Incorporating the Sympson-Hetter Exposure Control Method into the *a*-Stratified Method with Content Blocking

Qing Yi

ACT, Inc.

Paper Presented at the Annual Meeting of the American Educational Research Association, April 1-5, 2002, New Orleans, LA.

Abstract

Previous research (e.g., Leung, Chang, & Hau, 2001; Yi & Chang, 2001) has shown that the *a*-stratified method with content blocking (STR_C) performed well in content balancing constrained computerized adaptive tests (CATs). It effectively reduced the exposure rates of highly discriminating items, produced more balanced item usage within a pool, and maintained measurement precision. This method, however, does not have a mechanism that can fix (pre-specify) the maximum item exposure rate. In the current study, the Sympson-Hetter exposure control procedure (Sympson & Hetter, 1985) was incorporated into STR_C (denoted STR_C-SH). Under content balancing constraint conditions, this study investigated the effectiveness of STR_C-SH by comparing the STR_C-SH procedure with the original STR_C method and the maximum information selection method with Sympson-Hetter exposure control (MI-SH). This study also examined the effects of the test length, the number of strata, and the pre-specified maximum item exposure rate on the performance of each method. The results of this study have shown that STR_C-SH is an effective item exposure control method for CATs with practical content constraints.

Incorporating the Sympson-Hetter Exposure Control Method into the *a*-Stratified Method with Content Blocking

Introduction

In computerized adaptive testing (CAT), a traditional method of item selection is to select the most informative item corresponding to an examinee's current ability estimate (Lord, 1980). This item selection method has the advantage of measurement efficiency; however, it tends to cause highly discriminating items to be overexposed and low *a* items not to be administered frequently enough, or at all. Ideally, all items in an item pool should have similar usage (Mills & Stocking, 1996). If items are overexposed, it causes test security concerns; if items are included in the pool, but subsequently not administered frequently enough or at all, this can be inefficient economically. Thus, it is necessary to develop item exposure control methods that not only control highly discriminating items' exposure rates but also increase low *a* items' usage, while still maintaining measurement precision in CATs.

Different methods of item exposure control have been proposed by various researchers (e.g., Chang & Ying, 1999; Chang, Qian, & Ying, 2001; Davey & Parshall, 1995; Stocking & Lewis, 1998; Sympson & Hetter, 1985; Thomason, 1995; Yi & Chang, 2001). The Sympson-Hetter procedure uses exposure control parameters to probabilistically control the frequencies with which items are administered (Sympson & Hetter, 1985). Sympson-Hetter exposure control parameters are obtained through a series of simulated CATs administered to a target population. The *a*-stratified methods (Chang & Ying, 1999; Chang et al., 2001; Yi & Chang, 2001), take a different approach; no simulations are needed to obtain exposure control parameters. Instead, the methods select items from a stratified pool based on the closeness between item difficulty and a current ability estimate. In the *a*-stratified method with content blocking (STR_C) (Yi & Chang, 2001), an item pool is first divided into groups based on the content specifications of the pool. Within each content group, the steps of the *a*-stratified with *b* blocking method as described in Chang et al. (2001) are followed to obtain several strata. That is, in each content group, items are divided into different blocks in ascending order of the *b*-parameters. Within each block, items are sorted based on the *a*-parameters, from small to large. 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. Next, all items with the same stratum number from each content group are pooled to form a single stratum. For example, all stratum one items from each content group are pooled to form a single stratum one, and so forth. The resulting stratified pool has the following three characteristics: (1) the content coverage of each stratum is similar to that of the full item pool; (2) the distribution of *b*-parameters in each stratum is as similar as possible to that of the full item pool; and (3) the average value of *a*-parameters increases across strata. The test is divided into several stages, one per stratum. STR_C then selects items from the corresponding strata based on the match between item difficulty and an examinee's current ability estimate. Items with low *a* values are administered in the early stages of the test and high *a* items are used during the later stages.

Research has shown that STR_C outperformed the two previous *a*-stratified methods and the maximum information selection method with Sympson-Hetter exposure control (MI-SH).

STR_C lowered exposure rates for highly discriminating items while increasing the usage of less discriminating items without a loss in measurement precision (Leung et al., 2001; Yi & Chang, 2001). Although STR_C reduced exposure rates for highly discriminating items, this method does not have a mechanism that sets the maximum item exposure rate at a pre-specified level. Thus, STR_C cannot control the maximum observed item exposure rate at a certain level in a CAT.

In the current study, the Sympson-Hetter exposure control procedure was incorporated into the STR_C method (STR_C-SH); so that the maximum item exposure rate of STR_C-SH can be set at a pre-specified level. The purpose of the study was to investigate the effectiveness of STR_C-SH for a content balanced CAT. The performance of STR_C-SH was compared with the two item selection methods: STR_C and MI-SH. Additionally, the effects of the test length, the number of strata, and the pre-specified maximum item exposure rate on the performance of each method were also examined.

Item Exposure Control Methods

The Sympson-Hetter Method (SH)

The Sympson-Hetter (SH) procedure uses 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 approach 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 item 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, an exposure control parameter $P_i(A \mid S)$ is computed for each item in the pool through a series of simulated CATs administered to a target population. The simulated CATs are conducted by setting the initial exposure control parameter $P_i(A \mid S)$ to 1.0 for all the items in the pool. Based on the item selection criteria, an optimal item is selected. This item is administered if a uniform random number is less than or equal to this item's exposure control parameter. Otherwise, this item is set aside and the next optimal item is identified. This procedure is repeated until an item is found and administered. A new $P_i(A \mid S)$ for each item is then updated,

if $P_i(S) > R$, then new $P_i(A \mid S) = R \,/\, P_i(S)$, or

if $P_i(S) \le R$, then new $P_i(A \mid S) = 1.0$,

where $i = 1, 2, \ldots, N$ represents the items in a pool, $P_i(S)$ is the probability of item *i* being selected, $P_i(A)$ is the probability of this item being administered, $P_i(A \mid S)$ is the exposure control parameter of this item that is the probability of item *i* being administered given it has been selected, and *R* represents the maximum item exposure rate that is pre-specified. The CAT simulations are repeated until $P_i(A)$ approaches the pre-specified value of *R* and the values of

$P_i(A \mid S)$ for all items have been stabilized. The values of $P_i(A \mid S)$, obtained from the final round of iterations, are the exposure control parameters for items in the pool.

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. The SH procedure can be incorporated into different item selection methods to control the maximum observed item exposure rate at a certain level.

The *a*-Stratified Method with Content Blocking (STR_C)

Unlike the SH method, STR_C does not require simulations to obtain exposure control parameters for the items in a pool (Yi & Chang, 2001). Selecting items from a stratified pool based on the match between item difficulty and a current ability estimate reduces the exposure rates of highly discriminating items and increases low *a* items' exposure rates in STR_C.

STR_C is a further refinement of the original *a*-stratified method (Chang & Ying, 1999), which takes both the content specifications and the relationship between *a*- and *b*-parameters into consideration during the item pool stratification. Specific steps of STR_C are:

1. Divide an item pool into *G* groups based on the content specifications (i.e., one group per content specification);

2. Sort items in each of the *G* groups according to *b*-parameters, from small to large. Then, partition items in group *g* ($g = 1, 2, \ldots, G$) into P_g blocks based on *b*-parameters: Items with the lowest *b* going to the first block and the highest *b* items going to the last block;

3. In each of the *G* groups, sort the items within each of the P_g blocks in ascending order of the *a*-parameters. Assign the sorted items in the P_g blocks into the *K* strata based on *a*-parameters: 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 stratum;

4. Combine all the stratum one's from each content group into a single stratum one, all the stratum two's into a single stratum two, …, and all the stratum *K*'s into a single stratum *K*; and

5. 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 those selected items (note that $n_1 + \ldots + n_k$ equals the test length; and $k = 1, 2, 3, \ldots, K$).

Step 1 partitions an item pool into groups according to the content specifications, Steps 2 to 4 are similar to those of the *a*-stratified method with *b* blocking, and Step 5 is the same as that of the

original *a*-stratified method.

Incorporating SH into STR_C (STR_C-SH)

STR_C can effectively reduce the exposure rates of highly discriminating items, however, it does not have a mechanism to control the maximum observed item exposure rate at a certain level. The SH exposure control procedure can be incorporated into STR_C (STR_C-SH) to achieve the goal of limiting the maximum observed item exposure rate at a pre-specified level. STR_C-SH can be implemented in a similar way as STR_C, except the SH procedure is used within each stratum to control the maximum observed item exposure rates.

Similar to MI-SH, exposure control parameters are obtained through a series of simulated CATs and then they are used in the CAT to control the maximum observed item exposure rates. However, STR_C-SH selects items differently from MI-SH. Specifically, items are selected across strata based on the closeness between item difficulty and a current ability estimate. The exposure control parameter of the selected item is compared with a uniform random number to decide if the chosen item should be administered. After obtaining the exposure control parameters, they are used in the CAT to control the frequency with which items are administered.

Method

Simulation method was used to study the effectiveness of STR_C-SH; and to examine the effects of the test length, the number of strata, and the pre-specified maximum item exposure rate on the performance of different CAT methods. The item pool consists of 480 items from a large-scale achievement test. The three-parameter logistic (3-PL) item response theory (IRT) model was assumed and the BILOG computer program (Mislevy & Bock, 1990) was used to calibrate item parameters.

Ten thousand θ values were generated from a standard normal distribution. For each simulee, a fixed length CAT of 16, 20, 30, or 40 items was simulated. The ratio of the item pool size to the test length was 30, 24, 16, or 12, respectively. The rule-of-thumb for this ratio, recommended by Stocking (1994), is 12. Three content areas, denoted Content Area One, Content Area Two, and Content Area Three, were used for content consideration in the current research. A content control procedure that uses a modified multinomial model as described in Yi and Chang (2001) was implemented in each of the CAT methods. As implemented, the content balancing algorithm assured that each CAT consisted of about 40% Content Area One items and 30% each of Content Areas Two and Three items.

