To Stratify or Not: An Investigation of CAT Item Selection Procedures under Practical Constraints

Hui Deng and Tim Ansley

Paper presented at the 2003 NCME annual meeting, April 22-24, Chicago, IL

ABSTRACT

The maximum Fisher information procedure (F) is a commonly used algorithm for item selection in computerized adaptive testing. This approach leads to great enhancement in test efficiency, yet results in very unbalanced item usage. The a-stratified multistage CAT (STR) was developed to remedy the item usage problem with F, and was found to effectively balance item usage but yielded lower test efficiency. To address the efficiency loss, a refined stratification procedure has been proposed that allows more items to be selected from the high *a* strata and fewer items from the low *a* strata (USTR).

This study evaluated and compared the three item selection procedures, F, STR, and USTR, along with completely random item selection (RAN) with respect to test efficiency and item usage through CATs simulated under nine test conditions. The nine conditions resulted from combinations of three levels of practical constraints (no constraints, only exposure control, exposure control and content balancing), and three cases of the item selection space arising from combinations of various test lengths and target maximum exposure rates. The various item selection procedures were used to simulate CATs for an overall sample and for eleven conditional samples, and were compared in terms of error variances, item usage balance, overlap rates, and item pool utilization.

In general, the results showed that RAN consistently yielded the best item usage yet lowest efficiency. F showed an apparent efficiency advantage over STR and USTR under unconstrained item selection, but with very poor item usage. USTR reduced error variances for STR under various conditions, with small compromises in item usage. Compared to F, USTR enhanced item usage while achieving comparable efficiency; it achieved an efficiency level similar to F with improved item usage when items were selected under exposure

control, and the item selection space was restricted by long tests or a stringent security criterion. Under the extreme condition where item availability was severely limited by very stringent exposure control along with content constraints, USTR failed to increase efficiency relative to STR. The results provide implications for choosing an appropriate item selection procedure in applied settings.

To Stratify or Not: An Investigation of CAT Item Selection Procedures under Practical Constraints

Introduction

Item selection is a critical component of computer adaptive testing (CAT) (Weiss & Kingsbury, 1984; Wainer, 1990). A typical CAT item selection procedure strives to maximize test precision and efficiency by successively choosing items that provide optimal measurement at each level of estimated ability for an examinee. When applied in practical settings, however, some nonstatistical constraints usually need to be incorporated into the item selection process to ensure validity of the test. Content balancing is an important consideration, which ensures different tests across examinees cover the same proportion of content categories so that examinees are measured on the same composite of traits. Another factor is item exposure control. Because CATs are continuously administered from the same item pool over a period of time, without control for item exposure, some "popular" items may become known and no longer provide valid measurement. To ensure appropriate content coverage and prevent overexposure of items, some statistical algorithms are usually incorporated into the item selection process.

A good item selection algorithm should be compatible with content balancing and exposure control procedures, and achieve high precision and efficiency of the test while satisfying those constraints. In addition, for the sake of item pool maintenance, an ideal item selection algorithm should use all the items in the pool with nearly equal frequency (Way et al., 1998). As a commonly used algorithm, the maximum Fisher information procedure (F) selects items with maximum information at each of a succession of trait estimates. This

approach leads to great enhancement in test efficiency such that a CAT can achieve great precision with as few items as possible. This statistically optimal algorithm, however, was found to result in a subset of items (mostly high *a* items) exposed to almost every examinee, while another portion of the pool (items with low *a* values) was never used. This situation raised concerns about cost-efficiency of the pool utilization and posed security risks for CAT programs. When coupled with the Sympson-Hetter procedure (SH) (Sympson & Hetter, 1985), a commonly used exposure control method, the exposure rates of items that otherwise would be overexposed were effectively limited, but the overall distribution of exposure rates was still found drastically skewed (Chang & Ying, 1999; Hau & Chang, 1998). Specifically, some low *a* items had zero exposure rates, while many high *a* items achieved maximum allowable exposure rates at the early stages of CAT. This is because SH attempts to control the conditional probability that an item is administered when selected, but it has no direct control over the probability that an item is selected, therefore items with small probabilities of being selected may still be underexposed.

Chang and Ying (1999) proposed an alternative item selection procedure, the a-stratified multistage CAT (STR) to remedy the item usage problem with F. With this procedure the item pool is partitioned into strata according to *a* values and items are selected within one stratum at each stage of testing so that a natural balance of item usage can be achieved. It attempts to maintain test efficiency by using low *a* items in the early stages and deferring the use of high *a* items to the later stages of a CAT. The rationale is that the low *a* items provide more global information when trait estimates may be substantially deviant from the true level at the early stages of test, while high *a* items can better contribute to estimation when trait estimates are getting closer and closer to the true level at the later

stages (Chang and Ying, 1996). For item pools with correlated *a* and *b* values, the procedure has been modified into a two-stage process to take into account both *a* and *b* values in pool stratification (Chang, Qian & Ying, 2000). For CATs requiring content balance, the idea of the two-stage stratification was extended into a tri-stage process, so that content specification is also incorporated into pool stratification (Yi & Chang, 2001). It should be noted that STR has no direct control over exposure rates for individual items, therefore some items may still be overexposed unless some exposure control algorithm is imposed in item selection.

When compared to F through CATs simulated using both ideal and operational item pools, STR was found to result in more evenly distributed exposure rates and reduced overlap rates, while achieving somewhat lower test efficiency (Chang & Ying, 1999). Deng and Chang (2001) argued that the efficiency loss of STR could be attributed to insufficient use of high *a* items at later stages of the test. The point is that by drawing equal numbers of items from each stratum, the current STR procedure results in observed exposure rates for most items in the pool far below the target maximum exposure criterion, which leads to underuse of high *a* items later in the test and may degrade test efficiency to some extent. This consideration has led to a refined stratified procedure that allows more items to be selected from the high *a* strata and fewer items selected from the low *a* strata (USTR), which was found in a simulation study to effectively improve test efficiency over STR without unacceptably degrading item usage (Deng & Chang, 2001). The promising results about USTR have driven the current study for a systematic investigation of the comparative performances of F, STR, and USTR under a variety of realistic test conditions with item selection constrained by content coverage and exposure control.

In operational CATs, item selection often has to simultaneously address multiple and often conflicting demands, such as efficiency of ability estimation, content coverage, and exposure control. In addition, the behavior of an item selection procedure is influenced by the item selection space which can be viewed from the relationship between available and required item exposure. For a specific CAT, the required item exposure can be quantified as $n \times m$ where n is test length and m represents the examinee sample size. For a given item pool, the available item exposure is $N \times r \times m$ where N stands for item pool size and r is the target maximum exposure rate. With a fixed item pool and a fixed examinee sample, the relationship between the available and the required exposure is determined by test length and target maximum exposure rate.

