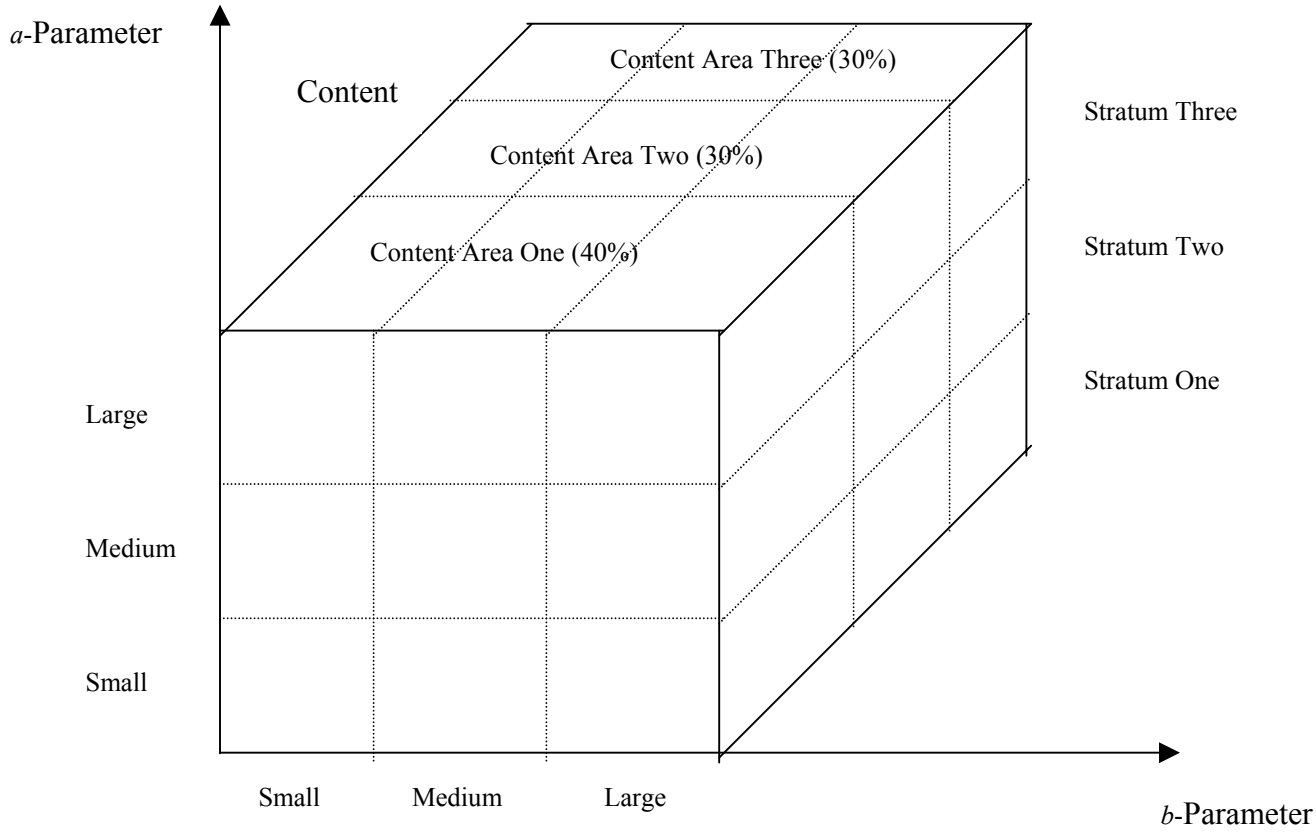
FIGURE 2. *Item exposure rates across MSH and the three stratified methods with four strata (Note: items were in the same sequence as they were in each of the pools)*