For STR_C and STR_C-SH, the item pool was stratified into two (denoted STR_C-2; STR_C-SH-2), four (denoted STR_C-4; STR_C-SH-4), or six (denoted STR_C-6; STR_C-SH-6) strata, respectively. The item pool was first divided into three groups based on the content specifications. The first group contains 192 items from Content Area One, and the second and third groups have 144 items from Content Areas Two and Three, respectively. Within each content group, following the steps of the *a*-stratified method with *b* blocking as described previously, two, four, or six strata were obtained, respectively. Next, all stratum one's from each content group were combined to form a single stratum one, all stratum two's were combined into

a single stratum two, and so on. Table 1 contains the descriptive statistics of the *a*- and *b*-parameters for the whole item pool and for the two, four, or six strata, respectively. The content coverage of each stratum is similar to that of the full item pool. As indicated in Table 1, the distribution of *b*-parameters closely matches that of the whole item pool, and the value of the *a*-parameters increases across the strata.

It is not uncommon in a CAT to select the first item to be of medium difficulty (Hambleton, Zaal, & Pieters, 1991; Hulin, Drasgow, & Parsons, 1983), assuming there is no prior knowledge about an examinee's ability. However, this may cause items with medium difficulty to be overexposed especially when examinees' ability is normally distributed with a mean close to the mean item difficulty for the pool. In practice, it maybe desirable to give an examinee a slightly easy item as the first item so as to help the examinee feel more comfortable with the testing situation, and to help reduce the exposure rates of middle difficulty items. In the current research, the first item was randomly selected from a list of ten optimal items assuming an examinee's initial ability estimate of -1, without content balancing constraints. More specifically, for STR_C and STR_C-SH, ten items were selected from the first stratum according to the closest match between item difficulty and the ability estimate of -1; for the MI-SH procedure, ten most informative items were selected at the ability estimate of -1. The first item was then randomly selected from those ten items.

The rest of the items were selected based on the item selection criteria endorsed by each of the methods from the designated content areas. For STR_C, two items were selected based on the same criteria, and one was then randomly chosen and administered. For STR_C-SH, the next item was selected if the following two conditions were satisfied: (1) item has the closest match between item difficulty and the current ability estimate; and (2) the uniform random number is less than or equal to the item exposure control parameter. For MI-SH, the next item was selected if the following two conditions were met: (1) item has the maximum information at the current ability estimate; and (2) the uniform random number is less than or equal to the item exposure control parameter. For both the STR_C-SH and MI-SH procedures, item exposure control parameters were obtained through a series of simulated CATs administered to 10,000 simulees; and the maximum item exposure rate was set at 0.10, 0.15, or 0.20, respectively. If the second condition of item selection was not met, then the next optimal item was selected and its exposure control parameter was compared to a new uniform random number.

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

Evaluation Criteria

The effectiveness of each method was evaluated in terms of overall item exposure control, item pool usage, and measurement precision in the simulated CATs. The number of items falling into a various range of the observed item exposure rate (*r*) was summarized. The χ^2 index, a measure to quantify the equalization of item exposure rates, was computed.

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

and

$$r_i = \frac{\text{number of times the } i^{th} \text{ item is used}}{m}, \quad (2)$$

where N represents the size of an item pool, L denotes the length of a test, and m is the number of examinees. Note that L/N denotes a desirable uniform rate for all items, and Equation (2) represents the observed item exposure rate.

The observed test overlap rate was computed in two steps: (1) calculate the total number of common items administered to each of the $m(m-1)/2$ pairs of examinees; and (2) divide this total count by $L[\,m(m-1)/2]$.

Bias and mean squared error (MSE) estimates were calculated as:

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

and

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

where $\hat{\theta}_m$ is m^{th} examinee's estimated ability, θ_m is the true ability. The correlation coefficient between $\hat{\theta}_m$ and θ_m ($\rho_{\hat{\theta}_m,\theta_m}$) was also calculated.

For the simulated CATs of 20 or 40 items, respectively, the conditional item exposure control and the conditional measurement precision were also obtained at nine equally spaced θ points from –2 to 2 in increments of 0.5 with 3,000 replications at each of the θ points.

Results

The results of this study are summarized in terms of overall item exposure control, item pool usage, and measurement precision. Additionally, the conditional item exposure control and measurement precision were also calculated at nine equally spaced θ points for CATs of 20 or 40 items, respectively. The effectiveness of each of the CAT methods is described first. The effects of the test length, the number of strata, and the pre-specified maximum item exposure rate on the performance of each method are then discussed in turn when it is applicable.

Overall Item Exposure Control, Item Pool Usage, and Measurement Precision

Tables 2 to 4 present the overall item exposure control, item pool usage, and measurement precision across the methods. The number of items falling into a various range of the observed item exposure rate is shown in Table 2. STR_C-SH and STR_C performed similarly in item exposure control, while MI-SH performed the worst among the methods. STR_C-SH and MI-SH both controlled the maximum observed item exposure rate at a pre-specified level, but the latter had considerably more items that exceeded the pre-specified maximum item exposure rate.

For STR_C, the number of items that were unexposed decreased, while the number of items having the maximum observed item exposure rates larger than 0.20 increased as the length of the CAT increased. The largest number of items having zero observed item exposure rates was 78 for STR_C-6 with the test length of 16, while the maximum number of items having observed item exposure rates larger than 0.20 was 16 for STR_C-4 with the test length of 40. For STR_C-2, all the items were administered at least once in the simulated CATs. For STR_C-4, there were some items that were not administered when the test length was small (e.g., 16 or 20). For STR_C-6, there were a few items that were never administered for different test lengths. The number of items falling into the other ranges of observed item exposure rates were similar across STR_C with different number of strata.

For STR_C-SH and MI-SH, the number of items having zero observed item exposure rates also decreased, while the number of items exceeding a pre-specified maximum item exposure rate increased as the test length increased. Compare to MI-SH, STR_C-SH and STR_C had a smaller number of items having zero observed item exposure rates, and a fewer number of items exceeding a pre-specified maximum item exposure rate. STR_C-SH and STR_C had a similar distribution of number of items falling into the various ranges of observed item exposure rates, except the former effectively controlled the maximum observed item exposure rate at a pre-specified level. The number of strata had a similar effect on the performance of STR_C-SH as of STR_C. For the methods that incorporated the SH procedure, the number of items that exceeded the pre-specified maximum item exposure rate increased when that rate decreased. MI-SH resulted in more items having zero exposure rates when a less stringent pre-specified maximum item exposure rate was implemented, while the number of unexposed items of STR_C-SH was unaffected.

Based on the items that were administered at least once in the simulated CATs, Table 3 contains descriptive information on observed item exposure rates across the methods. STR_C-SH and STR_C performed similarly on the means and minimums of observed item exposure rates, but the former effectively controlled the maximum observed item exposure rates at pre-specified levels. MI-SH had the largest mean and standard deviation (SD) for observed item exposure rate among the methods.

For STR_C, the number of items that was administered, the SD, and the maximum observed item exposure rate increased as the test length increased. STR_C-2 had the smallest means that equal to the ideal item exposure rates (i.e., test length / item pool size), while STR_C-6 had the largest means. There were little differences in the number of items administered, the SD, and the maximum observed exposure rate among STR_C with different number of strata; while STR_C-2 had the largest and STR_C-6 had the smallest minimum and maximum observed item exposure rates.

Similar to STR_C, for STR_C-SH and MI-SH, the number of items administered and the SD of observed item exposure rate increased as the test length increased. Except for the maximum observed item exposure rates, STR_C-SH follows pattern similar for that of STR_C. Both STR_C-SH and MI-SH controlled the maximum observed item exposure rate closely at a pre-specified level; however, MI-SH resulted in the fewest items that were used and the largest SD of observed item exposure rate. The tighter the pre-specified maximum item exposure rate, the smaller the SD of observed item exposure rate.

Table 4 presents the overall measurement precision and item pool usage across the methods. All the methods had very small overall bias. MI-SH had the best overall measurement precision, but the worst item pool usage. STR_C-SH and STR_C performed similarly. The difference of measurement precision was trivial, but the difference of item pool usage between MI-SH and STR_C-SH was substantial. MI-SH resulted in much larger χ^2 and observed test overlap rates than STR_C-SH.

For STR_C, the overall MSE decreased, while $\rho_{\hat{\theta}_m,\theta_m}$, χ^2 and observed test overlap rates increased as the test length increased; meaning the overall measurement precision improved and item pool usage was reduced when the length of the CAT increased. The differences in overall MSE, $\rho_{\hat{\theta}_m,\theta_m}$, χ^2 and observed test overlap rates among STR_C with different number of strata were small.

For STR_C-SH and MI-SH, a similar pattern was observed. MI-SH had the smallest MSE and the largest $\rho_{\hat{\theta}_m,\theta_m}$, which means that it had the best overall measurement precision. However, there is always a tread-off between measurement precision and item pool usage. MI-SH had the largest χ^2 and observed test overlap rates. A tighter pre-specified maximum item exposure rate did not affect the overall MSE and $\rho_{\hat{\theta}_m,\theta_m}$ of STR_C-SH and STR_C, but reduced χ^2 and observed test overlap rates for STR_C-SH in comparison to STR_C. A tighter pre-specified maximum item exposure rate, on the other hand, reduced MSE, increased $\rho_{\hat{\theta}_m,\theta_m}$, χ^2, and increased observed test overlap rates for MI-SH.