The purpose of this study was to evaluate and compare the three item selection procedures, F, STR, and USTR, along with the completely randomized item selection (RAN, serving as a baseline comparison) with respect to test efficiency and item pool usage through simulated CATs under systematically varied CAT design conditions. The test conditions varied in terms of absence or presence of practical constraints (content balancing and exposure control), and degree of restriction on item selection space due to various test lengths and target maximum exposure rates. The results from the study are intended to provide general guidelines for choosing an item selection procedure in realistic CAT settings.

Method

Design

In this study, the item selection space was manipulated by varying test length (n) and target maximum exposure rate (r). Since the item pool size was fixed at N=300 and examinee sample size was fixed at m=3000 for the overall sample and m=1000 for the conditional

samples, the ratio between the available item exposure ($N \times r \times m$) and the required exposure ($n \times m$) was determined by n and r. In general a high ratio signals a large item selection space which allows more freedom in item selection.

Case 1 represents a situation where available exposure is rich relative to the required exposure. With 300 items in the item pool, the test length was set to 20 items to make the ratio of the pool size to test length 15, more than satisfying the Stocking's rule (Stocking, 1994) that the pool size should be 12 times the test length. The maximum exposure rate was set at 0.20, a value commonly used by large-scale testing programs (Way, 1998). The required exposure was $20m$, while the available item exposure was $(300)(0.20)m=60m$, three times the required. In Case 2, the item selection space was limited relative to Case 1 due to lengthened tests, with $n=32$ and $r=0.2$. Compared to Case 1, the available exposure remained the same ($60m$) but the required exposure increased to $32m$. The ratio between the available and required exposure was reduced to 1.875. In Case 3, item availability was restricted by a more stringent security criterion, with $n=20$ and $r=0.125$. Compared to Case 1, the required exposure remained the same ($20m$) but the available exposure decreased to $(300)(0.125)m=37.5m$. The ratio between available and required exposure was 1.875, the same as Case 2, but the item selection space was limited due to highly restricted exposure instead of longer tests.

The study involved nine test conditions resulting from combinations of three levels of practical constraints (no constraints, exposure control only, exposure control and content balancing), and the three cases of item selection space specified above. With respect to practical constraints, the no constraint condition bears little practical relevance but provides base comparisons among the various item selection procedures.

The exposure- control-only condition represents the situation where a high-stakes test measures a single ability with homogeneous items for which no content balance is necessary but security is of major concern. Although the SH procedure has been extended to accommodate exposure control conditional on additional elements (Stocking & Lewis, 1998; Davey & Parshall, 1995), the refined procedures involve much more time-consuming iterative simulations and result in further lessened efficiency, therefore they were not chosen for this study. Since the current study was primarily concerned with item selection methods, the general SH procedure would suffice.

In the conditions that demand both exposure control and content balancing, the modified multinomial model (MMM) (Chen & Ankenmann, 1999) was used for content balancing, since it was found to achieve best item usage among various content balancing methods compared in a simulation study (Leing et al., 2001).

The layout of the nine test conditions is shown below:

Table 1. Specifications of Test Conditions

	No Constraints	With exposure control	With exposure control and content balancing
Case 1 (n=20, r=0.20)	RAN, F, STR, USTR	F, STR, USTR	RAN, F, STR, USTR
Case 2 (n=32, r=0.20)	RAN, F, STR, USTR	F, STR, USTR	RAN, F, STR, USTR
Case 3 (n=20, r=0.125)	RAN, F, STR, USTR	F, STR, USTR	RAN, F, STR, USTR

CATs were simulated using each of the three non-random item selection procedures (F, STR, USTR) under each of nine conditions. RAN was simulated under conditions except those requiring exposure control only, since random item selection itself can be considered an exposure control procedure that results in almost equalized exposure rates. Simulations under all the conditions were done for an overall sample of 3000 *q*s distributed as N(0,1), as well as for conditional samples of 1000 *q*s at each of eleven points equally spaced on the *q* scale, ranging form –2.5 to 2.5 in increments of 0.5. The three-parameter logistic IRT model (3PLM) (Lord, 1980; Lord & Novick, 1968) was used for item response generation. The *expected a posterior* (EAP) estimation (Bock & Aitkin, 1981) with a normal prior N(0,1) was used for ability estimation in the beginning of a test; once a response pattern scorable by MLE was obtained, MLE was used for the remainder of the test.

A real item pool consisting of 300 mathematical items from a large-scale assessment was used, with items classified into three major content areas. 120 items (40% of the pool) were from content area 1, and 90 items (30% of the pool) each from content areas 2 and 3. For CATs simulated with content constraints, the item selected were required to reflect the same content proportions. Table 2 shows the summary statistics of the item parameters for the entire pool and by content category.

Given correlated *a* and *b* values (.356, significant at .0001 level) for the items in the pool, the pool stratification must take into account both item *a* and *b* values. Two types of stratified item pools were prepared for conditions with or without content balancing, all with four strata of 75 items each. For conditions requiring exposure control only, the pool was stratified through a two-stage process, *b* blocking followed by *a* stratification (Chang et. al ,

2000), which resulted in *b* values similarly distributed in each stratum, with average *a* values increasing across strata. For the conditions requiring content balance, the item pool was stratified by item *a* and *b* values and content indices through a tri-stage process (Yi & Chang, 2001). Specifically, the pool was first stratified into blocks based on content area; then the two-stage stratification was performed within each content block. Finally the corresponding strata across content blocks were collapsed into a single stratum. The pool stratified through the tri-stage process had content coverage within each stratum that resembled the entire pool, while the pool stratified with only respect to *a* and *b* values did not have this quality. Summary statistics of item parameters for both types of item pools appear in Table 3.

Item selection

To allow an initial success experience for most examines, each examine was assigned an initial ability estimate of –1. For F, the first item was randomly selected from among the 10 items most informative at ability level –1; for STR and USTR, the first item was randomly selected from among the 10 items in the first stratum with *b* values most closely matching the *q* value -1. Random selection from 10 items for the first item was intended to eliminate the similar item sequences across examinees early in the test. This strategy for selecting the first item was used in one study of the stratified design (Yi, 2002) and differed from the procedure used in other studies (e.g, Chang & Ying, 1999; Chang, Qian & Ying, 2000; Leung, Chang & Hau, 2001), which used three artificial items to roughly "locate" each examinee on the ability continuum for the test to begin. The use of artificial items is unrealistic in that operational CATs never begin with several items that do not count toward the total test score.

After the first item administration, the ability estimate was updated, and the second and the subsequent items were selected, one at a time, according to maximum information criterion for F, and by matching *b* and *q* values for STR and USTR. For STR, 4 and 8 items were selected form each stratum for conditions with *n*=20 and *n*=32, respectively. For USTR, 3,5,6,7 items were selected from the four strata for conditions with *n*=20, and 5,7,9,11 items for conditions with *n*=32. For CATs simulated with RAN, items were randomly selected from the item pool at each provisional ability estimate, without targeting item attributes to ability estimates at all.