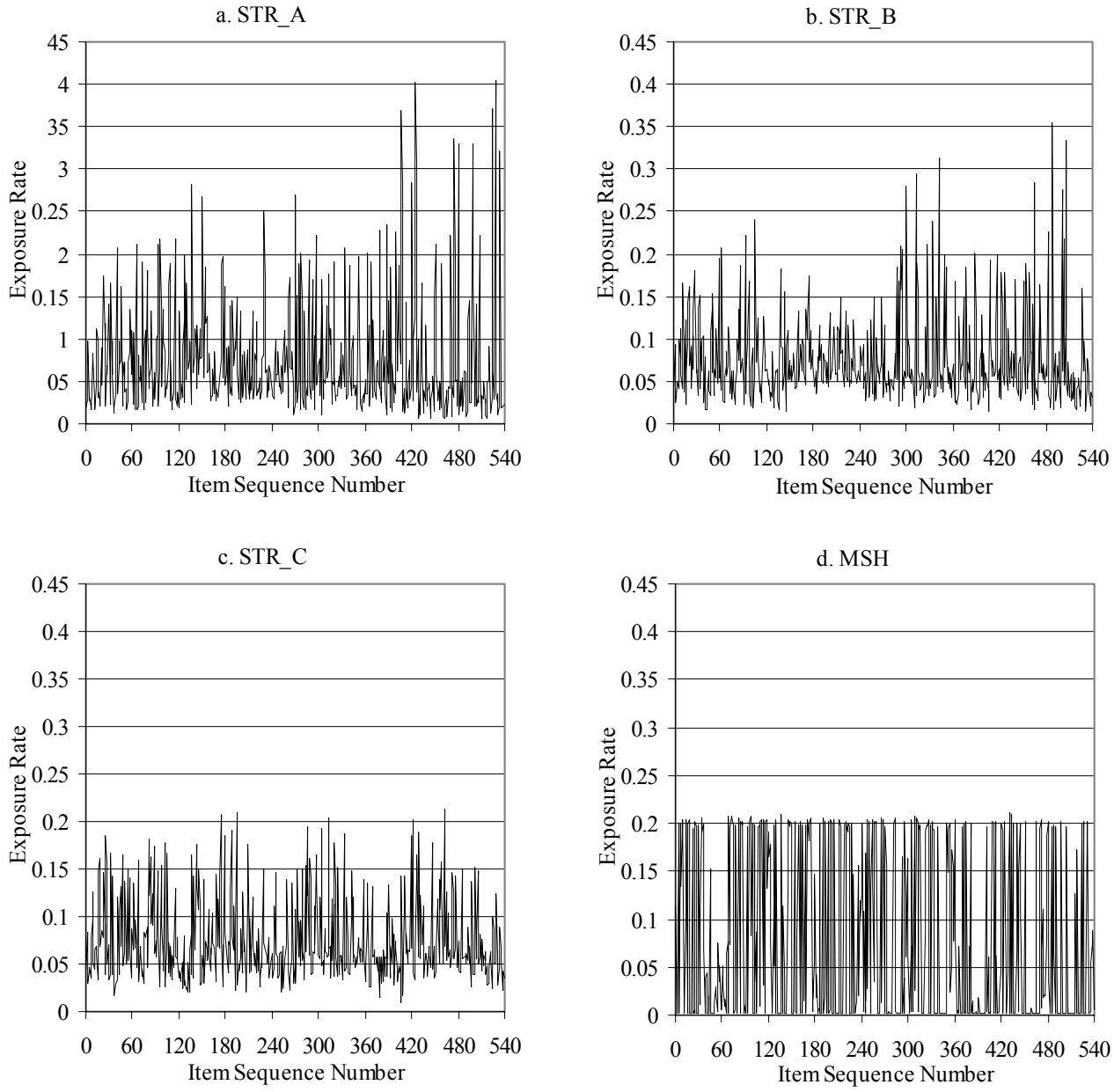
FIGURE 3. *Item exposure rates across the three stratified methods with six strata (Note: items were in the same sequence as they were in each of the pools)*

FIGURE 4. *Bias and MSE conditioned on examinees' ability across MSH and the three stratified methods with four strata*

