Implementing the *a*-Stratified Method with *b* Blocking in Computerized Adaptive Testing with the Generalized Partial Credit Model

Qing Yi
ACT, Inc.

Tianyou Wang
Independent Consultant

Shudong Wang
Harcourt Educational Measurement

Paper Presented at the Annual Meeting of the American Educational Research Association, April 21-25, 2003, Chicago, IL.

Abstract

The purpose of the current study was to generalize the use of the *a*-stratified method with *b* blocking (Chang, Qian, & Ying, 2001; denoted STR_B) to computerized adaptive testing (CAT) with the generalized partial credit model (Muraki, 1992). Chang et al. (2001) showed that STR_B performs well in reducing highly discriminating items' exposure rates and increasing lower *a* items' usage while maintaining measurement precision in CAT with a dichotomous item pool. The characteristics of STR_B have not yet been investigated in a CAT containing polytomous items. In the current study, the method of selecting items based on maximum information with Sympson-Hetter exposure control (Hetter & Sympson, 1997) and the original *a*-stratified method (Chang & Ying, 1999) were compared to STR_B in CAT with the generalized partial credit model. Additionally, this study examined the effect of the length of CATs on the performance of each method. The results of this study showed that applying STR_B in CAT with the generalized partial credit model increased item usage in the pool and controlled item exposure rates, while maintaining measurement precision, especially for longer tests.

Implementing the *a*-Stratified Method with *b* Blocking in Computerized Adaptive Testing with the Generalized Partial Credit Model

Introduction

Polytomous item response theory (IRT) models have been developed to model the relationship between an examinee's ability level and the probability of responding in a particular category for items with multiple response categories. In general, there are two advantages of using a polytomous IRT model. First, the amount of information provided by a polytomous item is greater than that of a dichotomous item (Bock, 1972; Samejima, 1969; Sympson, 1983; Thissen & Steinberg, 1984). Second, the chance of detecting a mis-measured examinee using a polytomous item is greater than that of a dichotomous item.

With the recent proliferation of computer technology and the development of psychometric knowledge, computerized adaptive testing (CAT) has become a popular mode of test administration in educational and psychological testing. As more performance-oriented items are used in large-scale assessments, the need for applying a polytomous IRT model in CAT is increasing. The advancement in developing procedures to score polytomous items using computers has made the application of CAT with polytomously scored items more plausible (Burstein & Boodoo, in preparation; Burstein & Chodorow, 1999; Burstein, Kukich, Wolff, Lu, & Chodorow, 1998).

The advantages of using a CAT as the test administration mode do not come without cost. Test security has been a special concern in CAT, especially for a polytomous CAT, due to its smaller item pool size. Unlike a paper-and-pencil test in which examinees take the same test items at the same time, CAT examinees are tested individually with items that will be reused for future examinees. CAT examinees have the potential to memorize the items administered to

them, and to share those items with their friends who may take the test at a later date. To reduce the occurrence of such a problem, the frequency with which items are administered to examinees needs to be controlled. That is, item exposure rates have to be monitored to prevent item overexposure.

Different methods of item exposure control have been proposed by various researchers (e.g., Chang & Ying, 1999; Chang, Qian, & Ying, 2001; Davey & Parshall, 1995; Hetter & Sympson, 1997; Stocking & Lewis, 1998; Thomasson, 1995). The Sympson-Hetter method (Hetter & Sympson, 1997) uses item exposure control parameters to probabilistically control the frequencies with which items are administered. Sympson-Hetter item 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 match between item difficulty and a current ability estimate. Items with lower *a*-values are administered in the early stages of the test and high *a* items are used during the later stages.

Most of the research that evaluates the properties of the item exposure control methods has focused on CATs containing dichotomous items. The findings obtained from this kind of research cannot be automatically applied to a CAT with polytomous items. Research is needed to explore the properties of item exposure control methods, especially the advantages of using the *a*-stratified methods in CAT with polytomous IRT models.

Pastor, Dodd, and Chang (2002) compared five item exposure control methods in CAT with the generalized partial credit model. In their study, two polytomous item pools were stratified based on the discrimination parameter for using the original *a*-stratified method (Chang

& Ying, 1999) and variations of it. However, unlike in a CAT containing dichotomous items that selects items from the stratified pool based on the close match between item difficulty and a current ability estimate, in Pastor et al.'s study, items were selected within a stratum based on the maximum information at an ability estimate. The results have shown that the merits of the *a*-stratified method were not realized under the conditions that were examined in this study.

In the current study, the *a*-stratified method with *b* blocking (Chang et al., 2001; denoted STR_B) was applied to a CAT with the generalized partial credit model. The effectiveness of using STR_B in a polytomous CAT was examined by comparing it with the use of the original *a*-stratified method (denoted STR_A) and the maximum item information method with Sympson-Hetter exposure control (denoted SH_MI) in the polytomous CAT.

Item Exposure Control Methods

The Sympson-Hetter Method

The Sympson-Hetter (SH) procedure uses exposure control parameters to probabilistically control the frequencies with which the selected items are administered. 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 automatically 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. In the simulated CATs administered to a target population (N), the frequency of each item being selected (S_i), and the frequency of its being administrated (A_i) need to be recorded. 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 ($P_i(S) = S_i / N$), $P_i(A)$ is the probability of this item being administered ($P_i(A) = A_i / N$), $P_i(A \mid S)$ is the exposure control parameter of this item (i.e., 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.

a-Stratified Method

The *a*-stratified method (Chang & Ying, 1999; STR_A) takes 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 *a*-parameters, and the test is then divided into several corresponding stages. The early stages of a test administer items with lower *a*s and the later stages use items with higher *a*s.

Research has shown that STR_A outperforms the maximum item information method with SH exposure control (SH-MI) 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, 2002). As emphasized by Chang and Ying (1999), a crucial requirement for STR_A to perform well is that *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 over selected by STR_A (Ban, Wang, & Yi, 1999; Parshall, Hogarty, & Kromrey, 1999).

a-Stratified Method with b-Blocking

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 are scarce for strata corresponding to the later stages. As a result, the shortage of such items may lead them to become overexposed. This problem can 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 method with *b*-blocking, 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 different blocks in ascending order of the *b*-parameters. Within each block, items are sorted based on the *a*-parameters, from small to large. 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, the first stratum from each block is combined to a single stratum one, the second stratum to a single stratum two, 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 pool usage, while improving measurement precision (Chang et al., 2001).