Conditional Item Exposure Control and Measurement Precision

The conditional item exposure control and measurement precision were calculated for CATs of 20 and 40 items. For CATs of 40 items with the pre-specified maximum item exposure rate of 0.10, the SH exposure control parameters could not be calculated for STR_C-SH-6. Tables 5 to 7 present the conditional item exposure control; and Figures 1 to 6 show the conditional measurement precision across the methods.

The number of items having zero observed item exposure rates conditional on ability across the methods is shown in Table 5. Across the ability points, MI-SH resulted in the most items that were never administered. STR_C-SH performed similarly or better at some ability points than STR_C.

For STR_C, the number of items having zero observed item exposure rates at each ability point increased as the test length increased from 20 to 40 items. There were few items having zero observed item exposure rates at the middle of the θ scale. The differences in the number of unexposed items at each ability point among STR_C with different number of strata were small.

MI-SH had the largest number of unexposed items, and across the ability points, the number of items having zero observed item exposure rates was distributed evenly. Similar to STR_C, STR_C-SH had a few unexposed items at the middle of the θ scale, and there was little difference in the number of unexposed items among STR_C-SH with different number of strata. The number of unexposed items increased as the pre-specified maximum item exposure rate increased (except for some ability points for the stratified methods with the test length of 20 items).

Table 6 presents the number of items having observed item exposure rate larger than or equal to a pre-specified maximum item exposure rate conditional on ability across the methods. STR_C resulted in fewer items than STR_C-SH that had observed item exposure rates larger than or equal to 0.20. MI-SH had more items exceeding a pre-specified maximum item exposure rate than STR_C-SH at most ability points.

For STR_C, there were fewer items that exceeded the exposure rate of 0.20 at the middle of the θ scale than at the two ends of the scale, for CATs of 20 items. When test length increased to 40 items, there were more items that had observed item exposure rates larger than or equal to 0.20 and they were more evenly distributed across the ability points. The differences in the number of items exceeding the exposure rate of 0.20 among STR_C with different number of strata were small.

For STR_C-SH and MI-SH, the number of items that exceeded a pre-specified maximum item exposure rate also increased as the test length increased. The tighter the pre-specified maximum exposure rate was, the more items that exceeded this rate. MI-SH resulted in the largest number of items that exceeded a pre-specified maximum item exposure rate, while there was little difference in the number of overexposed items among STR_C-SH with different number of strata.

The maximum observed item exposure rate conditional on ability across the methods is presented in Table 7. As indicated previously, STR_C could not control the maximum observed item exposure rate; thus, across the ability points, it had the largest observed item exposure rate among the methods. STR_C-SH and MI-SH closely controlled the maximum observed item exposure rate at a pre-specified level.

For STR_C, the maximum observed exposure rate was smaller at the middle of the θ scale than that of at the two ends of the scale. The maximum observed exposure rate increased as the test length increased from 20 to 40 items. The maximum observed item exposure rates were extremely large at the two ends of the θ scale, especially for CATs of 40 items. Even at the middle of the θ scale, the maximum observed item exposure rates were still quite large. The difference in the maximum observed exposure rate among STR_C with different number of strata was small. For methods with SH, the maximums of observed item exposure rates were close to their pre-specified levels. The differences across the methods were small.

Figures 1 to 6 display the conditional measurement precision for CATs of 20 and 40 items across the methods. For CATs of 20 items, there was more error at the lower end of the θ scale, while the methods with SH control had larger error than those methods without SH control. The differences of conditional measurement precision among STR_C with different strata were small, and the differences among methods with SH control were also negligible. There was more error with tighter exposure control.

For CATs of 40 items, the error was larger at the two ends of the θ scale, especially at the lower end of the scale. Similar to the results for CATs of 20 items, the difference in error among STR_C with different number of strata, and methods with SH control, was small. The amount of error reduced as the pre-specified maximum item exposure rate increased.

Summary of Results

The results of the study indicated that STR_C-SH is an effective method for achieving balanced item usage in a pool while maintaining measurement precision in CATs with practical content constraints. Both STR_C and STR_C-SH outperformed MI-SH in terms of overall item pool usage, while the two stratified methods resulted in similar performance; there were few items that were never administered, and few items that exceeded a pre-specified maximum item exposure rate. However, STR_C could not control the maximum observed item exposure rate at a certain level, especially when test length was long and item exposure control was evaluated conditional on ability.

Overall measurement precision improved while overall item exposure control and pool usage reduced, as the test length increased. The influence of different number of strata on the performance of the stratified methods was small; except there tended to be items that were unexposed when the item pool was stratified into more strata. Additionally, the conditional SH exposure control parameters could not be computed when the test length was longer and the pre-specified maximum item exposure rate was tighter for the larger number of strata. A tighter pre-specified maximum item exposure rate reduced the overall measurement precision, while improving the overall item exposure control and pool usage.

The methods with SH control performed similarly in conditional measurement precision, while STR_C with different number of strata also had similar conditional error. The former resulted in more error than the latter. As the test length increased, the conditional item exposure control reduced while conditional measurement precision increased. The conditional measurement precision reduced, while conditional item exposure control also reduced when a tighter pre-specified maximum item exposure rate was implemented.

Discussion

Many testing organizations are using or are considering using CAT as a test administration mode. CAT has the advantage of providing examinees with the flexibility of test scheduling and administering the test to small groups of examinees at more frequent time intervals. However, this advantage of CAT can also cause test security concerns: Some items may be more frequently exposed to examinees. Additionally, certain item selection algorithms of CAT may not administer some of the items in a pool frequently enough. Therefore, research

that investigates the possibility of developing item exposure control procedures that not only control highly discriminating items' exposure rates, but also increase less discriminating items' usage without sacrificing measurement precision, is very important. This kind of research will provide guidance for practitioners and researchers.

The Sympson-Hetter exposure control method was designed to control item exposure rate, and can be incorporated into different item selection methods. The maximum information selection method has the advantage of measurement efficiency. When the Sympson-Hetter exposure control procedure is incorporated into it, the maximum observed item exposure rate is closely controlled at a pre-specified level. However, this method can result in a large number of items not being administered. The *a*-stratified method with content blocking is a refinement of the original *a*-stratified method, which was developed to increase the usage of items with low *a*s and reduce the exposure rate of high *a* items without sacrificing measurement precision. Previous research (Leung et al., 2001; Yi & Chang, 2001) indicated that this method functions better than the previous *a*-stratified methods and the maximum information selection method with Sympson-Hetter exposure control. However, the *a*-stratified method with content blocking does not have a mechanism to control the upper limit of observed item exposure rates, so some items may still be overexposed. The Sympson-Hetter exposure control procedure can be combined with the *a*-stratified method with content blocking to retain its advantages and limit its disadvantages. Simulation studies showed that this method maintained the effectiveness of the *a*-stratified method with content blocking to produce balanced item usage within a pool while closely controlling the maximum observed item exposure rate at a pre-specified level. It also resulted in measurement precision that is comparable to that of the maximum information selection method with Sympson-Hetter exposure control.

For future studies, the performance of STR_C-SH can be examined with item pools that have different characteristics and with more complex content constraints. The possibility of applying STR_C-SH in computerized classification tests can also be investigated. The exposure control parameters were calculated based on a sample of examinees in the current study; in future, exposure control can be implemented conditioning on examinees' ability. The effectiveness of selecting more items from strata with higher *a* items can also be investigated.

References

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

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

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.

Hambleton, R. K., Zaal, N. J., & Pieters, J. P. (1991). Computerized adaptive testing: Theory, application, and standard. In R. K. Hambleton & N. J. Zaal (Eds.), *Advances in Educational and Psychological Testing: Theory and application* (pp. 341-366). Boston: Kluwer Academic Publishers.

Hulin, C. L., Drasgow, F., & Parsons, C. K. (1983). *Item response theory: Application to psychological measurement.* Dow Jones-Irwin: Homewood, IL.

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. (1980). *Applications of item response theory to particle testing problems*. Hillsale, NJ: Lawrence Erlbaum.

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

Mills, C. N., & Stocking, M. L. (1996). Practical issues in large-scale computerized adaptive testing. *Applied Measurement in Education, 9,* 287-304.

Stocking, M. (1994). *Three practical issues for modern adaptive testing item pools* (ETS Research Report No. 93-2). Princeton, NJ: Educational Testing Service.

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.

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.

Yi, Q., & Chang, H. (2001, June). *a-stratified computerized adaptive testing with content blocking.* Paper presented at the Annual Meeting of the Psychometric Society, King of Prussia, PA.