For conditions with exposure control, the use of SH requires a preliminary simulation phase to obtain exposure control parameters for all items in the pool. For each of the CAT design conditions with specified test length, maximum exposure rate, item selection procedure, and examinee sample (conditional or overall), a unique set of exposure control parameters was derived through iterative simulations. The results from the final round of the simulation were taken as the exposure control parameters for operational CAT administrations.

For conditions with both exposure control and content balance, the MMM content balancing procedure was incorporated into item selection in addition to the exposure control procedure SH. The MMM procedure forms a multinomial distribution from the target item proportions for the content areas (0.4, 0.3, and 0.3), and identifies a content area for the to-be-selected item by drawing a random number from U(0,1). Depending on where the random number fell in the distribution, a content area was identified, and an item was selected from this content area based on the appropriate item selection criterion. Whenever a target proportion was reached for a content area, it was dropped from the remainder of the test and

the rest of the target proportions were divided by their sum to form a new multinomial distribution. The item selection proceeded in the same manner until the target proportion was satisfied for each content area.

Simulation procedure

Item exposure control parameters were derived for each CAT design condition through separate iterative simulations. The SH procedure is based on the probability relationship $P(A)=P(A|S)P(S)$, where $P(A)$ is the probability that an item is administered, $P(S)$ is the probability that the item is selected by the CAT algorithm, and $P(A|S)$ is the conditional probability that the item is administered given that it has been selected. The SH procedure attempts to control item exposure through assigning each item an exposure control parameter k_i, which is the finalized value of $P(A_i|S_i)$ from the last round of the iterative simulations. For the condition with n=20 and r=.2, items selected using USTR with exposure and content constraints, and 1000 examinees with the same ability θ=1.0, item selection occurred within the pool stratified through the tri-stage process, and the steps for obtaining exposure control parameters were as follows:

Step 1. Set initial values $P(A|S)$=1.0 for all items in the pool.

Step 2. Simulate CATs for the 1000 examinees with θ=1.0. The first item was randomly selected from among the 10 best items targeted at an ability level of -1 from the first stratum. For the second item and on, items were selected based on content specification, exposure control algorithm, and the criterion of matching b with θ. For the 20-item test, 3,4,6, and 7 items were selected from each of the four strata, respectively. Each time to select an item, a random number sampled from U(0,1) was used to locate the appropriate content area in the current stratum, and an item i with b_i most closely matching the current θ estimate

was selected. Given the selected item *i*, another random number *u* was sampled from U(0,1) and compared to $P(A_i|S_i)$. Item *i* was administered only if $u \leq P(A_i|S_i)$, otherwise the next best item from the same content area in the same stratum was evaluated. Item *i* was dropped for the remainder of the test no matter if it was administered or not. The procedure was repeated until an item was administered. After all CATs were finished, the frequency of selection $P(S_i)$ and the frequency of administration $P(A_i)$ were computed for all items in the pool.

Step 3. Adjust $P(A_i|S_i)$ according to the prespecified maximum exposure rate, r=.2.

If $P(S_i) > r$, then $P(A_i|S_i) = r / P(S_i)$;

If $P(S_i) \leq r$, then $P(A_i|S_i) = 1.0$.

The resulting new $P(A|S)$s were sorted. To ensure a complete test of 20 items with appropriate content coverage for each examinee, the 8 largest $P(A_i|S_i)$s from content area one, and the 6 largest $P(A_i|S_i)$s from each of content areas two and three were set to 1.0.

Step 4. Repeat Steps 2 and 3 until all $P(A_i|S_i)$ have stabilized and the maximum $P(A)$ value approximately equals r. The $P(A_i|S_i)$ values from the final iteration were the exposure control parameters (k_i) to be used in operational CAT administrations.

The operational CAT simulations proceeded in much the same way as the preliminary simulations except that no adjustments of parameters were needed and all simulations occurred in one cycle that resulted in a final ability estimate for each examinee.

The above procedure for simulating CATs also applies to the STR conditions, except that unequal numbers of items were to be selected from each stratum. For the F conditions, the simulation procedure differed from the above in three ways: (1) it did not require a stratified pool, (2) it did not involve multiple stages of testing, and (3) it adopted a maximum information criterion for item selection.

Evaluation criteria

To evaluate test efficiency, bias, standard error of measurement (SEM) and root mean square error (RMSE) of ability estimates were examined for the overall sample, and RMSE was also evaluated for each conditional sample.

To evaluate item usage, observed item exposure rates were graphed and examined for an overall impression of item usage balance. A $?^2$ statistic, defined as follows, was computed as a measure of the discrepancy of the observed and the uniform exposure rates:

$$c^2 = \sum_{j=1}^{N} (er_j - \overline{er_j})^2 \Big/ \overline{er_j}$$

where er_j is the observed exposure rate, and $\overline{er_j} = n/N$ is the uniform exposure rate with n being the test length and N being the pool size.

The observed maximum item exposure rate for all items in the pool was taken as an indication of the overall security achieved (Stocking & Lewis, 1995); the number of items never used for the overall sample was reported as a measure of item pool utilization efficiency. Test overlap rate is another important summary index for measuring exposure control. For fixed length n-item CATs and for a sample of m examinees, the calculation of test overlap rates involves the following steps: (1) counting the number of common items for each of the $m(m-1)/2$ pairs of examinees, (2) summing all the $m(m-1)/2$ counts, and (3) dividing the total counts by $nm(m-1)/2$. To ensure test security, the number of common items shared by two randomly sampled examinees should be minimized. For the current study, the test-retest overlap rates were obtained for each item selection procedure conditional on ability level with CATs simulated for 1000 examinees (m=1000), while the peer-to-peer

overlap rates were obtained with CATs simulated for the overall sample of 3000 examinees (*m*=3000).

Results

The global comparison of the various item selection procedures was based on CATs simulated for the overall sample of 3000 examinees with abilities distributed as N (0,1), while the comparison at specific ability levels was based on conditional samples of 1000 examinees at each of the eleven ability points. It should be noted that the global and conditional results were based on independent simulations. In addition, for conditions with exposure control, the CAT simulations for the global and conditional examinee samples each involved unique sets of exposure control parameters. For these reasons, the global and conditional results have no direct correspondence to each other, but they provide complementary information about the comparative performance of each procedure.

No matter whether observed marginally or conditionally, RAN was found to consistently result in the lowest test efficiency and best item usage across test conditions. Completely randomized item selection achieves a natural balance of item usage by generating item exposure rates that closely matched the uniform distribution, and yielded minimum overlap rates across examinees. This procedure, however, showed the poorest performance on test efficiency due to not matching item difficulty to ability estimates. Since RAN forfeited the basic property of a CAT, its inclusion in this study was simply for baseline comparison. Summarized below are the comparative performances of the three non-random procedures, F, STR, and USTR, organized by conditions of practical constraints.