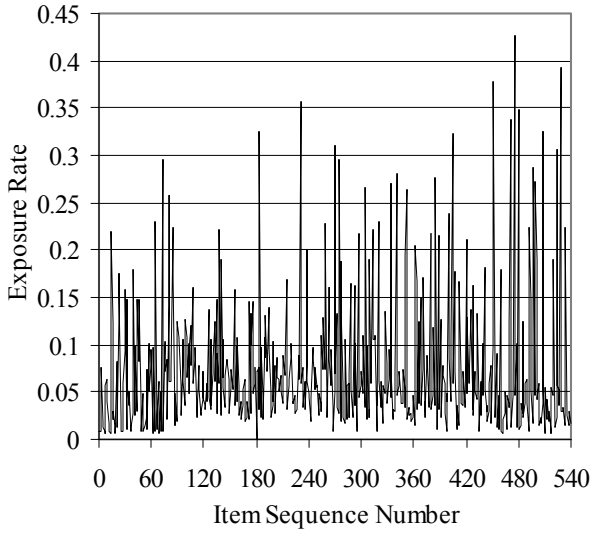
FIGURE 5. *Bias and MSE conditioned on examinees' ability across the three stratified methods with six strata*

a-stratified with content-blocking

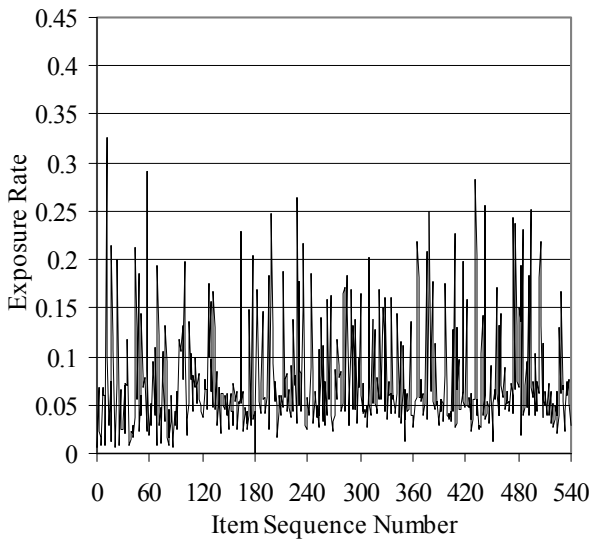



a. STR_A
Exposure Rate
45
4
35
3
0.25
2
0.15
1
05
0
0 60 120 180 240 300 360 420 480 540
Item Sequence Number
b. STR_B
Exposure Rate
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
0 60 120 180 240 300 360 420 480 540
Item Sequence Number
c. STR_C
Exposure Rate
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
0 60 120 180 240 300 360 420 480 540
Item Sequence Number
d. MSH
Exposure Rate
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
0 60 120 180 240 300 360 420 480 540
Item Sequence Number