Generalized Partial Credit Model

Various models have been developed for scoring polytomous items (Andrich, 1978; Bock, 1972; Masters, 1982; Muraki, 1992; Samejima, 1969). The generalized partial credit model (Muraki, 1992) is one of the most commonly used IRT models for polytomously scored items. The probability an examinee responds in a given category according to the generalized partial credit model can be expressed as:

$$P_{ik}(\theta) = \frac{\exp\left[\sum_{r=1}^{k} Da_i(\theta - b_i + d_{ir})\right]}{\sum_{u=1}^{m_i} \exp\left[\sum_{r=1}^{u} Da_i(\theta - b_i + d_{ir})\right]} \qquad (1)$$

where D is a scaling constant that equals 1.7, a_i is a discrimination parameter, b_i is an item location parameter, and d_{ir} is a category parameter for response category r. For a dichotomous item, the item location parameter, b_i, is called an item difficulty parameter. The category parameter is interpreted as the relative difficulty of getting response category r in comparison with other response categories for an item. For the generalized partial credit model, the following location constraint is imposed on the category parameters:

$$\sum_{k=2}^{m_i} d_k = 0. \qquad (2)$$

Method

Simulation studies were conducted to study the effectiveness of applying STR_B in a polytomous CAT and the effects of the length of CATs on the performance of each of the CAT methods. A real item pool consisting of 263 polytomously scored 1996 National Assessment of Educational Progress (NAEP) Science items was used for this study. The item pool was calibrated by Educational Testing Service (ETS) using the generalized partial credit model (Muraki, 1992). Table 1 lists the descriptive statistics of the item bank. Of the 263 items in the pool, 55 items have three categories, eight items include four categories, and the rest of the 263 items have two categories. The descriptive statistics of the item discrimination and location parameters in each of the 21 ranges of the location parameters are presented in Table 2. As shown in Table 2, this item bank includes very few items with low values of the location parameters. Details of how the item pool was calibrated can be found in Allen, Carlson, and Zelenak (1999).

As indicated in Equation 1, there is not a single item difficulty parameter in the generalized partial credit model as in the dichotomous three-parameter logistic IRT model.

Therefore, in the current study, the location parameters obtained through the generalized partial credit model were used similar to the *b*-parameters in the CAT with dichotomous items. The procedures of STR_A and STR_B as described in the previous research (see Chang & Ying, 1999; Chang et al. 2001) were followed to stratify the NAEP polytomous item pool to four strata. More specifically, for STR_A, items were sorted in ascending order according to the value of the discrimination parameters and then items were assigned to the four strata. The first stratum contained the items with the lowest *a*-values and the fourth stratum had the items with the highest *a*-values. When applying STR_B, the polytomous item pool was divided into blocks in an ascending order of the location parameters. Within each block, items were sorted based on the *a*-parameters, from the smallest to the largest. Then, across all the blocks, items with the lowest *a*s were 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, the first stratum from each block is combined to a single stratum one, and so forth. The first two strata contained 65 items, the third stratum included 66 items, and the last stratum had 67 items. Table 1 also presents the descriptive statistics of the item parameters for STR_A and STR_B across the four strata.

The simulations were conducted conditional on 21 equally spaced ability points from -4.0 to 4.0 in increments of 0.4 with 1000 simulees at each of the ability points. Wang and Wang (2001) used a simulation study to compare several ability estimation methods in a CAT with the generalized partial credit model, and discovered that the Warm's (1989) weighted likelihood estimate (WLE) method was most accurate. Thus, WLE was used as the ability estimation method in this study. Fixed-length CATs of 8, 12, and 20 items were simulated for the item pool with 263 items. Content balancing control was intentionally ignored in this study to avoid confounding effects.

For the simulated CATs, the two stratification methods selected items based on the close match between the item location parameter and the current ability estimate. For MI-SH, the next item was selected if the following two conditions were met: (1) item has the maximum Fisher information at the current ability estimate; and (2) a uniform random number is less than or equal the item exposure control parameter. For the MI-SH procedure, item exposure control parameters were obtained through a series of simulated CATs administered to the 250 simulees at each of the 21 ability points; and the maximum item exposure rate was set at 0.25. 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 uniform random number. This study did not implement a conditional SH procedure, thus, across the 21 ability points, the same set of item exposure control parameters were used.

The effectiveness of each method was evaluated in terms of item exposure control, item pool usage, and measurement precision in the simulated CATs. The number of items with exposure rate of zero or larger than 0.25 are summarized at each of the 21 ability points. The average between test overlap rate conditional on the 21 ability points is computed. Chen, Ankenmann, and Spray (1999) indicate that the average between test overlap rate can be mathematically defined as

$$\overline{T} = \frac{\sum_{i=1}^{n} m_i(m_i - 1)}{kp(p-1)} \tag{3}$$

where p is the number of fixed-length CATs administered, k represents the test length of the CATs, n is the item pool size, m_i denotes the frequency of item i was administered across the p CATs, and $\overline{T}$ is the between test overlap rate. Additionally, the overall item exposure rates and between test overlap rates across the CAT methods are also summarized.

For measurement precision, the bias, standard error (SE), and root mean square error (RMSE) are computed at each of the 21 ability points.

$$Bias(\hat{\theta}) = \sum_{p=1}^{P}(\hat{\theta}_p - \theta) \tag{4}$$

$$SE(\hat{\theta}) = \sqrt{\frac{1}{P}\left(\hat{\theta}_p - \frac{\sum_{t=1}^{P}\hat{\theta}_t}{P}\right)^2} \tag{5}$$

$$RMSE(\hat{\theta}) = \sqrt{\frac{1}{P}\sum_{p=1}^{P}(\hat{\theta}_p - \theta)^2} \tag{6}$$

where θ is simulees' true ability, $\hat{\theta}_p$ is the estimated ability for the p^{th} replication, and P is the number of replications.

Results

The results of this study are summarized in terms of the conditional item exposure control and measurement precision at 21 equally spaced θ points for CATs of 8, 12, or 20 items, respectively. The overall item exposure rates and average between test overlap rates are also computed. The effectiveness of each of the CAT methods is described first and the effects of the test length on the performance of each method are then discussed.

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

Figures 1a to 1c display the number of items with exposure rate of zero at each of the 21 θ points for CAT with 8, 12, or 20 items, respectively. MI_SH resulted in similar numbers of items with exposure rate of zero across the 21 θ points. The two stratification methods, on the other hand, had more items with zero exposure rates at the two ends of the ability continuum especially at the low end. There were few items with exposure rate of zero at the middle of the

ability scale. For CATs with 8 or 12 items, MI_SH had more items with exposure rate of zero along the ability continuum, while STR_A and STR_B either performed similarly or STR_B performed better at the high end of the ability scale. As the test length increased to 20, MI_SH still had more items with exposure rate of zero at ability above -1.6 while the three methods performed similarly at ability below -1.6. The number of items with zero exposure rates increased as the test length increased for the two stratified methods, but decreased for MI_SH.

Figures 2a to 2c present the number of items with exposure rate larger than 0.25 at each of the 21 ability points for the two stratified methods. The MI_SH exposure control parameters were not calculated conditionally on θ in this study, therefore, Figures 2a to 2c do not include MI-SH. The two stratified methods performed similarly, few items exceeded the exposure rate of 0.25 at the middle of the ability scale while more items had exposure rates larger than 0.25 at the two ends. For CATs with test length of 8 or 12, none of the items had exposure rate larger than 0.25 at the middle of the ability continuum for the two stratified methods. As the test length increased, the number of items with large exposure rate also increased.

Figures 3a to 3c present the average between test overlap rate conditional on the 21 ability points for CATs with 8, 12, or 20 items, respectively. MI_SH had the highest between test overlap rate for ability above zero, while STR_B had the lowest overlap rate. For CATs with 8 or 12 items, STR_B and MI_SH had similar average between test overlap rates that were lower than that of STR_A for ability below zero. When the test length increased to 20, MI_SH had the lowest overlap rate while the two stratified methods performed similarly at the low end of the ability scale. As the test length increased, the average between test overlap rate also increased.