TABLE 1
Descriptive Statistics for a- and b-Parameters of the Whole Item Pool and the Two, Four, and Six Strata

	Parameter	N	Mean	SD	Minimum	Maximum
Item Pool						
	a	480	1.056	0.347	0.193	2.685
	b	480	0.111	1.060	-2.970	2.475
Two Strata						
1st Stratum	*a*	240	0.999	0.346	0.193	2.478
2nd Stratum		240	1.112	0.338	0.375	2.685
1st Stratum	*b*	240	0.111	1.064	-2.970	2.475
2nd Stratum		240	0.111	1.057	-2.800	2.308
Four Strata						
1st Stratum	*a*	120	0.766	0.220	0.193	1.550
2nd Stratum		120	0.971	0.230	0.444	1.609
3rd Stratum		120	1.138	0.256	0.567	1.852
4th Stratum		120	1.348	0.363	0.705	2.685
1st Stratum	*b*	120	0.111	1.058	-2.540	2.308
2nd Stratum		120	0.110	1.087	-2.970	2.475
3rd Stratum		120	0.111	1.051	-2.800	2.098
4th Stratum		120	0.113	1.055	-2.370	2.323
Six Strata						
1st Stratum	*a*	80	0.710	0.179	0.375	1.115
2nd Stratum		80	0.875	0.203	0.477	1.461
3rd Stratum		80	0.981	0.213	0.418	1.527
4th Stratum		80	1.091	0.237	0.193	1.627
5th Stratum		80	1.240	0.275	0.708	1.852
6th Stratum		80	1.437	0.367	0.829	2.685
1st Stratum	*b*	80	0.108	1.055	-2.970	2.109
2nd Stratum		80	0.115	1.095	-2.420	2.308
3rd Stratum		80	0.114	1.063	-2.240	2.475
4th Stratum		80	0.111	1.071	-2.540	2.064
5th Stratum		80	0.109	1.047	-2.370	2.133
6th Stratum		80	0.110	1.060	-2.800	2.323

TABLE 2

Number of Items Falling into Various Ranges of Observed Item Exposure Rates (r) across Methods

	Observed Item Exposure Rate Range					
Methods	r = 0.00	0.00 < r ≤ 0.05	0.05 < r ≤ 0.10	0.10 < r ≤ 0.15	0.15 < r ≤ 0.20	r > 0.20
Test Length = 16						
STR_C-2	0	402	61	15	2	0
STR_C-4	34	368	63	14	1	0
STR_C-6	78	307	80	15	0	0
Pre-Specified Rate = 0.10						
STR_C-SH-2	0	395	71	14	0	0
STR_C-SH-4	34	369	62	15	0	0
STR_C-SH-6	72	313	86	9	0	0
MI-SH	246	75	100	59	0	0
Pre-Specified Rate = 0.15						
STR_C-SH-2	0	404	58	18	0	0
STR_C-SH-4	34	366	64	15	1	0
MI-SH	280	85	25	53	37	0
Pre-Specified Rate = 0.20						
MI-SH	299	81	25	18	33	24
Test Length = 20						
STR_C-2	0	340	120	13	7	0
STR_C-4	19	318	118	22	3	0
STR_C-6	38	295	130	14	3	0
Pre-Specified Rate = 0.10						
STR_C-SH-2	0	337	126	17	0	0
STR_C-SH-4	20	312	130	18	0	0
STR_C-SH-6	37	289	138	16	0	0
MI-SH	195	93	108	84	0	0
Pre-Specified Rate = 0.15						
STR_C-SH-2	0	341	116	17	6	0
STR_C-SH-4	20	312	121	26	1	0
STR_C-SH-6	36	289	135	18	2	0
MI-SH	249	84	36	67	44	0
Pre-Specified Rate = 0.20						
MI-SH	273	87	27	18	48	27

Table 2 Continued

Methods	Observed Item Exposure Rate Range					
	r = 0.00	0.00 < r ≤ 0.05	0.05 < r ≤ 0.10	0.10 < r ≤ 0.15	0.15 < r ≤ 0.20	r > 0.20
Test Length = 30						
STR_C-2	0	220	195	52	6	7
STR_C-4	0	230	185	50	11	4
STR_C-6	12	212	198	42	14	2
Pre-Specified Rate = 0.10						
STR_C-SH-2	0	206	206	68	0	0
STR_C-SH-4	0	208	209	63	0	0
STR_C-SH-6	11	192	222	55	0	0
MI-SH	103	79	173	125	0	0
Pre-Specified Rate = 0.15						
STR_C-SH-2	0	224	181	64	11	0
STR_C-SH-4	0	226	186	54	14	0
STR_C-SH-6	12	206	198	56	8	0
MI-SH	167	99	35	106	73	0
Pre-Specified Rate = 0.20						
STR_C-SH-2	0	225	196	43	8	8
STR_C-SH-4	0	230	190	44	13	3
STR_C-SH-6	13	212	199	39	16	1
MI-SH	199	103	33	31	67	47
Test Length = 40						
STR_C-2	0	135	196	117	18	14
STR_C-4	0	118	214	115	17	16
STR_C-6	6	96	225	117	22	14
Pre-Specified Rate = 0.10						
STR_C-SH-2	0	91	211	178	0	0
STR_C-SH-4	0	86	235	159	0	0
STR_C-SH-6	4	65	229	182	0	0
MI-SH	19	63	206	192	0	0
Pre-Specified Rate = 0.15						
STR_C-SH-2	0	135	175	134	36	0
STR_C-SH-4	0	114	199	139	28	0
STR_C-SH-6	7	87	211	152	23	0
MI-SH	105	93	33	137	112	0
Pre-Specified Rate = 0.20						
STR_C-SH-2	0	133	195	115	25	12
STR_C-SH-4	0	127	197	118	26	12
STR_C-SH-6	6	102	217	116	33	6
MI-SH	148	103	38	31	75	85

TABLE 3
Descriptive Statistics of Observed Item Exposure Rates across Methods

		Descriptive Statistics*			
Methods	N	Mean	SD	Minimum	Maximum
Test Length = 16					
STR_C-2	480	0.0333	0.0255	0.0006	0.1664
STR_C-4	446	0.0359	0.0244	0.0004	0.1574
STR_C-6	402	0.0398	0.0247	0.0005	0.1449
Pre-Specified Rate = 0.10					
STR_C-SH-2	480	0.0333	0.0238	0.0011	0.1071
STR_C-SH-4	446	0.0359	0.0230	0.0004	0.1069
STR_C-SH-6	408	0.0392	0.0234	0.0003	0.1087
MI-SH	234	0.0684	0.0414	0.0001	0.1095
Pre-Specified Rate = 0.15					
STR_C-SH-2	480	0.0333	0.0253	0.0010	0.1479
STR_C-SH-4	446	0.0359	0.0244	0.0002	0.1510
MI-SH	200	0.0800	0.0629	0.0001	0.1566
Pre-Specified Rate = 0.20					
MI-SH	181	0.0884	0.0808	0.0001	0.2080
Test Length = 20					
STR_C-2	480	0.0417	0.0285	0.0058	0.1916
STR_C-4	461	0.0434	0.0279	0.0001	0.1922
STR_C-6	442	0.0452	0.0274	0.0002	0.1583
Pre-Specified Rate = 0.10					
STR_C-SH-2	480	0.0417	0.0258	0.0054	0.1068
STR_C-SH-4	460	0.0435	0.0253	0.0004	0.1103
STR_C-SH-6	443	0.0451	0.0256	0.0005	0.1051
MI-SH	285	0.0702	0.0406	0.0001	0.1077
Pre-Specified Rate = 0.15					
STR_C-SH-2	480	0.0417	0.0282	0.0049	0.1547
STR_C-SH-4	460	0.0435	0.0275	0.0001	0.1512
STR_C-SH-6	444	0.0450	0.0275	0.0001	0.1538
MI-SH	231	0.0866	0.0619	0.0001	0.1602
Pre-Specified Rate = 0.20					
MI-SH	207	0.0966	0.0817	0.0001	0.2092

* Descriptive statistics were obtained based on the items that were administered at least once

TABLE 3 Continued

Methods	N	Descriptive Statistics Mean	SD	Minimum	Maximum
Test Length = 30					
STR_C-2	480	0.0625	0.0375	0.0115	0.2493
STR_C-4	480	0.0625	0.0372	0.0003	0.2175
STR_C-6	468	0.0641	0.0356	0.0001	0.2058
Pre-Specified Rate = 0.10					
STR_C-SH-2	480	0.0625	0.0295	0.0126	0.1129
STR_C-SH-4	480	0.0625	0.0302	0.0008	0.1091
STR_C-SH-6	469	0.0640	0.0279	0.0001	0.1060
MI-SH	377	0.0796	0.0358	0.0001	0.1087
Pre-Specified Rate = 0.15					
STR_C-SH-2	480	0.0625	0.0345	0.0117	0.1555
STR_C-SH-4	480	0.0625	0.0354	0.0005	0.1564
STR_C-SH-6	468	0.0641	0.0334	0.0002	0.1581
MI-SH	313	0.0958	0.0612	0.0001	0.1588
Pre-Specified Rate = 0.20					
STR_C-SH-2	480	0.0625	0.0367	0.0105	0.2062
STR_C-SH-4	480	0.0625	0.0370	0.0009	0.2063
STR_C-SH-6	467	0.0642	0.0354	0.0001	0.2044
MI-SH	281	0.1068	0.0831	0.0001	0.2104
Test Length = 40					
STR_C-2	480	0.0833	0.0468	0.0210	0.3161
STR_C-4	480	0.0833	0.0453	0.0168	0.2799
STR_C-6	474	0.0844	0.0452	0.0003	0.2474
Pre-Specified Rate = 0.10					
STR_C-SH-2	480	0.0833	0.0259	0.0188	0.1087
STR_C-SH-4	480	0.0833	0.0251	0.0168	0.1084
STR_C-SH-6	476	0.0840	0.0259	0.0001	0.1123
MI-SH	461	0.0868	0.0306	0.0001	0.1079
Pre-Specified Rate = 0.15					
STR_C-SH-2	480	0.0833	0.0402	0.0210	0.1581
STR_C-SH-4	480	0.0833	0.0395	0.0175	0.1566
STR_C-SH-6	473	0.0846	0.0398	0.0004	0.1582
MI-SH	375	0.1067	0.0588	0.0001	0.1616
Pre-Specified Rate = 0.20					
STR_C-SH-2	480	0.0833	0.0439	0.0165	0.2080
STR_C-SH-4	480	0.0833	0.0440	0.0166	0.2128
STR_C-SH-6	474	0.0844	0.0436	0.0003	0.2063
MI-SH	332	0.1205	0.0818	0.0001	0.2106