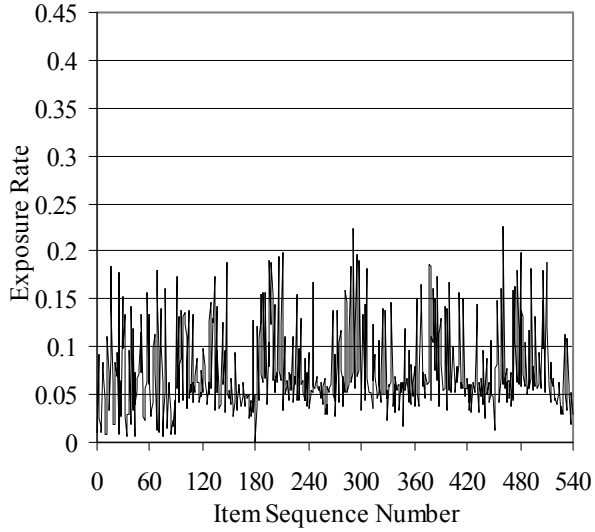
a. STR_A

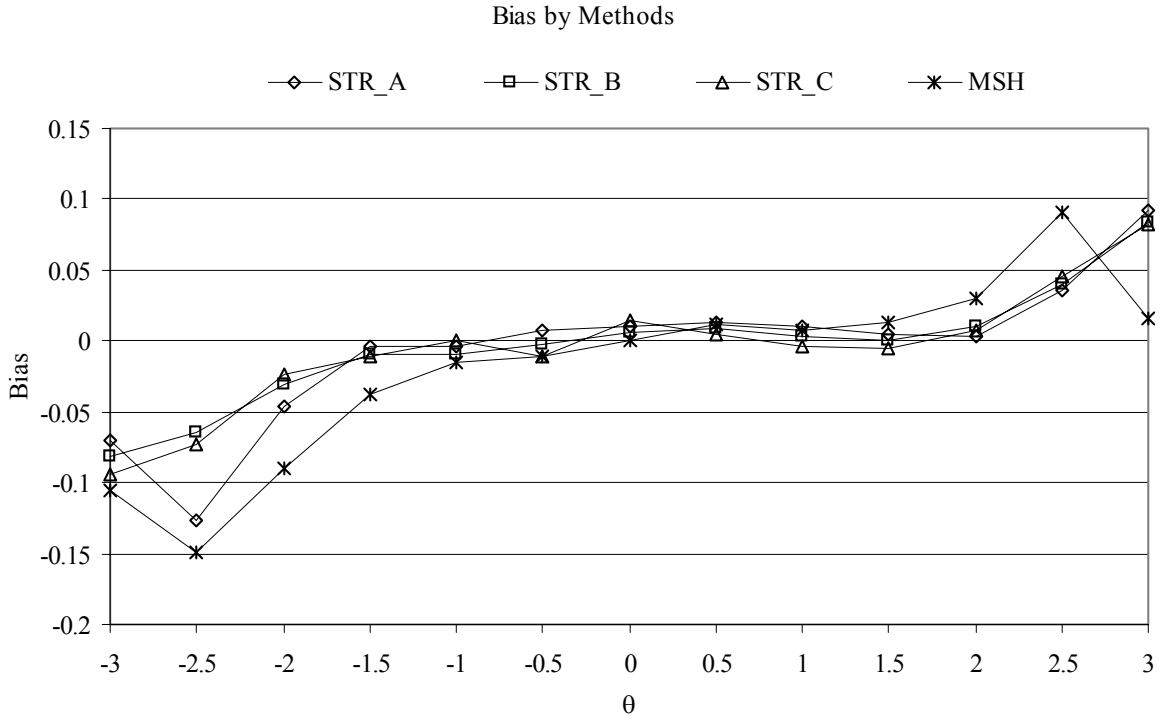


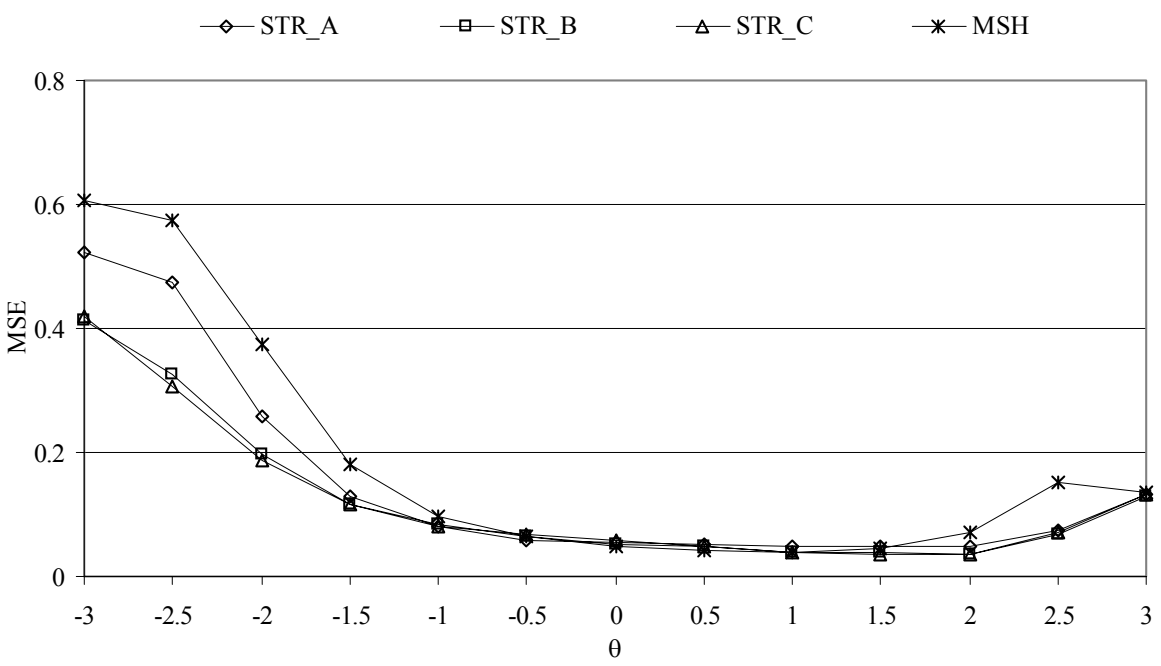
b. STR_B

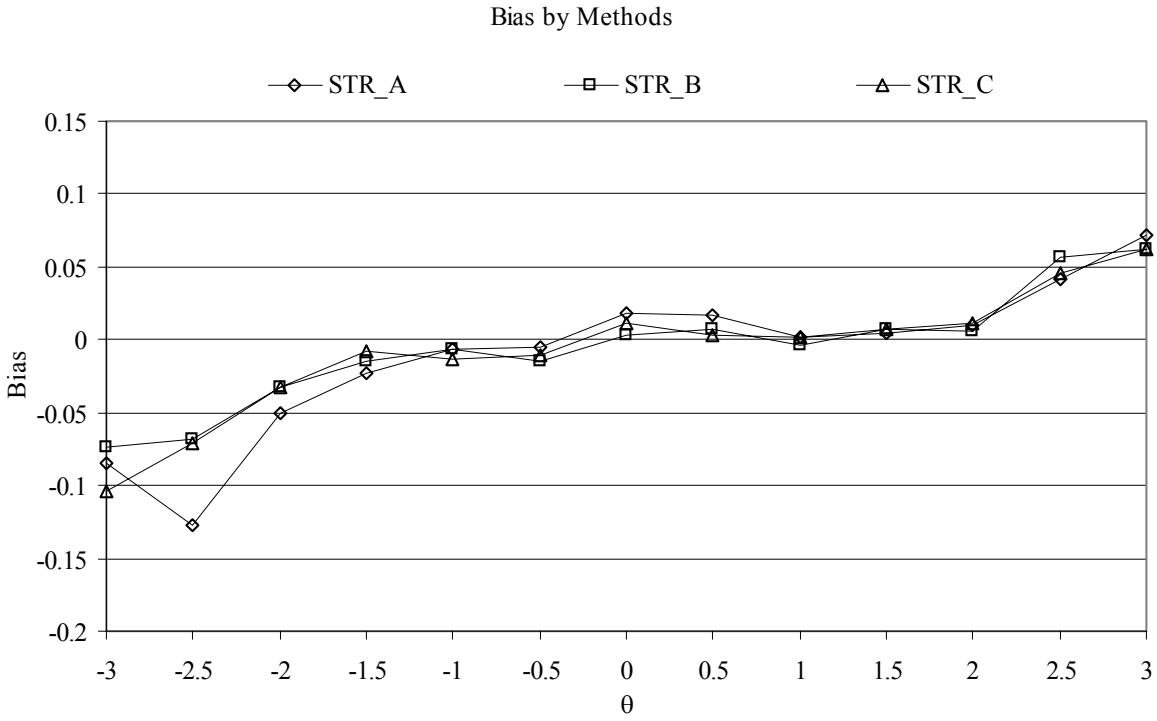


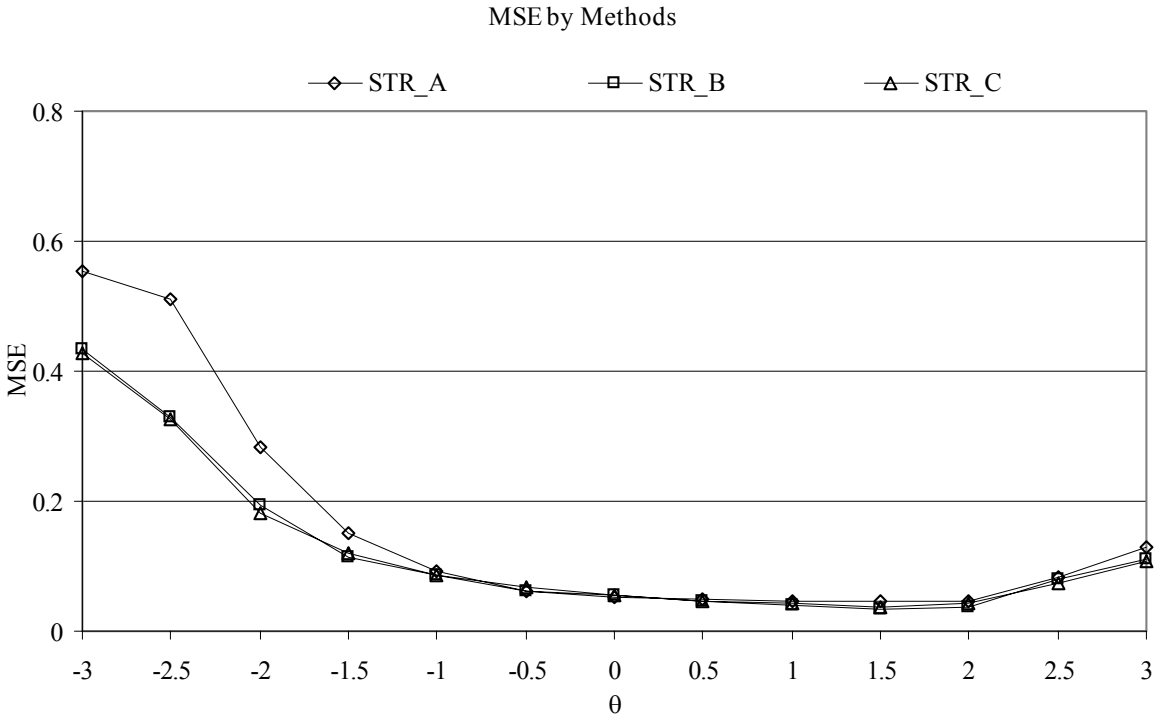
c. STR_C




Bias by Methods
STR_A
STR_B
STR_C
MSH
0.15
0.1
0.05
0
-0.05
-0.1
-0.15
-0.2
Bias
-3
-2.5
-2
-1.5
-1
-0.5
0
0.5
1
1.5
2
2.5
3
θ


MSE by Methods
STR_A
STR_B
STR_C
MSH
0.8
0.6
0.4
0.2
0
MSE
-3
-2.5
-2
-1.5
-1
-0.5
0
0.5
1
1.5
2
2.5
3
θ


Bias by Methods
STR_A
STR_B
STR_C
0.15
0.1
0.05
0
-0.05
-0.1
-0.15
-0.2
Bias
-3
-2.5
-2
-1.5
-1
-0.5
0
0.5
1
1.5
2
2.5
3
θ


MSE by Methods
STR_A
STR_B
STR_C
0.8
0.6
0.4
0.2
0
MSE
-3
-2.5
-2
-1.5
-1
-0.5
0
0.5
1
1.5
2
2.5
3
θ