The conditional measurement precision is displayed in figures 4a to 4c, 5a to 5c, and 6a to 6c. Figures 4a to 4c present the conditional bias. The three methods generated very small bias especially at the middle of the ability continuum. For CATs with 8 or 12 items, STR_B had relatively larger bias at the two ends of the ability scale. The three methods had similar bias for CATs with 20 items. As the test length increased, the conditional bias decreased. Figures 5a to 5c and 6a to 6c display the SE and RMSE at each of the 21 ability points across the methods. MI_SH had the smallest error at the high end of the ability scale; at the low end, the three methods performed similarly. STR_B had the largest error for CATs with 8 or 12 items at the high end of ability scale. As the test length increased, the conditional error decreased.

Figures 7a to 7c present the overall item exposure rate, and Table 3 contains the number of items falling into various ranges of overall exposure rates for CATs with different length across the methods. Figures 7a to 7c show that MI_SH controlled the overall exposure rate relatively well, with few items slightly exceeding the pre-specified maximum item exposure rate of 0.25 (the maximum observed exposure rate is less than 0.28). The two stratified methods had some items with exposure rates larger than 0.25. However, the two stratified methods did not have any items with an exposure rate of zero (except STR_B with test length of 12, having one such item). As the test length increased, the number of items exceeding the exposure rate of 0.25 increased, while the number of items with exposure rate of zero decreased (see also Table 3). Table 4 contains the descriptive statistics of overall observed item exposure rates and overall average between test overlap rates across methods. The descriptive statistics presented in Table 4 are computed based on the number of items that were administered at least once. The two stratified methods resulted in the mean of the overall observed item exposure rates that equals the value of test length divided by item pool size, while MI_SH had a larger mean and standard

deviation due to a large number of items with zero exposure rates. The maximum overall observed item exposure rates for the two stratified methods are larger than that of MI_SH. STR_B had the smallest overall average between test overlap rates while MI_SH had the largest overlap rate. As test length increased, this overlap rate also increased.

Discussion

The merits of applying the *a*-stratified methods in a dichotomous CAT have been investigated by several researchers (e.g., Chang & Ying, 1999; Chang et. al., 2001; Hau & Chang, 2001; Leung et al., 2002; Yi & Chang, in press). However, there is only a handful of research that has examined the possibility of using the idea of stratification in a polytomous CAT. The purpose of the current study was to generalize the use of the *a*-stratified method with *b* blocking (Chang et. al., 2001) to CAT with the generalized partial credit model (Muraki, 1992).

The results suggest that applying the stratified methods in a polytomous CAT could reduce the item exposure rate and increase item usage in a pool when treating the location parameter of the generalized partial credit model similarly to the difficulty parameter of the dichotomous IRT model. MI_SH had better conditional measurement precision; however, as test length increased to 20, the three methods resulted in very similar measurement precision. Results of this study are similar to those obtained from applying the *a*-stratified methods in a CAT with dichotomous items (Yi, 2002); that is, as test length increased, item exposure rates and item pool usage decreased, while measurement precision increased. The results of this study also indicated that the richness of the location parameters affected the performance of the stratified methods. In this study, there were limited numbers of items that had low values of the location

parameters; thus, resulted in more items with zero exposure rates, and larger measurement error at the low end of ability scale.

For future studies, the possibility of applying the stratified methods to different polytomous models can be explored. The size of a polytomous item pool tends to be small, the effects of the size of an item pool can be examined.

References

Allen, N. L., Carlson, J. E., & Zelenak, C. A. (1999*). The NAEP 1996 Technical Report, NCES 1999-452*. Washington, DC: National Center for Educational Statistics.

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika, 43,* 69-81.

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.

Bock, R. D. (1972). Estimating item parameters and latent ability when response are scored in two or more normal categories. *Psychometrika, 37,* 29-51.

Bock, R. D. (1972). Estimating item parameters and latent ability when response are scored in two or more normal categories. *Psychometrika, 37,* 29-51.

Burstein, J., & Boodoo, G. M. (in preparation). *Automated essay scoring for Advanced Placement.* Princeton, NJ: Educational Testing Service.

Burstein, J. C., & Chodorow, M. (1999). *Automated essay scoring for nonnative English speakers.* Proceedings of Workshop on Computer-Mediated Language Assessment and Evaluation of Natural Language Processing, Joint Symposium of the Association of Computational Linguistics and the International Association of Language Learning Technologies College Park, Maryland.

Burstein, J. C., Kukich, K., Wolff, W., Lu, C., & Chodorow, M. (1998). *Enriching automated scoring using discourse marking.* Proceedings of the Workshop on Discourse Relations and Discourse Marking, Annual Meeting of the Association of Computational Linguistics, Montreal, Canada.

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.

Chen, S., Ankenmann, R., & Spray, J. (1999). *Exploring the relationship between item exposure rate and test overlap rate in computerized adaptive testing.* ACT Research Report (RR-99-5).

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.

Gall, M. D., Borg, W. R., & Gall, J. P. (1996). *Educational research: An introduction.* New York: Longman.

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.

Leung, C., Chang, H., & Hau, K. (2002). Item selection in computerized adaptive testing: Improving the a-stratified design with the Sympson-Hetter algorithm. *Applied Psychology Measurement, 26,* 376-392.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47,* 149-174.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16,* 159-176.

Pastor, D., Dodd, P., & Chang, H. (2002). A comparison of item selection techniques and exposure control mechanisms in CATs using the generalized partial credit model. *Applied Psychological Measurement, 26,* 147-163.

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.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika monograph, No. 17*.

Sands, W., Water, B., & McBride, J. (Eds.) (1997). *Computerized adaptive testing: From inquiry to operation.* Washington, DC: American Psychological Association.

Segall, D. O. (1995, April). *Equating the CAT-ASVAB: Experiences and lessons learned*. Paper presented at the meeting of the National Council on Measurement in Education, San Francisco, CA.

Segall, D. O., & Carter, G. (1995, April). *Equating the CAT-GATB: Issues and approach*. Paper presented at the meeting of the National Council on Measurement in Education, San Francisco, CA.

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. B. (1983, June). *A new IRT model for calibrating multiple choice items.* Paper presented at the annual meeting of the Psychometric Society, Los Angeles CA.

Thissen, D. J., & Steinberg, L. (1984). A response model for multiple choice items. *Psychometrika, 49,* 501-519.

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.

Wang, S. & Wang, T. (2001). Precision of Warm's weighted likelihood estimation of ability for a polytomous model in CAT. *Applied Psychological Measurement, 25,* 317-331.

Warm, T. A. (1989). Weighted likelihood estimation of ability in the item response theory. *Psychometrika, 54*, 427-450.

Yi, Q. (2002, April). *Incorporating the Sympson-Hetter exposure control method into the a-stratified method with content blocking.* Paper presented at the annual meeting of the American Educational Research Association, April 1-5, New Orleans, LA.