TABLE 4
Overall Measurement Precision and Item Pool Usage across Methods

Methods	Measurement Precision and Item Pool Usage				
	Bias	MSE	$\rho_{\hat{\theta}_m,\theta_m}$	χ^2	Overlap
Test Length = 16					
STR_C-2	-0.0047	0.1838	0.9214	9.367	5.275%
STR_C-4	-0.0042	0.1894	0.9208	9.154	5.231%
STR_C-6	-0.0005	0.1786	0.9244	10.425	5.496%
Pre-Specified Rate = 0.10					
STR_C-SH-2	0.0076	0.1902	0.9196	8.112	5.014%
STR_C-SH-4	0.0007	0.1849	0.9220	8.287	5.050%
STR_C-SH-6	0.0038	0.1798	0.9235	9.496	5.302%
MI-SH	-0.0025	0.1353	0.9405	28.819	9.328%
Pre-Specified Rate = 0.15					
STR_C-SH-2	0.0003	0.1852	0.9214	9.234	5.248%
STR_C-SH-4	0.0031	0.1835	0.9209	9.178	5.236%
MI-SH	0.0046	0.1266	0.9447	45.993	12.907%
Pre-Specified Rate = 0.20					
MI-SH	-0.0041	0.1182	0.9476	61.643	16.167%
Test Length = 20					
STR_C-2	-0.0010	0.1489	0.9361	9.305	6.096%
STR_C-4	0.0036	0.1434	0.9376	9.447	6.125%
STR_C-6	0.0004	0.1340	0.9419	9.645	6.167%
Pre-Specified Rate = 0.10					
STR_C-SH-2	-0.0022	0.1462	0.9375	7.656	5.752%
STR_C-SH-4	0.0006	0.1381	0.9398	7.938	5.811%
STR_C-SH-6	-0.0026	0.1399	0.9396	8.599	5.949%
MI-SH	-0.0015	0.1114	0.9508	24.906	9.346%
Pre-Specified Rate = 0.15					
STR_C-SH-2	-0.0045	0.1445	0.9375	9.134	6.060%
STR_C-SH-4	0.0048	0.1410	0.9383	9.175	6.069%
STR_C-SH-6	0.0018	0.1373	0.9398	9.689	6.176%
MI-SH	-0.0036	0.0987	0.9559	42.740	13.062%
Pre-Specified Rate = 0.20					
MI-SH	0.0013	0.0951	0.9576	59.368	16.527%

TABLE 4 Continued

Methods	Measurement Precision and Item Pool Usage				
	Bias	MSE	$\rho_{\hat{\theta}_m,\theta_m}$	χ^2	Overlap
Test Length = 30					
STR_C-2	-0.0015	0.0870	0.9610	10.765	8.484%
STR_C-4	0.0004	0.0829	0.9628	10.590	8.447%
STR_C-6	-0.0041	0.0915	0.9588	10.236	8.373%
Pre-Specified Rate = 0.10					
STR_C-SH-2	-0.0014	0.0937	0.9586	6.666	7.630%
STR_C-SH-4	-0.0025	0.0909	0.9592	7.007	7.701%
STR_C-SH-6	-0.0077	0.0915	0.9591	6.541	7.603%
MI-SH	-0.0085	0.0829	0.9627	15.889	9.551%
Pre-Specified Rate = 0.15					
STR_C-SH-2	-0.0066	0.0907	0.9596	9.123	8.141%
STR_C-SH-4	0.0015	0.0871	0.9610	9.597	8.240%
STR_C-SH-6	-0.0038	0.0906	0.9595	9.099	8.136%
MI-SH	-0.0041	0.0728	0.9673	34.720	13.475%
Pre-Specified Rate = 0.20					
STR_C-SH-2	0.0005	0.0882	0.9605	10.346	8.396%
STR_C-SH-4	-0.0051	0.0885	0.9603	10.515	8.432%
STR_C-SH-6	-0.0050	0.0888	0.9602	10.194	8.365%
MI-SH	-0.0030	0.0659	0.9699	52.152	17.107%
Test Length = 40					
STR_C-2	-0.0030	0.0652	0.9705	12.600	10.949%
STR_C-4	-0.0015	0.0641	0.9708	11.788	10.780%
STR_C-6	-0.0040	0.0638	0.9708	12.083	10.842%
Pre-Specified Rate = 0.10					
STR_C-SH-2	-0.0013	0.0746	0.9659	3.847	9.126%
STR_C-SH-4	-0.0044	0.0725	0.9672	3.608	9.076%
STR_C-SH-6	-0.0081	0.0750	0.9660	4.147	9.188%
MI-SH	0.0010	0.0734	0.9664	6.821	9.745%
Pre-Specified Rate = 0.15					
STR_C-SH-2	-0.0005	0.0637	0.9704	9.267	10.255%
STR_C-SH-4	-0.0050	0.0644	0.9705	8.983	10.196%
STR_C-SH-6	0.0008	0.0642	0.9705	9.574	10.319%
MI-SH	-0.0051	0.0560	0.9742	26.695	13.886%
Pre-Specified Rate = 0.20					
STR_C-SH-2	-0.0056	0.0680	0.9690	11.093	10.636%
STR_C-SH-4	-0.0021	0.0635	0.9711	11.117	10.640%
STR_C-SH-6	-0.0007	0.0653	0.9703	11.293	10.677%
MI-SH	-0.0003	0.0523	0.9756	44.398	17.575%

TABLE 5

Number of Items Having Zero Observed Item Exposure Rate Conditional on Ability Across Methods

Methods	Ability								
	-2.0	-1.5	-1.0	-0.5	0	0.5	1.0	1.5	2.0
Test Length = 20									
STR_C-2	128	54	58	24	5	50	91	127	180
STR_C-4	138	90	74	40	37	51	59	136	192
STR_C-6	100	107	72	51	56	64	91	160	19
Pre-Specified Rate = 0.10									
STR_C-SH-2	113	67	45	24	12	38	87	121	160
STR_C-SH-4	129	96	68	43	30	46	85	119	170
STR_C-SH-6	107	89	77	55	51	67	95	137	167
MI-SH	234	227	230	229	226	213	232	241	247
Pre-Specified Rate = 0.15									
STR_C-SH-2	92	116	63	18	11	29	89	138	172
STR_C-SH-4	120	105	71	41	38	48	92	115	182
STR_C-SH-6	137	111	78	57	48	66	108	150	168
MI-SH	276	284	281	278	285	283	291	293	297
Pre-Specified Rate = 0.20									
STR_C-SH-2	109	75	59	23	15	33	80	115	179
STR_C-SH-4	112	98	69	35	36	55	95	158	194
STR_C-SH-6	114	94	77	63	50	64	99	165	209
MI-SH	314	309	308	311	308	308	320	329	332
Test Length = 40									
STR_C-2	157	117	108	67	43	71	108	179	216
STR_C-4	115	143	90	80	42	55	107	176	239
STR_C-6	154	140	110	66	65	70	122	164	232
Pre-Specified Rate = 0.10									
STR_C-SH-2	38	33	34	26	15	28	36	38	43
STR_C-SH-4	19	26	19	12	13	11	18	25	27
MI-SH	49	48	50	42	39	40	54	51	50
Pre-Specified Rate = 0.15									
STR_C-SH-2	119	87	69	44	45	64	91	125	135
STR_C-SH-4	110	84	85	61	41	47	71	122	130
STR_C-SH-6	84	76	62	61	48	59	78	106	103
MI-SH	161	145	158	154	140	152	161	161	170
Pre-Specified Rate = 0.20									
STR_C-SH-2	140	137	111	46	60	66	113	154	165
STR_C-SH-4	130	118	88	53	52	58	100	149	171
STR_C-SH-6	122	112	88	67	54	63	93	131	146
MI-SH	212	211	216	201	201	202	221	228	227

TABLE 6

Number of Items Having Observed Item Exposure Rate Larger than or Equal to the Pre-Specified Maximum Item Exposure Rate Conditional on Ability Across Methods

	Ability								
Methods	-2.0	-1.5	-1.0	-0.5	0	0.5	1.0	1.5	2.0
Test Length = 20									
Observed Exposure Rate ≥ 0.20									
STR_C-2	24	17	12	10	2	3	8	21	25
STR_C-4	32	23	18	7	1	4	9	23	30
STR_C-6	30	21	20	10	0	5	9	24	33
Pre-Specified Rate = 0.10									
Observed Exposure Rate ≥ 0.10									
STR_C-SH-2	87	75	76	81	74	57	60	83	89
STR_C-SH-4	96	71	75	77	45	38	61	80	80
STR_C-SH-6	87	91	73	72	55	47	70	81	82
MI-SH	91	101	89	92	95	96	101	105	98
Pre-Specified Rate = 0.15									
Observed Exposure Rate ≥ 0.15									
STR_C-SH-2	53	39	30	34	8	17	23	35	41
STR_C-SH-4	51	38	28	19	4	8	19	40	46
STR_C-SH-6	45	31	17	22	11	10	17	45	39
MI-SH	51	53	60	63	54	57	54	70	55
Pre-Specified Rate = 0.20									
Observed Exposure Rate ≥ 0.20									
STR_C-SH-2	29	23	15	7	2	1	7	21	32
STR_C-SH-4	29	18	13	5	1	1	7	21	32
STR_C-SH-6	29	16	14	8	0	3	7	23	27
MI-SH	41	45	41	38	40	44	39	34	43