Without constraints

From Tables 4 - 6, it was observed that for CATs simulated for the overall sample of examinees under non-constrained conditions, F resulted in lower SEM (equivalently, RMSE, since $RMSE^2=Bias^2+SEM^2$ and bias values were small across all conditions) than STR and USTR, and thus F had higher efficiency. However, F had dramatically skewed item exposure rates (see Figures 1-3), poor utilization of the item pool (many unused items), and high overlap rates. USTR showed slightly improved efficiency over STR (see SEM), with a larger proportion of unused items. This pattern of results transcends different combinations of test lengths and target maximum exposure rates. The efficiency advantage of F was clearly seen when items were selected based solely on statistical properties.

A similar pattern of results was also observed for conditional samples (see Figures 4-9). F had lower RMAE, thus higher efficiency than STR and USTR along the ability scale, but it had high test-retest overlap rates at all ability levels. USTR showed slightly improved efficiency over STR at all ability levels except the low extreme, and had slightly higher overlap rates than STR along the ability scale.

With exposure control

For CATs simulated with respect to the overall sample, the relative performance of F, STR, and USTR varied with test length and target exposure rate (see Tables 4-6). As shown in Table 4, when available exposure was rich relative to the required, F showed somewhat higher efficiency yet poorer item usage than STR and USTR; USTR had slightly improved efficiency over STR with slightly compromised item usage. When the item selection space was restricted by either lengthened tests or a lower target maximum exposure rate, as shown in Tables 5 and 6, however, poor item usage for F no longer led to an efficiency gain in

return. While USTR yielded slightly higher efficiency than STR with slightly compromised item usage, it approached the same level of efficiency as F. It seemed that under conditions with restricted item availability, USTR was a good alternative to F since it resulted in a similar level of efficiency with better item usage.

Observed conditionally, the relative performance of F, STR, and USTR also showed different patterns across conditions of different test lengths and target maximum exposure rates (see Figures 4-9). In Case 1 where item availability was relatively large, the three procedures showed a similar trend as those observed for the unconstrained condition, although the differences among the three procedures were quite small. F showed the highest efficiency and poorest item usage across ability levels, while STR showed the opposite; at all ability levels except the low extreme, USTR had slightly higher efficiency than STR with slightly inflated overlap rates. In Case 2 and Case 3 where the item selection space was relatively restricted, all three procedures performed similarly in terms of efficiency along the ability scale, while F had higher overlap rates at the middle ability levels than STR and USTR. The results from conditional samples were not completely consistent to those observed from the overall sample, partly because the item pool was poorly targeted to ability levels beyond the middle portion of the scale, while the configuration of the *b* values approximated reasonably well the overall ability distribution N(0,1). Had sufficient number of items appropriate for all ability levels been included in the item pool, more distinction might have been observed across various procedures on the conditional samples.

With exposure control and content balancing

With respect to the overall sample, when item selection involved both exposure control and content balancing, the comparative performance of various procedures very much

resembled the pattern observed for the condition with only exposure control, except for the situation with very stringent exposure control (see Tables 4-6). Specifically, with a relatively large item selection space in Case 1, the use of F resulted in efficiency gain over the STR and USTR, with poor item usage; USTR slightly improved efficiency over STR with small compromises on item usage. When the item selection space was limited by lengthened tests as examined in Case 2, F lost its efficiency gain. STR still had the lowest efficiency and best item usage. The use of USTR seemed to be a viable option since it achieved a similar efficiency level as F with improved item usage. Under the condition where more stringent exposure control was imposed along with content balancing, USTR lost its efficiency gain over STR; the two procedures achieved the same level of efficiency with STR showing better item usage. Relative to the stratification procedures, F retained its efficiency advantage, with still problematic item usage.

When examined conditionally on ability levels, the patterns of relative performance of the three procedures were in general similar to those observed under the conditions with only exposure control. Specifically, in Case 1 where item availability was relatively large, F showed slightly higher efficiency and poorer item usage than the stratification methods along the ability scale; USTR had slightly higher efficiency than STR with slightly inflated overlap rates at the middle ability levels. In Case 2 and Case 3 where the item selection space was relatively restricted, all three procedures performed similarly in terms of efficiency along the ability scale, while F showed higher overlap rates at the middle ability levels than STR and USTR.

To summarize, the results showed that USTR reduced error variances for STR under a variety of test conditions, with small compromises in item usage. F had an apparent

efficiency advantage over STR and USTR only when item selection occurred with no constraints in a relatively large selection space. Compared to F, USTR enhanced the balance of item exposure, reduced overlap rate, and made better utilization of the entire pool, while achieving comparable test efficiency, especially when items were selected under exposure control and the item selection space was restricted by long tests or a stringent security criterion. It should also be noted that, if item usage is of major concern, STR seems to be the best procedure since it always outperformed the other procedures on item usage balance and pool utilization.

With respect to the effects of constraints, it was found that exposure control had a general tendency to dramatically decrease efficiency for F, yet had minimal impact on efficiency for STR and USTR, which implies that USTR had large potential for achieving efficiency comparable to F under exposure constrained conditions. On the other hand, the content constraints were found to decrease efficiency for all non-random procedures, including STR and USTR. In addition, a noteworthy finding was that under the condition where item availability was severely limited by very stringent exposure control along with content constraints, USTR failed to realize its efficiency potential. Finally, the results based on conditional indices highlighted the importance of including an adequate number of items appropriate for the ability range the test was intended to measure.

Discussion

The results from the current study confirmed the preliminary findings from Deng & Chang (2001), and extended the work by examining the three procedures (F, STR, and USTR) along with a baseline RAN, both conditionally and unconditionally, under

systematically varied conditions. The results from this study have implications for choosing an appropriate item selection procedure in applied settings.

The efficiency advantage of F was evident when items were selected without any constraints, but its high efficiency came at the price of very poor item usage. In addition to a large proportion of unused items that could cause an economic concern, overexposed items, mostly high *a* items, compromised test security and validity. To maintain an item pool that supports effective CATs, there must be a substantial number of high *a* items continuously input into the pool, while at the same time those items must satisfy test specifications and other constraints. That is a very difficult task for item writers, which even if achievable, may not be economically feasible at all. It is likely that few operational CAT programs could afford the use of maximum information item selection without any constraints.

For CAT programs that implement exposure control in item selection, the results suggest that the choice of an item selection procedure should be based on the relationship between the available and required item exposure. When the item pool size, test length and security criterion permit a large item selection space, as in Case 1 of the current study, the use of F resulted in an efficiency gain over the stratification procedures, yet the efficiency gain was accompanied by unbalanced exposure rates, high overlap rates, and a large number of unused items. When the item selection space was restricted by lengthened tests or more stringent exposure control, USTR seemed to be a promising alternative to F, since it to a large extent overcame the item usage problems with F while at the same time maintained its efficiency level. In addition, the results showed that the stratification procedures had stable efficiency levels no matter if item exposure was constrained or not; on the other hand,

exposure control consistently led to substantial efficiency loss for F, and the more stringent the exposure control, the greater the loss.