Yi, Q., & Chang, H. (in press). *a*-Stratified CAT design with content blocking. *British Journal of Mathematical and Statistical Psychology.*

TABLE 1

Descriptive Statistics for Item Parameter Estimates of Item Bank and Four Strata

	Parameter	N	Mean	SD	Minimum	Maximum
Item Bank						
	a	263	0.5486	0.2290	0.1046	1.8711
	b	263	1.0424	1.6226	-4.8191	9.1434
	d_1	263	0.3301	1.7891	-14.5631	4.7828
	d_2	263	-0.2273	1.6532	-4.7828	14.5631
	d_3	55	-0.4445	1.7138	-3.7152	3.3598
	d_4	8	-0.3227	2.2363	-3.6275	3.3371
1st Stratum						
STR_A	a	65	0.3176	0.0670	0.1046	0.4141
STR_B		65	0.4030	0.1435	0.1046	1.0529
2nd Stratum						
STR_A	a	65	0.4682	0.0319	0.4141	0.5189
STR_B		65	0.4749	0.1226	0.2127	0.9886
3rd Stratum						
STR_A	a	66	0.5659	0.0310	0.5213	0.6225
STR_B		66	0.5606	0.1273	0.2475	0.8537
4th Stratum						
STR_A	a	67	0.8338	0.2452	0.6230	1.871
STR_B		67	0.7497	0.3005	0.2338	1.8711
1st Stratum						
STR_A	b	65	1.1864	2.5889	-4.8191	9.1434
STR_B		65	1.0559	1.7530	-3.6839	9.1434
2nd Stratum						
STR_A	b	65	1.1605	1.3365	-1.3070	5.2699
STR_B		65	1.0163	1.6670	-4.8191	5.4140
3rd Stratum						
STR_A	b	66	0.7781	1.1184	-2.2705	3.2679
STR_B		66	1.0243	1.5439	-3.4298	5.9140
4th Stratum						
STR_A	b	67	1.0485	0.9405	-0.9140	3.7920
STR_B		67	1.0725	1.5581	-3.5584	4.4083
1st Stratum						
STR_A	d_1	65	0.1807	3.0772	-14.5631	4.7828
STR_B		65	0.2649	2.9058	-14.5631	4.7828
2nd Stratum						
STR_A	d_1	65	0.3952	1.3129	-2.1744	3.2204
STR_B		65	0.1647	1.3937	-4.0618	3.2204
3rd Stratum						
STR_A	d_1	66	0.4543	1.0566	-1.8219	3.7848
STR_B		66	0.2984	1.2358	-1.8219	4.2224

TABLE 1 Cont'd

	Parameter	N	Mean	SD	Minimum	Maximum
4th Stratum						
STR_A	d_1	67	0.2894	0.8421	-2.0469	2.3736
STR_B		67	0.5850	1.0161	-2.8915	2.9872
1st Stratum						
STR_A	d_2	65	-0.3525	2.8162	-4.7828	14.5631
STR_B		65	-0.1587	2.6705	-4.7828	14.5631
2nd Stratum						
STR_A	d_2	65	-0.1280	1.3140	-2.9872	4.4576
STR_B		65	-0.1158	1.3166	-3.2895	4.4576
3rd Stratum						
STR_A	d_2	66	-0.2313	0.9039	-1.9804	1.7182
STR_B		66	-0.1248	1.1272	-4.2224	1.7182
4th Stratum						
STR_A	d_2	67	-0.1983	0.8086	-1.9552	2.0469
STR_B		67	-0.5030	0.9470	-2.9872	1.8520
1st Stratum						
STR_A	d_3	18	0.3053	2.3514	-3.4196	3.3598
STR_B		17	-0.7216	1.8945	-3.4196	3.0334
2nd Stratum						
STR_A	d_3	15	-1.1108	0.9501	-3.0056	0.4806
STR_B		20	0.1094	1.7042	-3.0056	3.3598
3rd Stratum						
STR_A	d_3	13	-0.7503	1.6226	-3.7152	1.2214
STR_B		9	-1.2326	1.5226	-3.7152	0.2259
4th Stratum						
STR_A	d_3	9	-0.3919	0.5438	-1.083	0.7070
STR_B		9	-0.3639	1.3401	-2.0771	2.8110
1st Stratum						
STR_A	d_4	3	1.8909	1.5173	0.3114	3.3371
STR_B		2	2.6806	0.9284	2.0242	3.3371
2nd Stratum						
STR_A	d_4	1	-0.7108		-0.7108	-0.7108
STR_B		4	-1.3405	1.6689	-3.6275	0.3114
3rd Stratum						
STR_A	d_4	2	-2.4813	1.6210	-3.6275	-1.3350
STR_B		1	-0.3595		-0.3595	-0.3595
4th Stratum						
STR_A	d_4	2	-1.2904	1.3165	-2.2214	-0.3595
STR_B		1	-2.2214		-2.2214	-2.2214

TABLE 2

Descriptive Statistics of a- and b-Parameters in the Ranges of b-Parameter

Range	Parameter	N	Mean	SD	Minimum	Maximum
$b \leq -4.0$	*a*	1	0.213		0.213	0.213
	b	1	-4.819		4.819	4.819
$-4.0 < b \leq -3.6$	*a*	1	0.203		0.203	0.203
	b	1	-3.684		3.684	3.684
$-3.6 < b \leq -3.2$	*a*	2	0.241	0.100	0.234	0.248
	b	2	-3.494	0.091	-3.558	-3.430
$-3.2 < b \leq -2.8$	*a*	3	0.245	0.122	0.105	0.324
	b	3	-2.939	0.196	-3.165	-2.818
$-2.8 < b \leq -2.4$	*a*					
	b					
$-2.4 < b \leq -2.0$	*a*	2	0.427	0.166	0.310	0.545
	b	2	-2.322	0.073	-2.374	-2.270
$-2.0 < b \leq -1.6$	*a*	2	0.238	0.080	0.182	0.295
	b	2	-1.733	0.066	-1.779	-1.686
$-1.6 < b \leq -1.2$	*a*	4	0.484	0.083	0.373	0.565
	b	4	-1.405	0.119	-1.561	-1.307
$-1.2 < b \leq -0.8$	*a*	5	0.497	0.138	0.296	0.652
	b	5	-0.996	0.070	-1.076	-0.914
$-0.8 < b \leq -0.4$	*a*	23	0.554	0.203	0.239	1.241
	b	23	-0.585	0.114	-0.763	-0.409
$-0.4 < b \leq 0.0$	*a*	13	0.563	0.109	0.440	0.836
	b	13	-0.160	0.128	-0.356	-0.003
$0.0 < b \leq 0.4$	*a*	21	0.537	0.152	0.214	0.923
	b	21	0.185	0.115	0.010	0.381
$0.4 < b \leq 0.8$	*a*	39	0.615	0.302	0.262	1.871
	b	39	0.614	0.105	0.402	0.799
$0.8 < b \leq 1.2$	*a*	28	0.640	0.276	0.317	1.438
	b	28	1.005	0.127	0.806	1.185
$1.2 < b \leq 1.6$	*a*	38	0.591	0.192	0.244	1.053
	b	38	1.401	0.114	1.604	1.967
$1.6 < b \leq 2.0$	*a*	20	0.576	0.216	0.311	1.053
	b	20	1.801	0.127	1.604	1.967
$2.0 < b \leq 2.4$	*a*	20	0.548	0.251	0.253	1.429
	b	20	2.220	0.083	2.083	2.357
$2.4 < b \leq 2.8$	*a*	16	0.529	0.169	0.342	0.981
	b	16	2.523	0.114	2.411	2.774
$2.8 < b \leq 3.2$	*a*	5	0.551	0.300	0.267	1.055
	b	5	2.973	0.138	2.847	3.195
$3.2 < b \leq 3.6$	*a*	5	0.439	0.090	0.309	0.523
	b	5	3.385	0.092	3.268	3.479
$3.6 < b$	*a*	15	0.353	0.122	0.215	0.704
	b	15	4.725	1.403	3.601	9.143