TABLE 6 Continued

Methods	Ability -2.0	-1.5	-1.0	-0.5	0	0.5	1.0	1.5	2.0
Test Length = 40									
Observed Exposure Rate ≥ 0.20									
STR_C-2	55	60	72	74	80	59	69	66	50
STR_C-4	69	69	69	73	76	55	64	69	60
STR_C-6	72	76	69	70	72	56	69	72	62
Pre-Specified Rate = 0.10									
Observed Exposure Rate ≥ 0.10									
STR_C-SH-2	198	212	196	184	200	188	205	204	198
STR_C-SH-4	190	186	185	205	197	207	190	196	210
MI-SH	200	193	196	206	214	209	207	196	190
Pre-Specified Rate = 0.15									
Observed Exposure Rate ≥ 0.15									
STR_C-SH-2	134	128	127	110	108	106	104	127	116
STR_C-SH-4	124	119	106	127	115	95	98	111	117
STR_C-SH-6	119	120	122	127	98	85	102	130	115
MI-SH	122	136	128	124	135	116	131	127	121
Pre-Specified Rate = 0.20									
Observed Exposure Rate ≥ 0.20									
STR_C-SH-2	82	79	80	82	74	57	67	84	82
STR_C-SH-4	81	89	74	88	52	48	67	81	84
STR_C-SH-6	83	64	67	74	46	54	63	66	70
MI-SH	88	89	81	101	92	94	94	99	90

TABLE 7

Maximum Observed Item Exposure Rate Conditional on Ability Across Methods

Methods	Ability -2.0	-1.5	-1.0	-0.5	0	0.5	1.0	1.5	2.0
Test Length = 20									
STR_C-2	0.7883	0.7060	0.5380	0.2883	0.2487	0.2600	0.2380	0.5460	0.8420
STR_C-4	0.5687	0.5343	0.4417	0.2733	0.2090	0.2193	0.2693	0.5243	0.6537
STR_C-6	0.4820	0.4773	0.4330	0.2950	0.1763	0.2467	0.2973	0.4837	0.6033
Pre-Specified Rate = 0.10									
STR_C-SH-2	0.1167	0.1163	0.1157	0.1140	0.1177	0.1137	0.1240	0.1147	0.1163
STR_C-SH-4	0.1133	0.1163	0.1200	0.1197	0.1170	0.1167	0.1140	0.1217	0.1190
STR_C-SH-6	0.1147	0.1150	0.1160	0.1170	0.1150	0.1200	0.1247	0.1160	0.1183
MI-SH	0.1177	0.1193	0.1147	0.1177	0.1173	0.1140	0.1190	0.1147	0.1150
Pre-Specified Rate = 0.15									
STR_C-SH-2	0.1680	0.1720	0.1693	0.1720	0.1703	0.1683	0.1660	0.1670	0.1683
STR_C-SH-4	0.1677	0.1767	0.1630	0.1703	0.1613	0.1577	0.1717	0.1687	0.1720
STR_C-SH-6	0.1720	0.1630	0.1973	0.1687	0.1667	0.1597	0.1610	0.1737	0.1643
MI-SH	0.1663	0.1670	0.1720	0.1743	0.1693	0.1647	0.1697	0.1690	0.1683
Pre-Specified Rate = 0.20									
STR_C-SH-2	0.2230	0.2183	0.2147	0.2147	0.2060	0.2150	0.2147	0.2150	0.2180
STR_C-SH-4	0.2223	0.2200	0.2180	0.2207	0.2080	0.2110	0.2077	0.2290	0.2167
STR_C-SH-6	0.2133	0.2220	0.2163	0.2120	0.1783	0.2103	0.2167	0.2217	0.2203
MI-SH	0.2197	0.2197	0.2137	0.2203	0.2227	0.2187	0.2113	0.2210	0.2250
Test Length = 40									
STR_C-2	0.9643	0.9453	0.8743	0.6300	0.3900	0.5420	0.6260	0.8990	0.9880
STR_C-4	0.8277	0.8057	0.7360	0.5753	0.3987	0.5357	0.6257	0.8417	0.8950
STR_C-6	0.7133	0.7010	0.6543	0.5590	0.4400	0.5403	0.5837	0.7370	0.7960
Pre-Specified Rate = 0.10									
STR_C-SH-2	0.1170	0.1187	0.1160	0.1187	0.1140	0.1153	0.1233	0.1170	0.1160
STR_C-SH-4	0.1173	0.1217	0.1167	0.1160	0.1217	0.1217	0.1180	0.1173	0.1190
MI-SH	0.1183	0.1207	0.1177	0.1157	0.1153	0.1167	0.1200	0.1200	0.1187
Pre-Specified Rate = 0.15									
STR_C-SH-2	0.1717	0.1700	0.1707	0.1663	0.1693	0.1703	0.1710	0.1720	0.1707
STR_C-SH-4	0.1693	0.1703	0.1743	0.1650	0.1647	0.1707	0.1750	0.1727	0.1680
STR_C-SH-6	0.1697	0.1693	0.1733	0.1703	0.1730	0.1763	0.1690	0.1680	0.1700
MI-SH	0.1667	0.1697	0.1680	0.1670	0.1683	0.1663	0.1690	0.1673	0.1740
Pre-Specified Rate = 0.20									
STR_C-SH-2	0.2430	0.2240	0.2187	0.2200	0.2257	0.2210	0.2157	0.2180	0.2257
STR_C-SH-4	0.2207	0.2197	0.2193	0.2260	0.2177	0.2237	0.2223	0.2227	0.2197
STR_C-SH-6	0.2170	0.2293	0.2197	0.2207	0.2160	0.2170	0.2287	0.2213	0.2277
MI-SH	0.2267	0.2260	0.2247	0.2253	0.2193	0.2187	0.2240	0.2257	0.2197

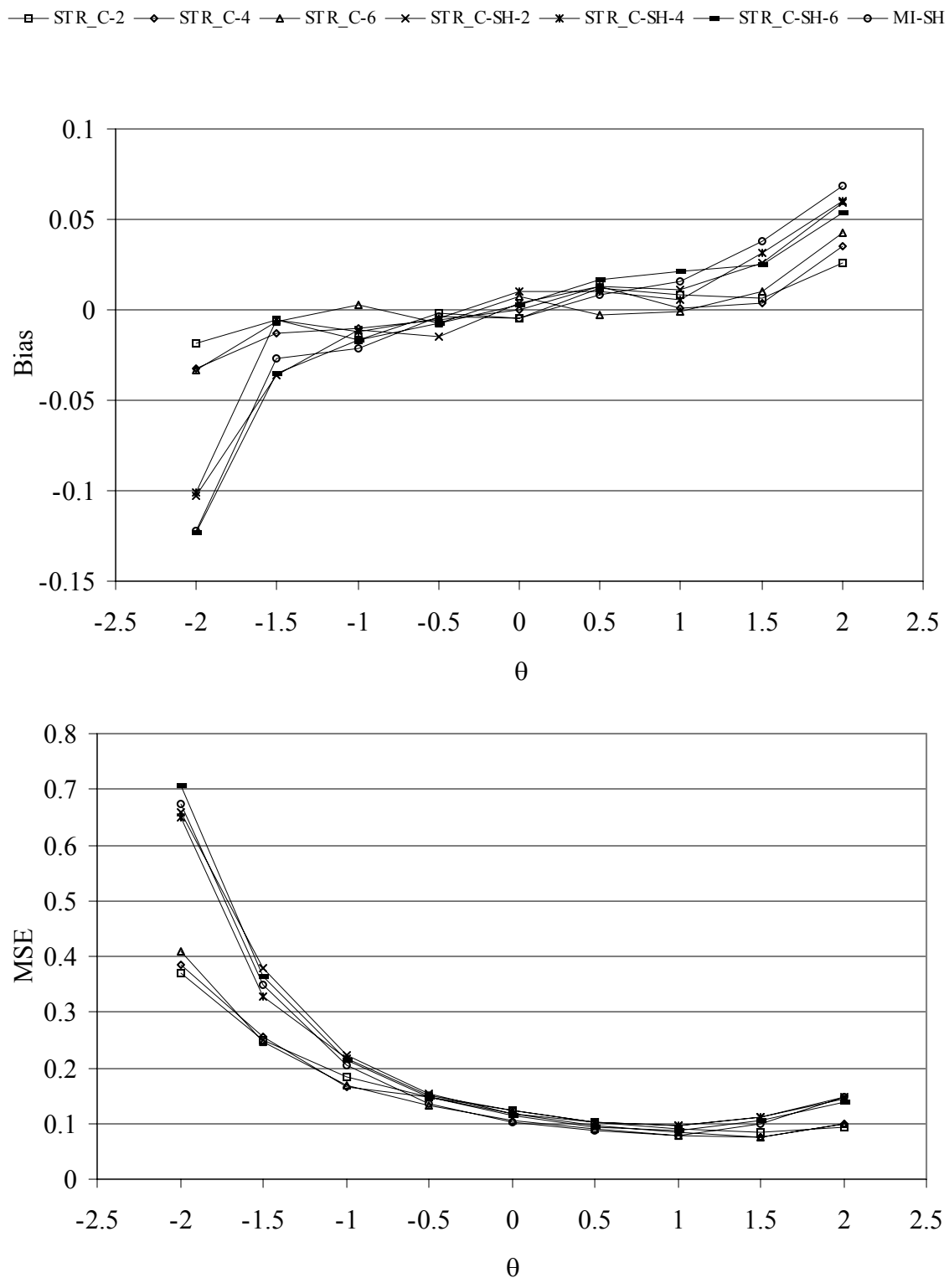

Figure 1. *Conditional bias and MSE for test length of 20 and pre-specified maximum item exposure rate of 0.10 across methods.*