For CAT programs that implement content constraints as well as exposure control, the results suggest that the choice of an item selection procedure requires careful examination of the relationship between available and required item exposure. Unlike exposure control that only influenced test efficiency for F, content constraints tended to decrease efficiency for all non-random procedures, including STR and USTR. When the item selection space is limited by stringent exposure control, imposing content constraints may further decrease item selection space to a point where USTR fails to gain efficiency over STR.

In reality, any CAT operation necessitates compromises among competing goals. This is clearly evident in the choice of an item selection procedure. The results from this study provide some tentative guidelines for choosing an item selection procedure in CAT and should encourage further studies using different item pools and CAT design structures. The results also highlight the importance of viewing the merits and shortcomings of each item selection procedure in relation to important practical concerns and CAT design conditions, and emphasize the necessity of choosing a procedure based on informed judgments from balancing many factors in a manner that best satisfies the needs of the testing program.

Table 2. Descriptive Statistics of the Item Parameters by Content Category

Item Parameters	N	Mean	SD	Minimum	Maximum	Skewness	Kurtosis
			Content 1				
a	120	0.923	0.281	0.359	1.635	0.345	-0.374
b	120	-0.001	1.048	-2.515	2.005	-0.227	-0.759
c	120	0.148	0.042	0.031	0.251	-0.039	0.140
			Content 2				
a	90	1.012	0.305	0.349	1.605	-0.159	-0.915
b	90	0.587	0.847	-1.977	2.582	-0.460	0.122
c	90	0.156	0.049	0.042	0.271	-0.140	-0.512
			Content 3				
a	90	1.037	0.290	0.296	1.933	0.210	0.399
b	90	0.442	0.785	-1.809	1.971	-0.473	0.344
c	90	0.141	0.048	0.037	0.267	0.102	-0.324
			The Entire Pool				
a	300	0.984	0.295	0.296	1.933	0.147	-0.419
b	300	0.308	0.949	-2.515	2.582	-0.480	-0.191
c	300	0.148	0.046	0.031	0.271	-0.017	-0.282

Table 3. Descriptive Statistics of the Item Parameters by Stratum

Stratum	Item Parameters	N	Mean	SD	Minimum	Maximum	Skewness	Kurtosis
Stratified by *a*, *b*								
1	a	75	0.696	0.184	0.296	1.114	0.033	-0.423
	b	75	0.312	0.972	-2.348	2.582	-0.430	-0.001
	c	75	0.160	0.046	0.042	0.267	-0.357	0.030
2	a	75	0.911	0.202	0.487	1.407	0.146	-0.707
	b	75	0.303	0.967	-2.515	2.199	-0.566	0.052
	c	75	0.151	0.040	0.073	0.252	0.258	-0.013
3	a	75	1.072	0.205	0.493	1.505	-0.264	-0.077
	b	75	0.308	0.939	-2.132	2.005	-0.479	-0.354
	c	75	0.145	0.053	0.031	0.271	0.145	-0.656
4	a	75	1.259	0.247	0.584	1.933	-0.262	0.284
	b	75	0.310	0.937	-2.185	1.971	-0.474	-0.311
	c	75	0.137	0.043	0.037	0.251	-0.086	-0.067
Stratified by *a*, *b* & content								
1	a	75	0.705	0.182	0.296	1.084	-0.078	-0.353
	b	75	0.308	0.971	-2.348	2.582	-0.451	0.009
	c	75	0.157	0.043	0.042	0.267	-0.027	0.125
2	a	75	0.893	0.195	0.487	1.420	0.102	-0.174
	b	75	0.313	0.949	-2.515	2.199	-0.537	0.070
	c	75	0.159	0.043	0.062	0.240	-0.437	-0.395
3	a	75	1.061	0.197	0.493	1.505	-0.451	0.244
	b	75	0.295	0.965	-2.132	2.005	-0.501	-0.295
	c	75	0.143	0.047	0.055	0.271	0.383	-0.234
4	a	75	1.278	0.244	0.584	1.933	-0.612	1.106
	b	75	0.317	0.931	-2.185	1.871	-0.460	-0.371
	c	75	0.135	0.049	0.031	0.251	0.155	0.089

Table 4. Performance Summary for Various Methods under Case 1 Conditions

	n=20, *r*=.20					
	Bias	SEM	RMSE	χ^2	# Zero Exposure	Overlap Rate
Without Constraints						
F	-0.004	0.297	0.297	71.456	177	0.305
STR	0.007	0.369	0.369	9.767	11	0.099
USTR	0.003	0.350	0.350	9.285	33	0.097
RAN	-0.094	0.689	0.695	0.112	0	0.067
With Exposure Control						
F-SH	-0.008	0.331	0.331	32.242	139	0.174
STR-SH	-0.005	0.373	0.373	8.158	11	0.094
USTR-SH	-0.004	0.351	0.351	8.945	34	0.096
With Exposure Control and Content Control						
F-SH-C	-0.005	0.346	0.346	31.591	120	0.172
STR-SH-C	-0.004	0.387	0.387	9.293	9	0.097
USTR-SH-C	0.002	0.369	0.369	9.710	29	0.099
RAN-C	-0.084	0.686	0.691	0.093	0	0.067

Table 5. Performance Summary for Various Methods under Case 2 Conditions

	n=32, *r*=.20					
	Bias	SEM	RMSE	χ^2	# Zero Exposure	Overlap Rate
Without Constraints						
F	-0.001	0.245	0.245	66.054	121	0.327
STR	0.003	0.290	0.290	10.244	0	0.141
USTR	0.000	0.278	0.278	11.228	11	0.144
RAN	-0.062	0.528	0.532	0.092	0	0.107
With Exposure Control						
F-SH	-0.005	0.271	0.271	22.740	75	0.182
STR-SH	-0.002	0.281	0.281	6.214	0	0.127
USTR-SH	0.010	0.270	0.270	9.922	11	0.140
With Exposure Control and Content Control						
F-SH-C	-0.006	0.285	0.285	22.387	64	0.181
STR-SH-C	0.002	0.307	0.307	7.568	0	0.132
USTR-SH-C	-0.007	0.289	0.289	10.161	9	0.140
RAN-C	-0.045	0.518	0.520	0.185	0	0.107