TABLE 3

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

	Overall Observed Item Exposure Rate Range				
Methods	$r = 0.00$	$0.00 < r \leq 0.05$	$0.05 < r \leq 0.15$	$0.15 < r \leq 0.25$	$r > 0.25$
Test Length = 8					
STR_A	0	225	34	1	3
STR_B	1	224	34	3	1
MI_SH	144	86	5	24	4
Test Length = 12					
STR_A	0	194	55	10	4
STR_B	0	188	63	10	2
MI_SH	125	87	10	34	7
Test Length = 20					
STR_A	0	139	87	25	12
STR_B	0	133	89	30	11
MI_SH	87	86	17	64	9

TABLE 4

Overall Average Between Test Overlap Rates and Descriptive Statistics of Overall Observed Item Exposure Rates across Methods

			Descriptive Statistics*			
Methods	Overlap	N	Mean	SD	Minimum	Maximum
Test Length = 8						
STR_A	0.0803	263	0.0304	0.0394	0.0003	0.3403
STR_B	0.0668	262	0.0305	0.0338	0.0010	0.2919
MI_SH	0.1841	119	0.0672	0.0902	0.0000	0.2762
Test Length = 12						
STR_A	0.1015	263	0.0456	0.0510	0.0014	0.3190
STR_B	0.0914	263	0.0456	0.0463	0.0003	0.3190
MI_SH	0.2007	138	0.0870	0.1012	0.0001	0.2589
Test Length = 20						
STR_A	0.1469	263	0.0760	0.0740	0.0024	0.3898
STR_B	0.1403	263	0.0760	0.0705	0.0002	0.3787
MI_SH	0.2129	176	0.1136	0.1082	0.0000	0.2609

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

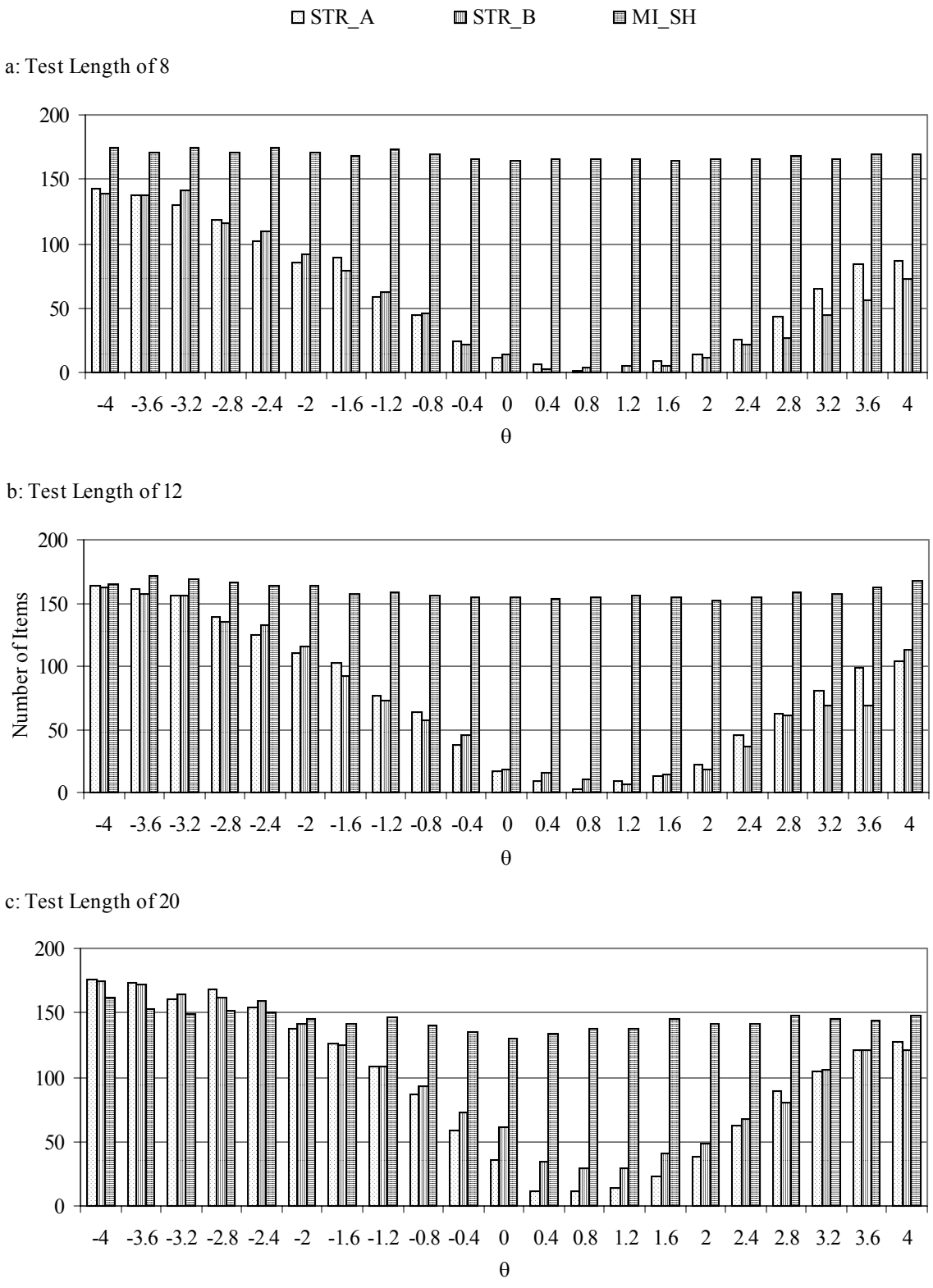

Figure 1. Number of items with exposure rate of zero at 21 ability points across methods.