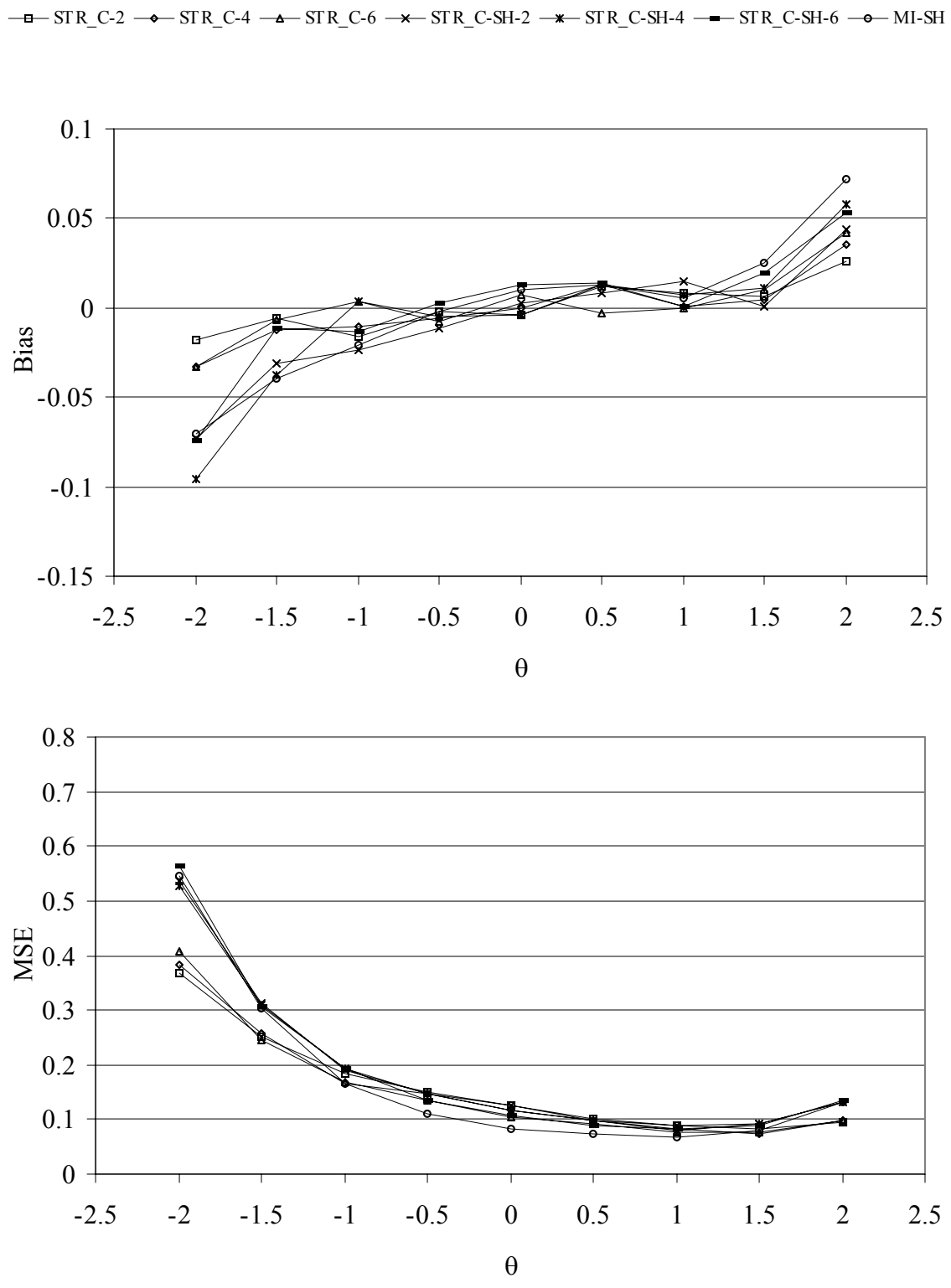

Figure 2. *Conditional bias and MSE for test length of 20 and pre-specified maximum item exposure rate of 0.15 across methods.*

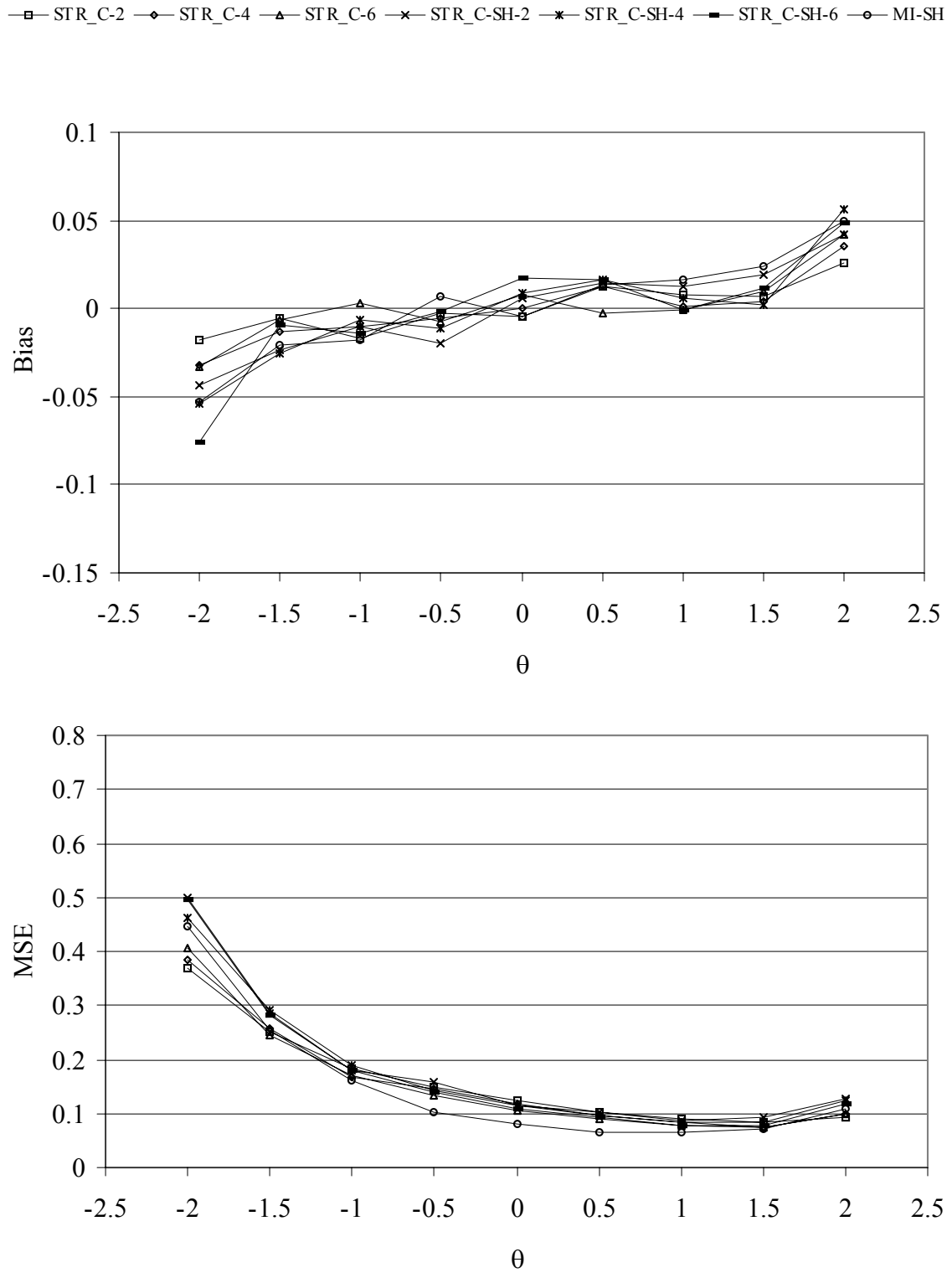

Figure 3. *Conditional bias and MSE for test length of 20 and pre-specified maximum item exposure rate of 0.20 across methods.*