Table 6. Performance Summary for Various Methods under Case 3 Conditions

	n=20, *r*=.125					
	Bias	SEM	RMSE	χ^2	# Zero Exposure	Overlap Rate
Without Constraints						
F	-0.004	0.294	0.294	71.012	177	0.303
STR	0.002	0.372	0.372	9.838	11	0.099
USTR	0.013	0.351	0.351	9.405	33	0.096
RAN	-0.066	0.695	0.698	0.097	0	0.067
With Exposure Control						
F-SH	-0.007	0.348	0.348	15.231	96	0.117
STR-SH	0.013	0.369	0.369	5.395	11	0.084
USTR-SH	-0.002	0.353	0.353	7.813	34	0.092
With Exposure Control and Content Control						
F-SH-C	-0.021	0.373	0.374	14.912	86	0.116
STR-SH-C	-0.008	0.392	0.392	6.121	9	0.087
USTR-SH-C	-0.011	0.393	0.393	7.511	29	0.091
RAN-C	-0.077	0.687	0.692	0.085	0	0.067

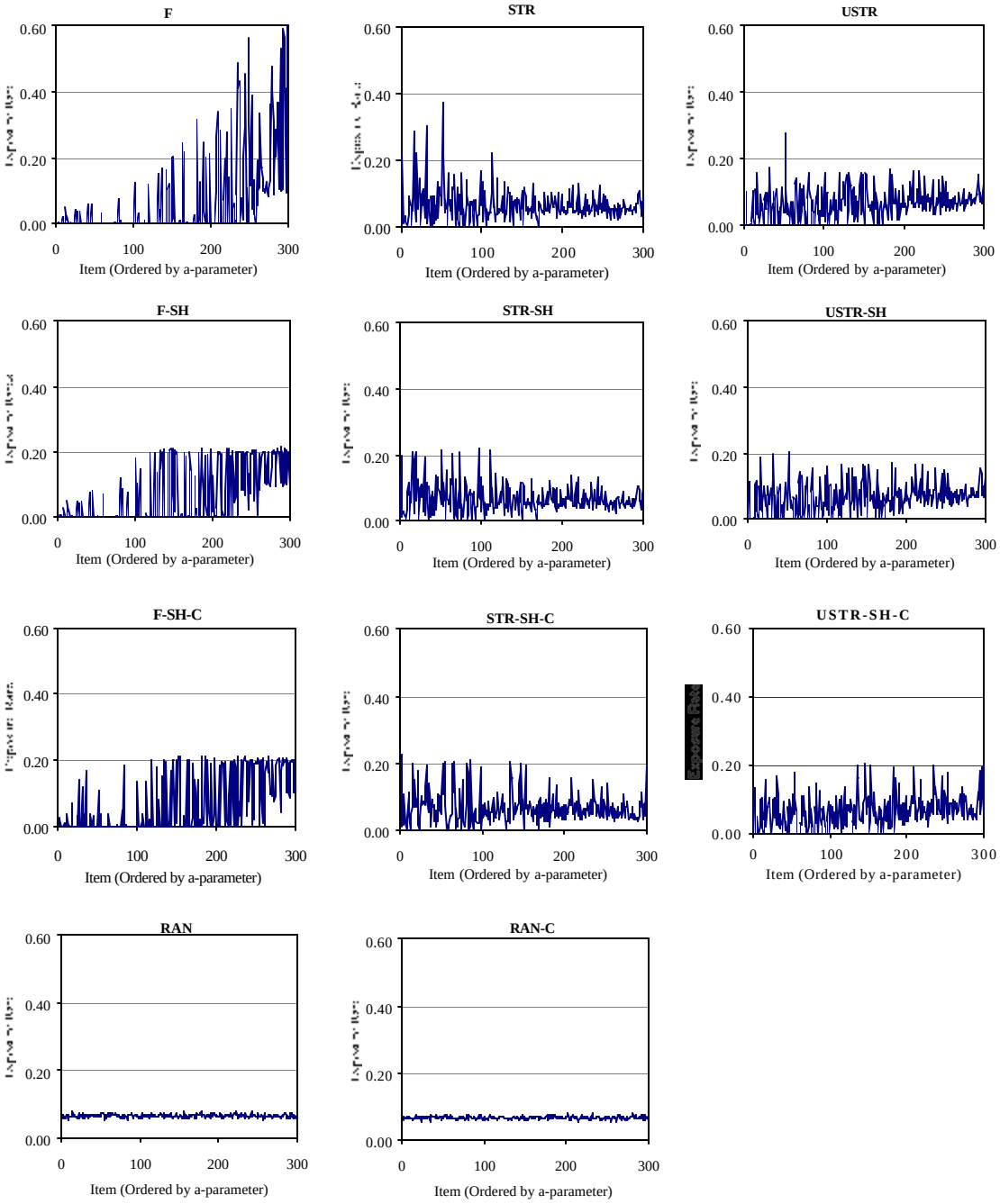

Figure 1. Item Exposure Rates for Various Methods in Case 1 (n=20, r=.20)