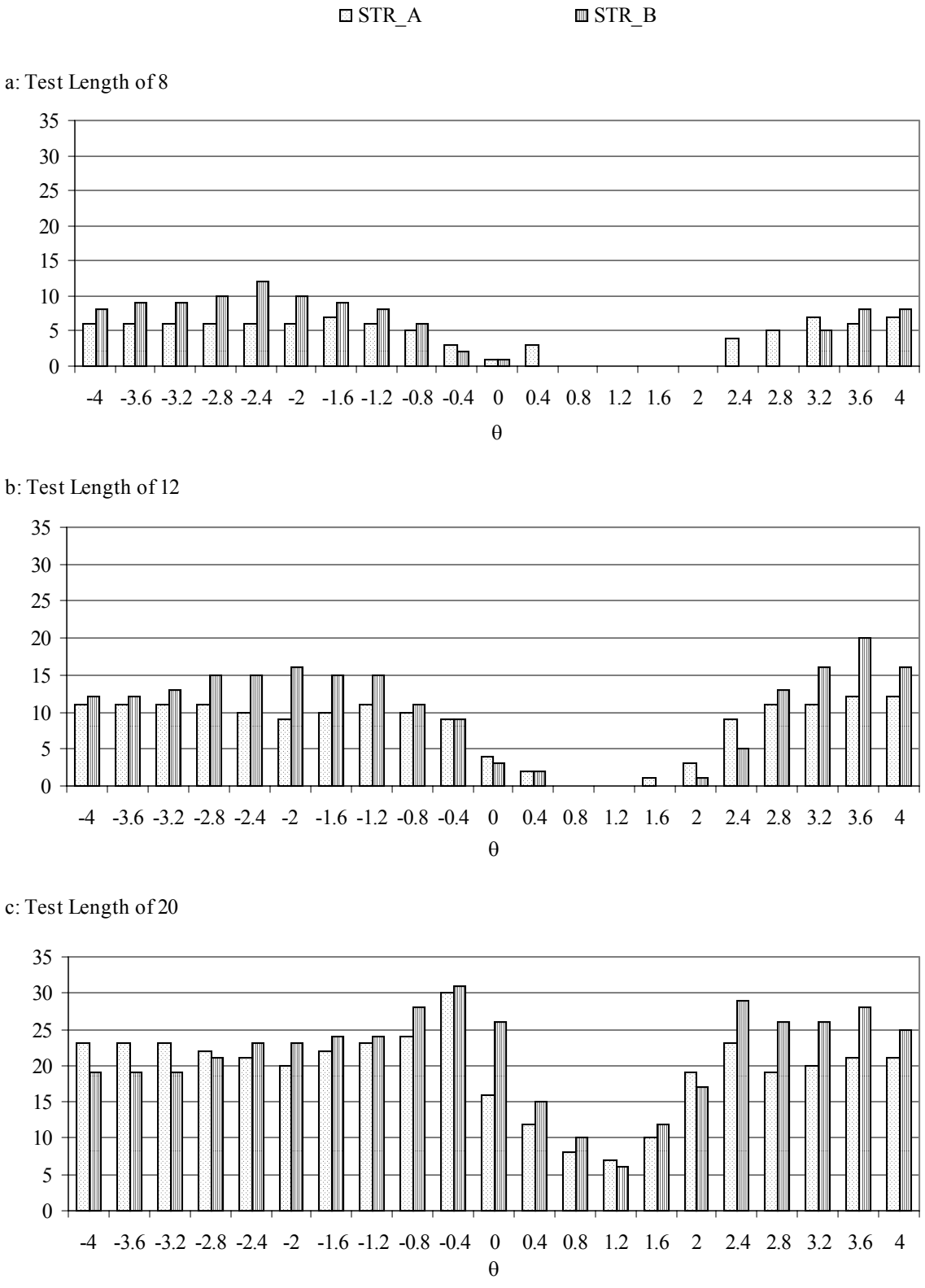

Figure 2. Number of items with exposure rates > 0.25 at 21 ability points for the two stratified methods.

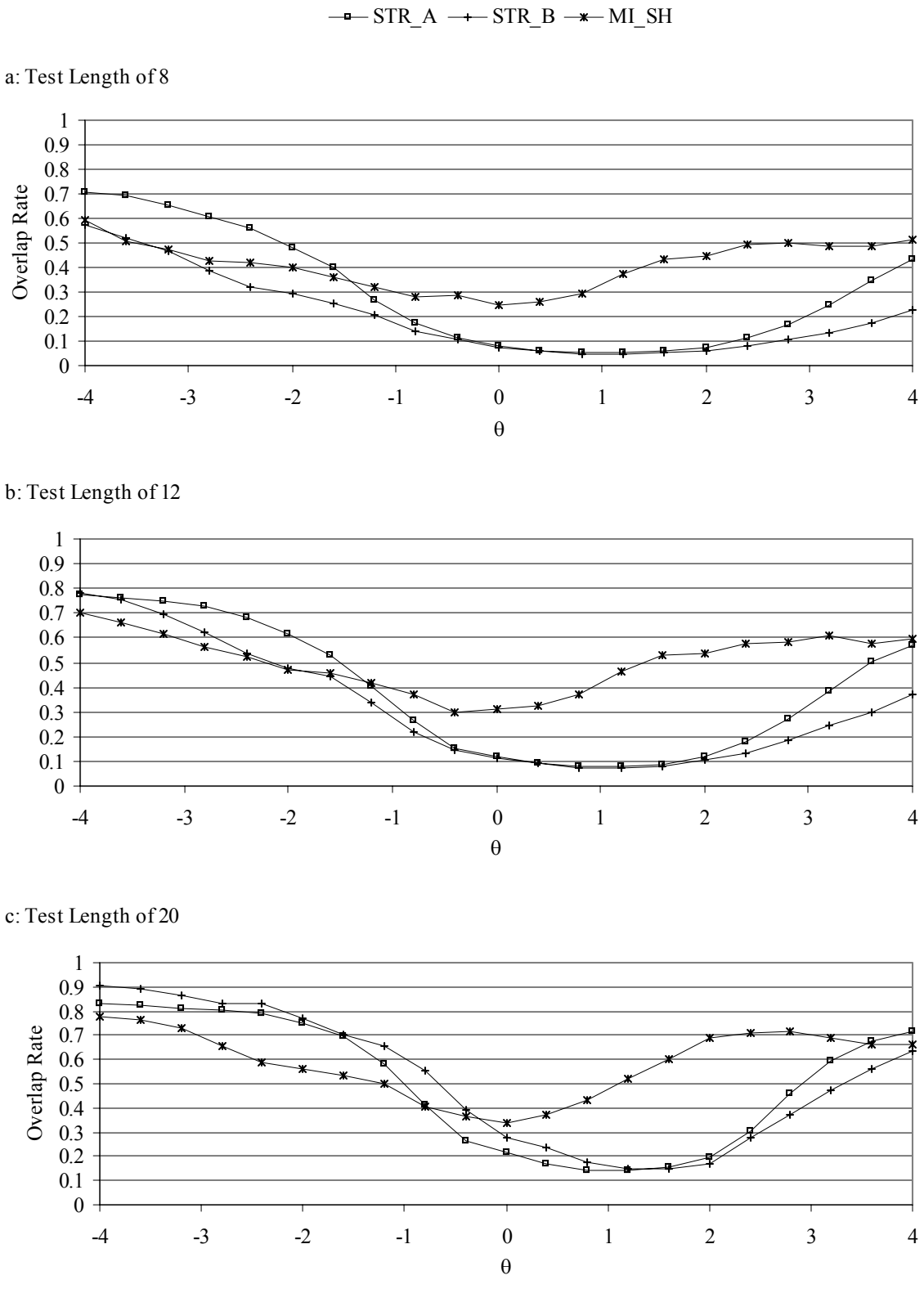

Figure 3. Average between test overlap rate at 21 ability points across methods.