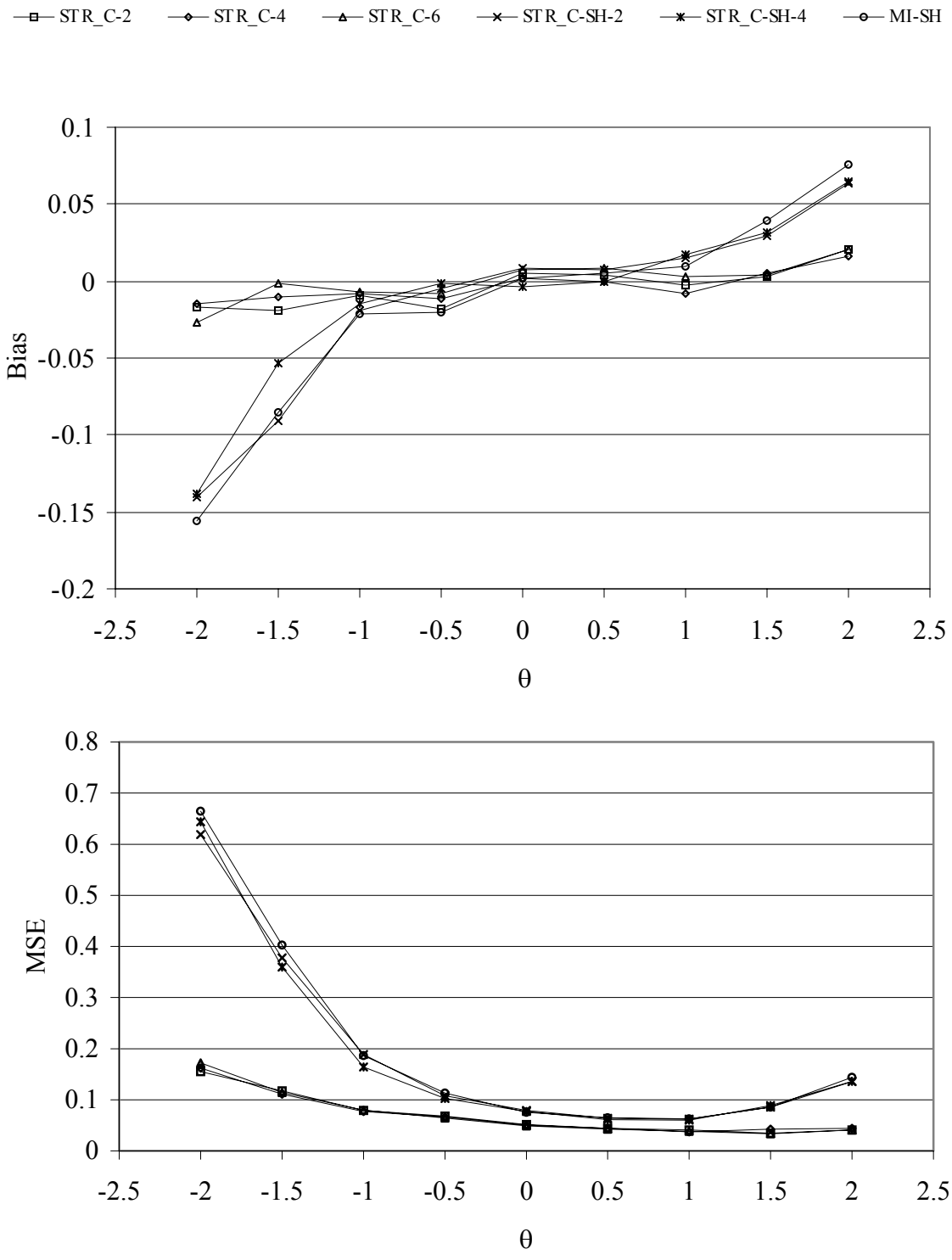

Figure 4. *Conditional bias and MSE for test length of 40 and pre-specified maximum item exposure rate of 0.10 across methods.*

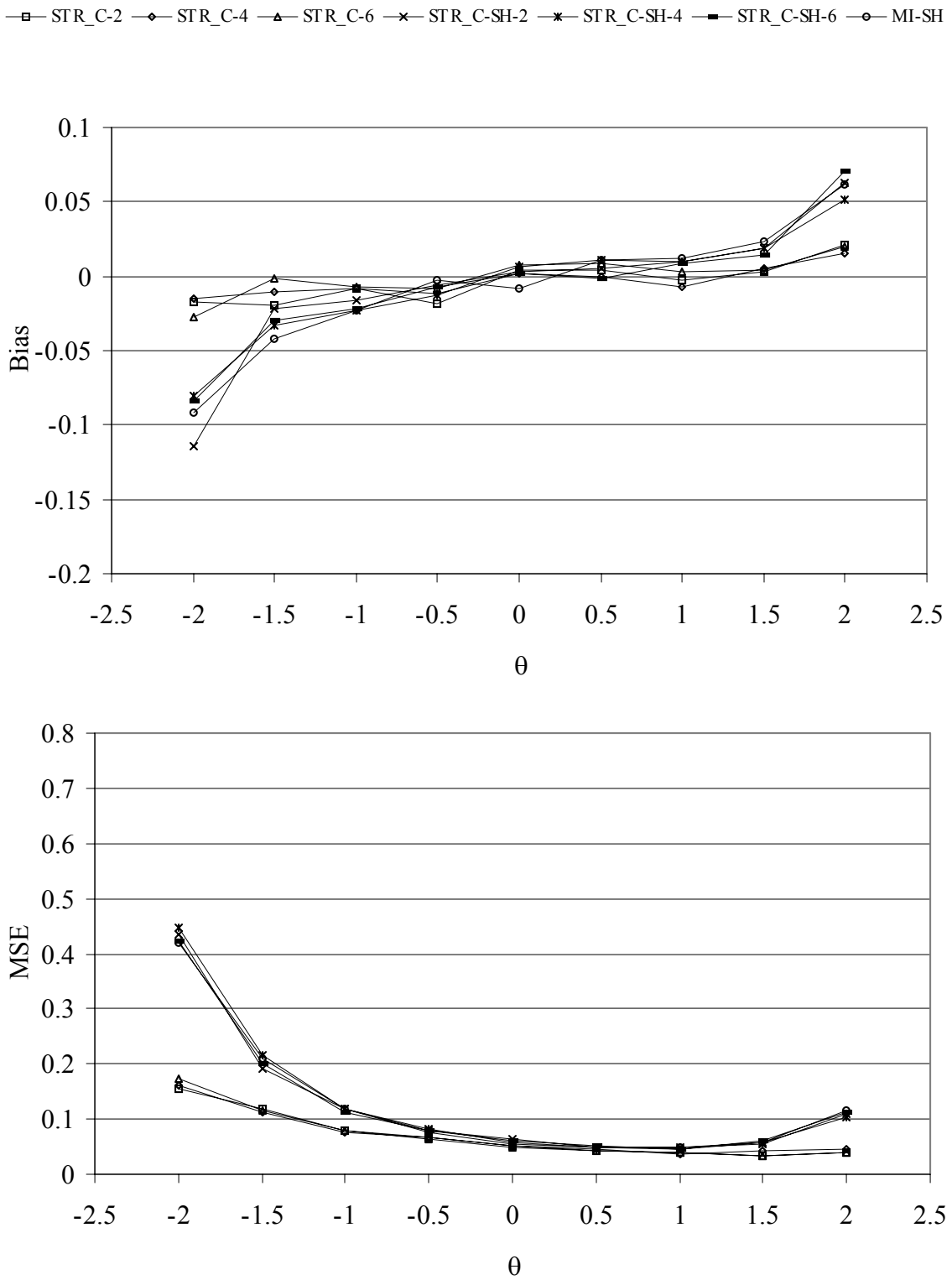

Figure 5. *Conditional bias and MSE for test length of 40 and pre-specified maximum item exposure rate of 0.15 across methods.*

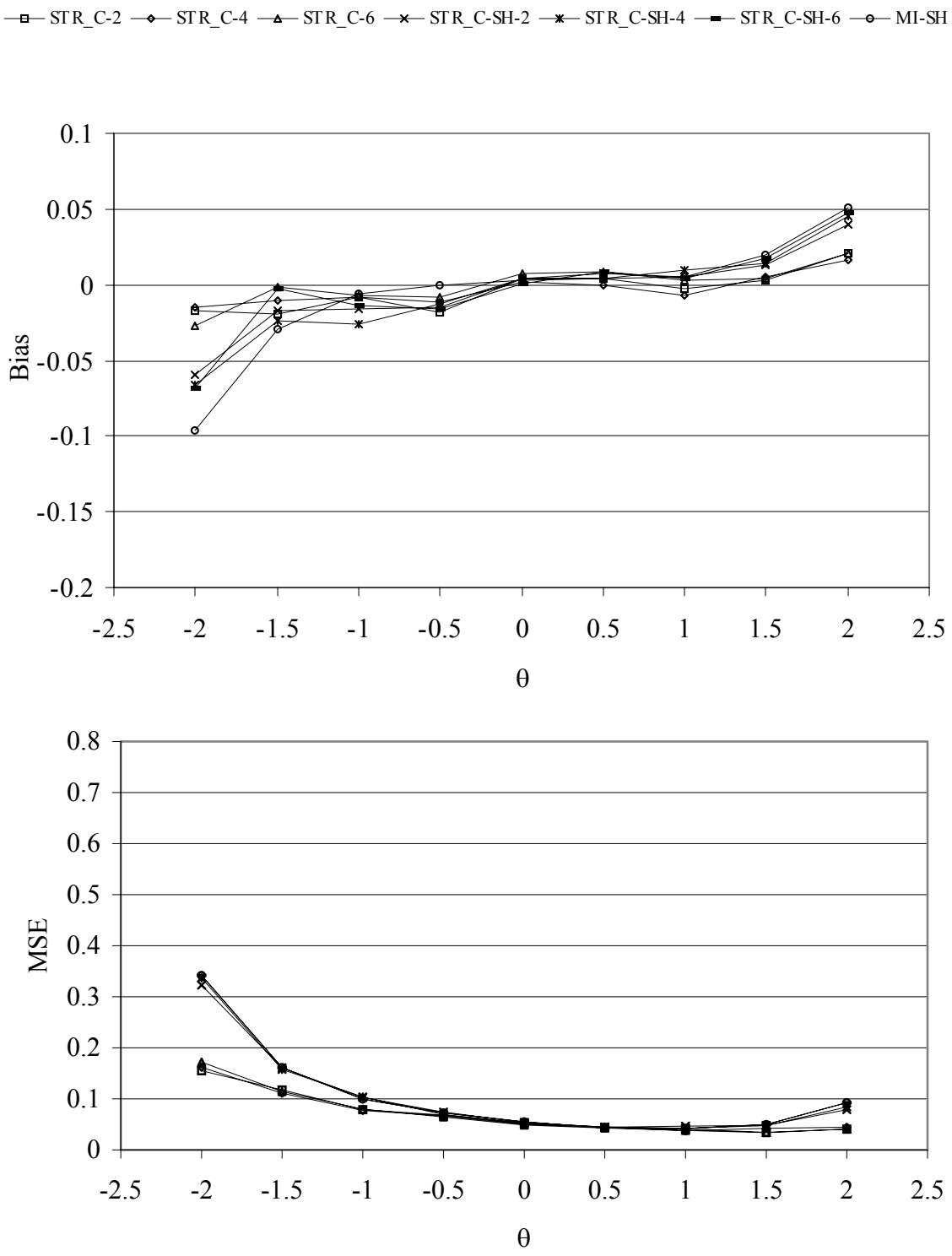

Figure 6. *Conditional bias and MSE for test length of 40 and pre-specified maximum item exposure rate of 0.20 across methods.*