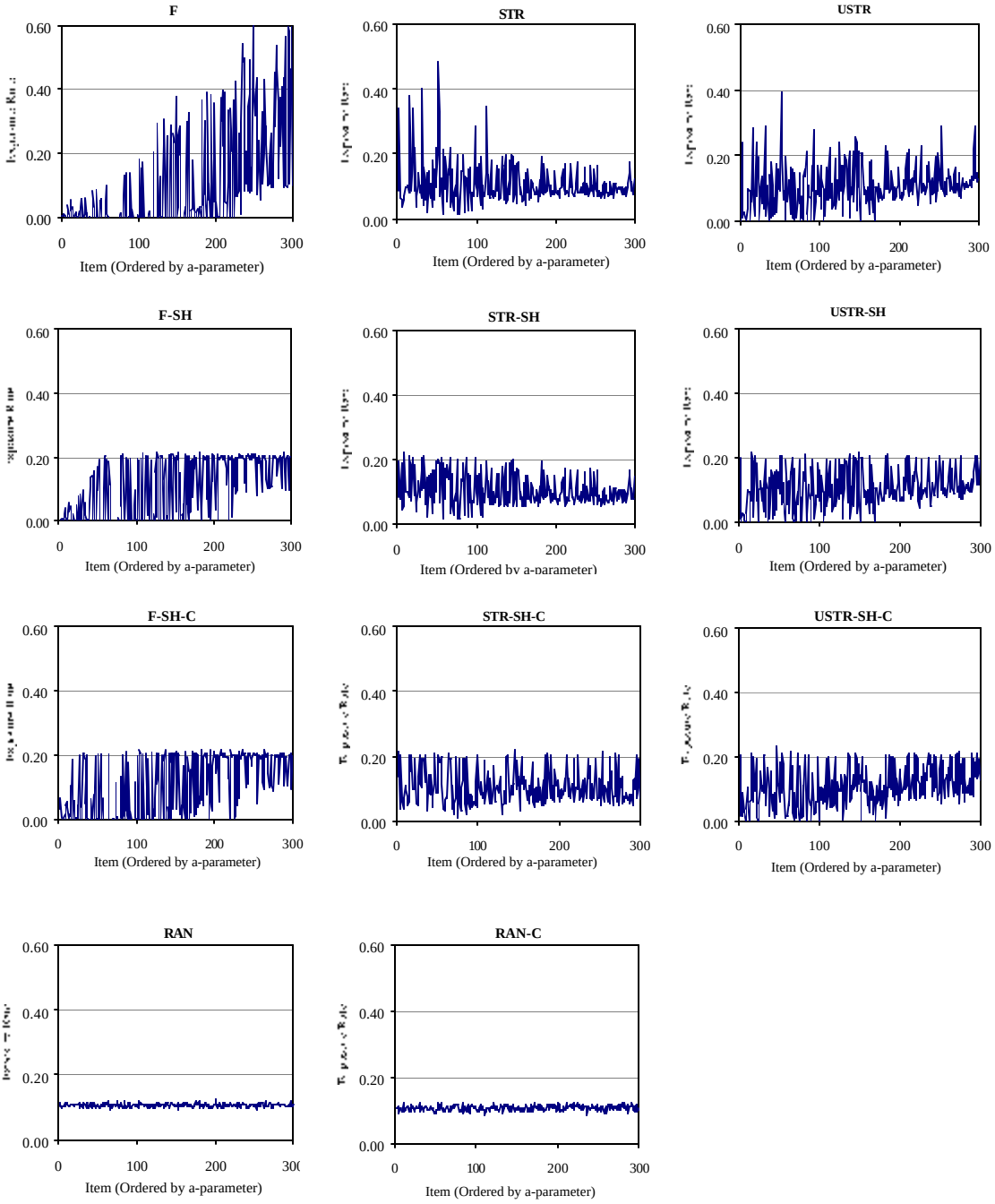

Figure 2. Item Exposure Rates for Various Methods in Case 2 (n=32, r=.20)

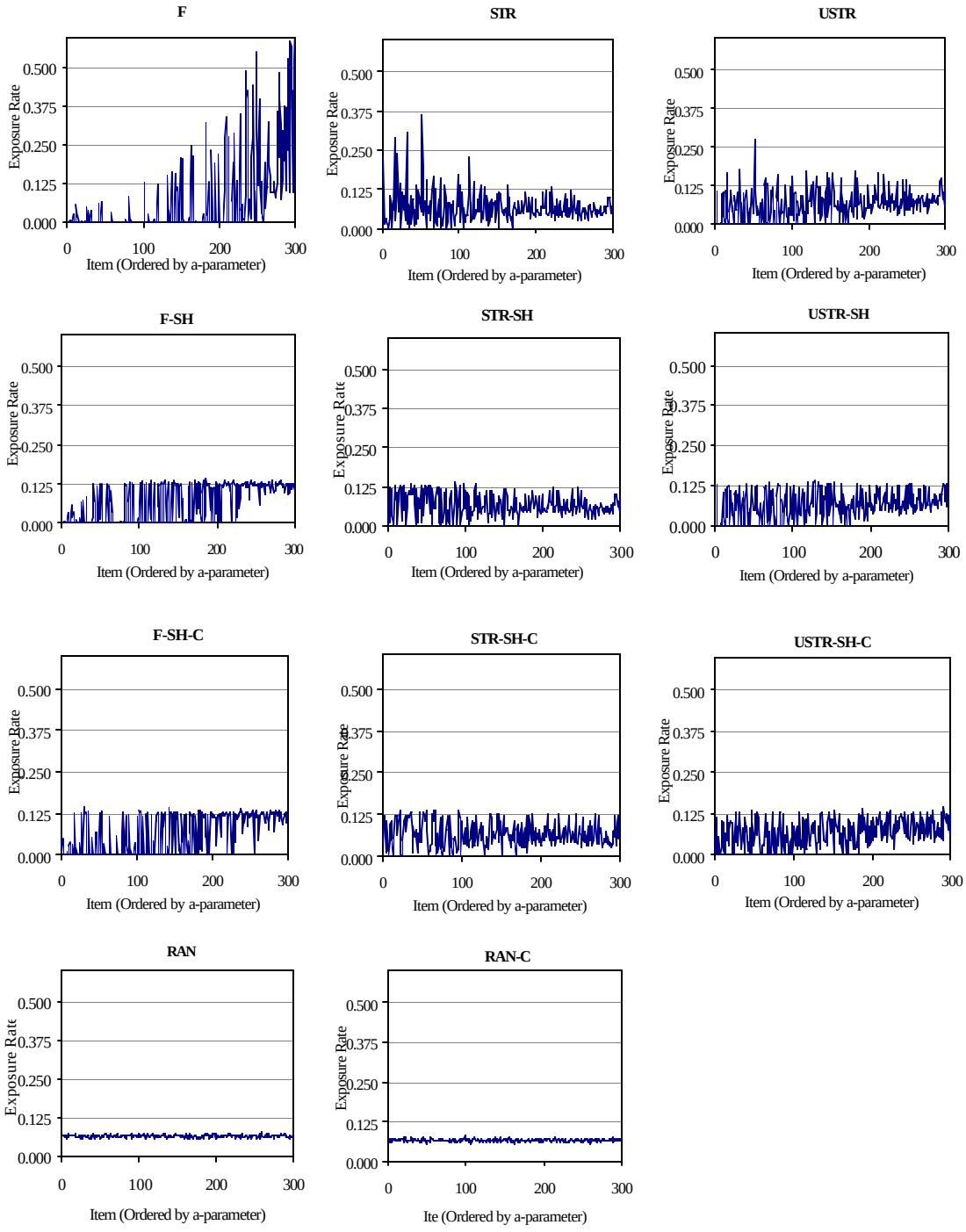

Figure 3. Item Exposure Rates for Various Methods in Case 3 (n =20, r =.125)