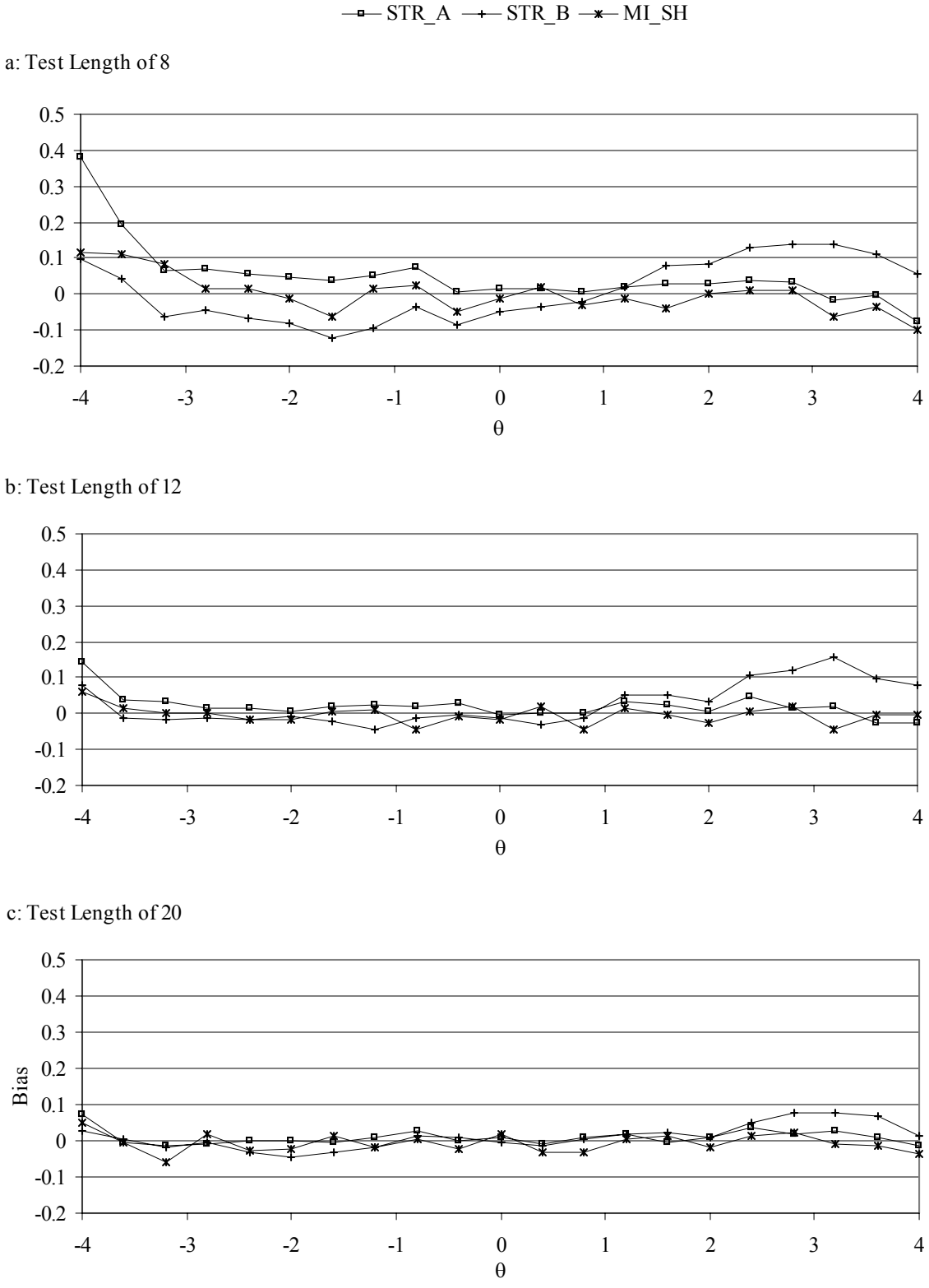

Figure 4. Bias at 21 ability points across methods.

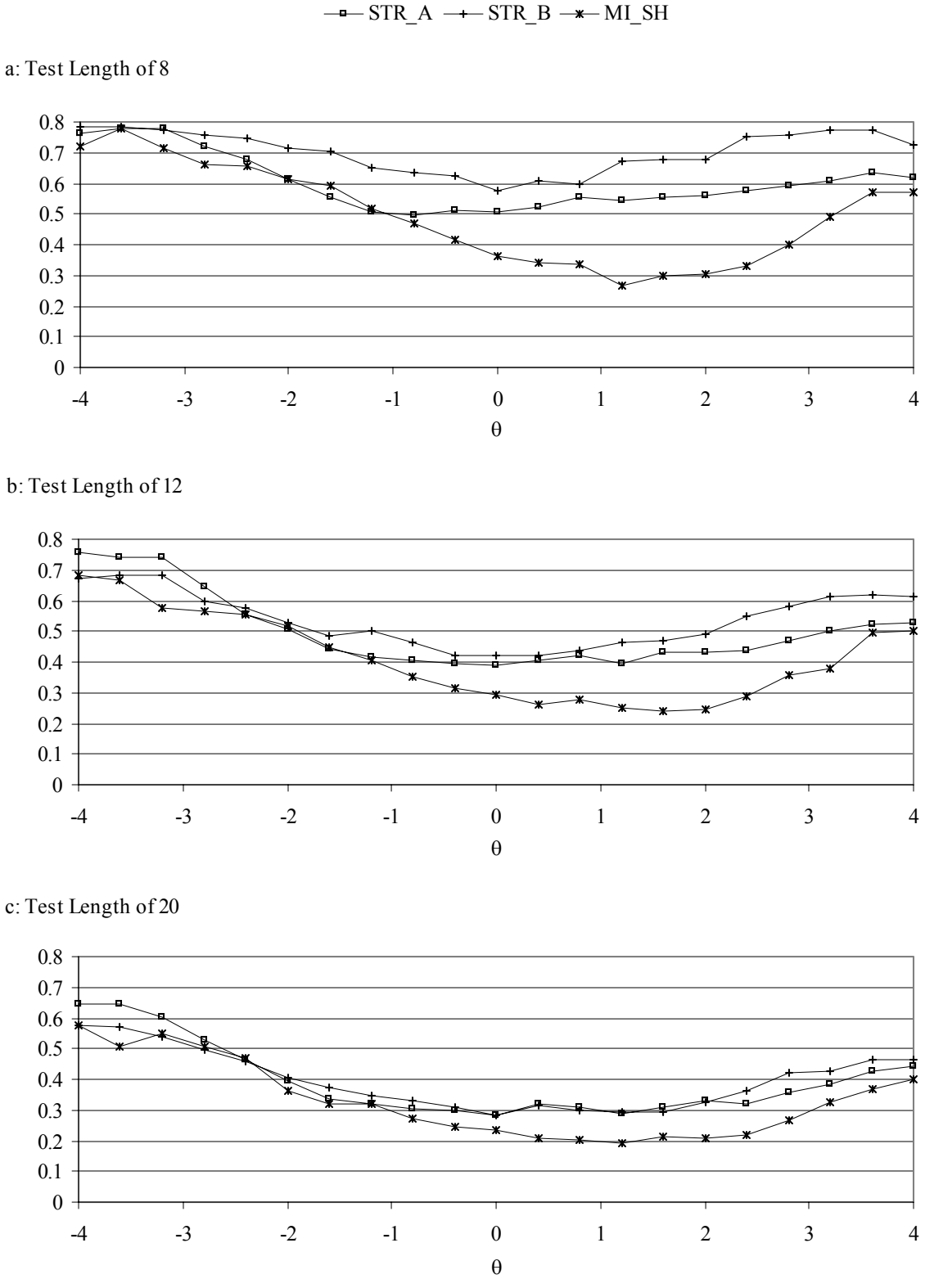

Figure 5. SE at 21 ability points across methods.

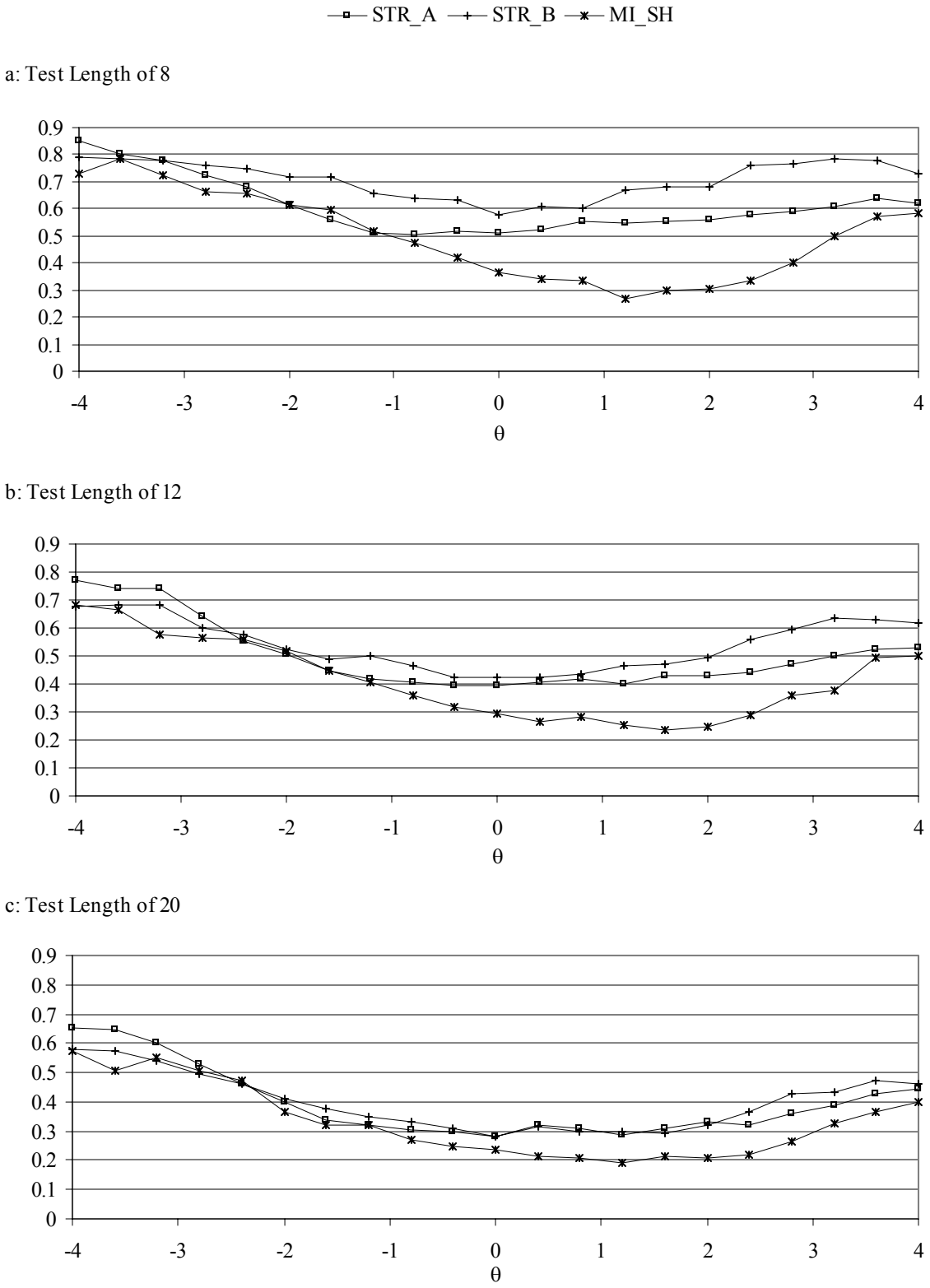

Figure 6. RMSE at 21 ability points across methods.

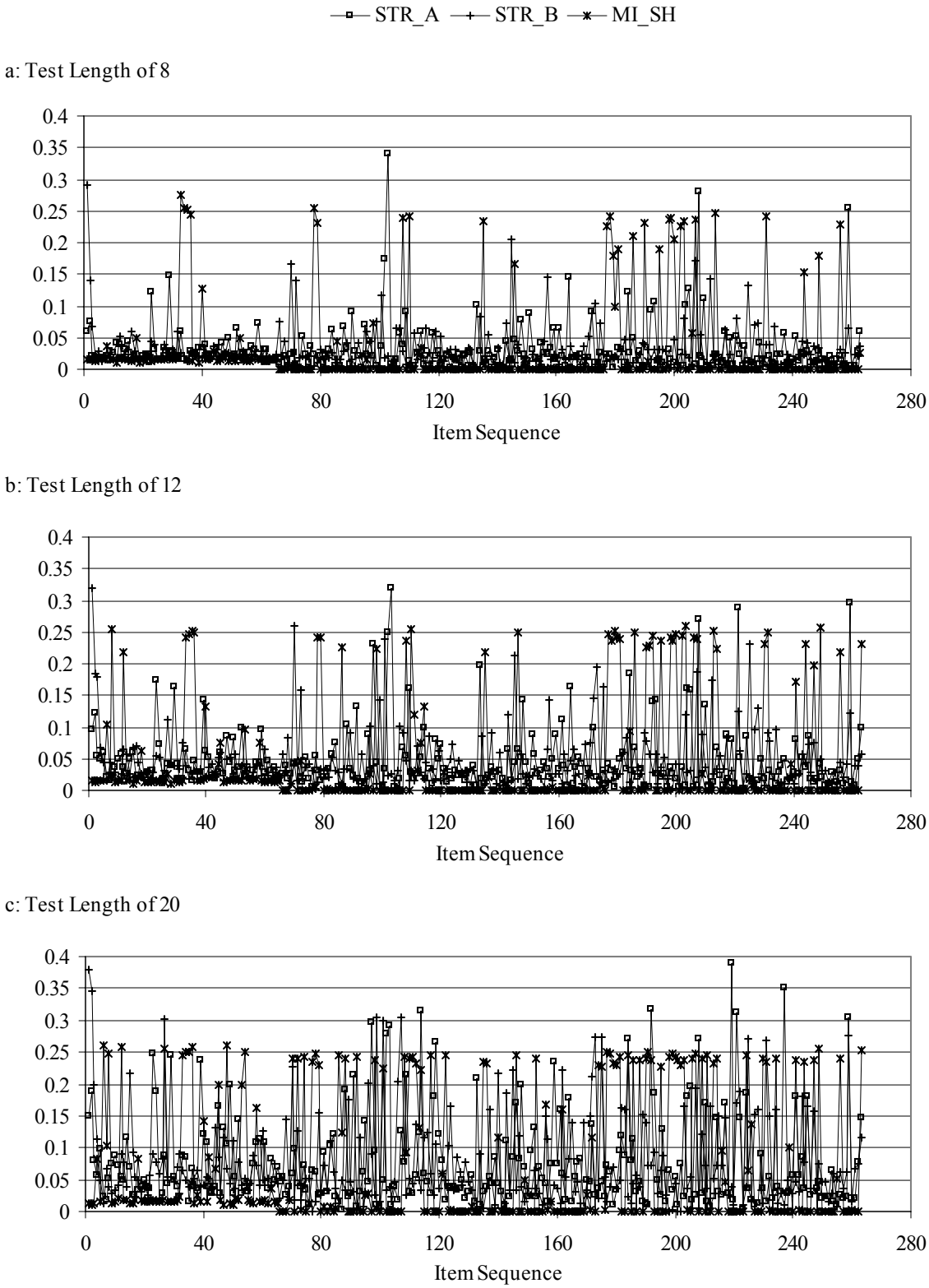

Figure 7. Overall item exposure rate distribution across methods.