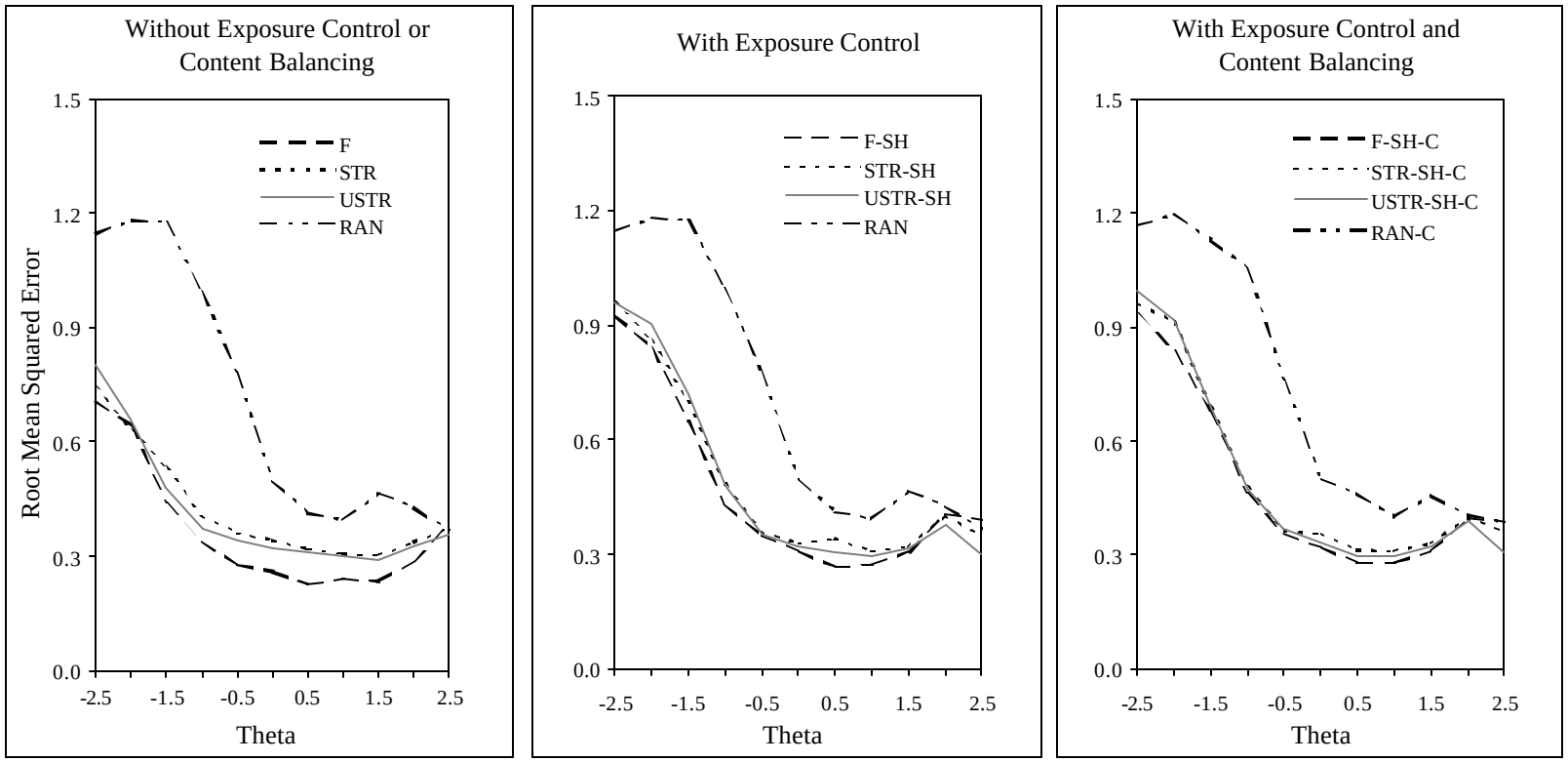

Figure 4. Conditional Root Mean Squared Errors for Variouos Methods under Case 1 Conditions (n =20, r =.20)

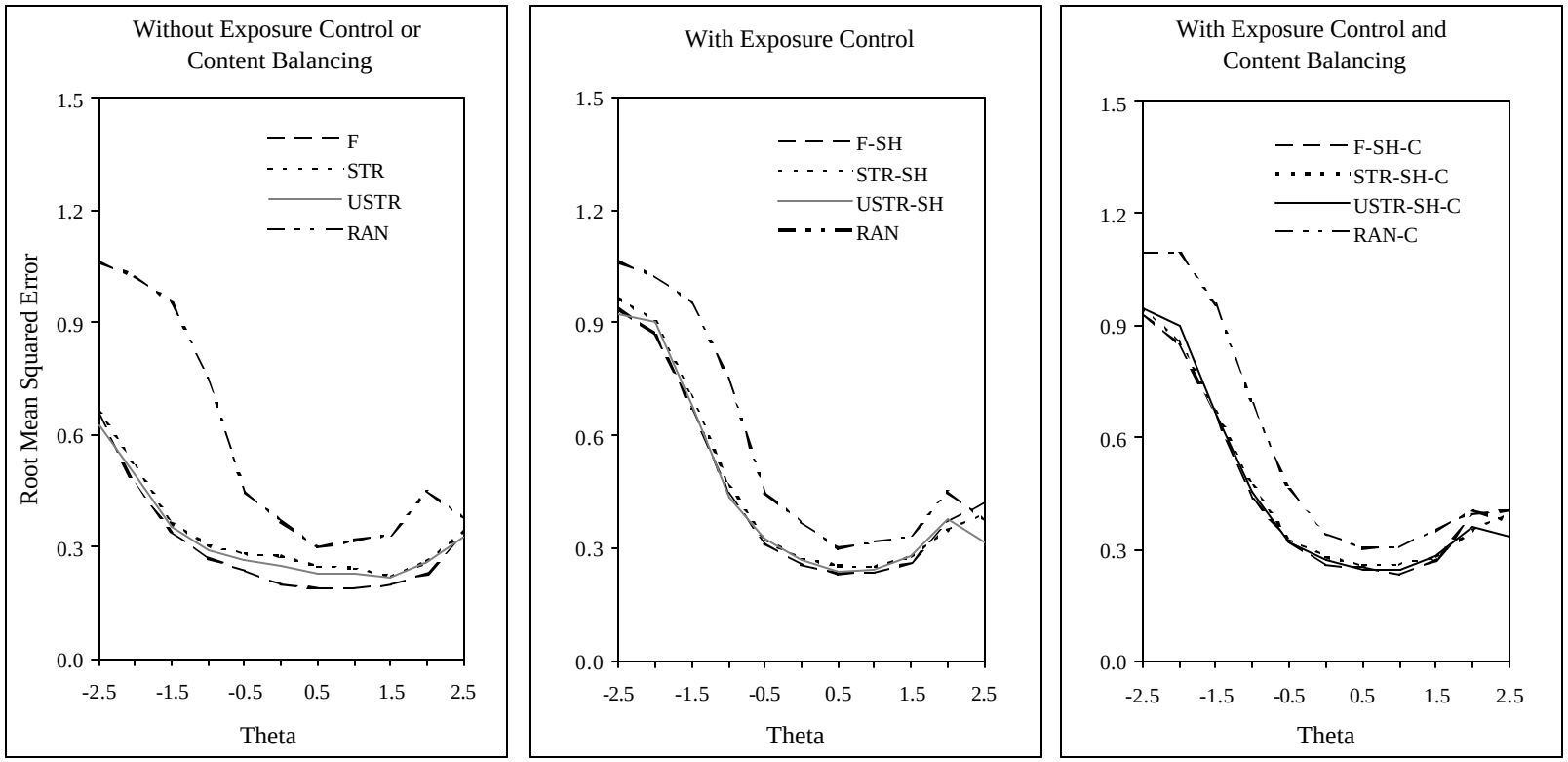

Figure 5. Conditional Root Mean Squared Errors for Variouos Methods under Case 2 Conditions (n =32, r =.20)

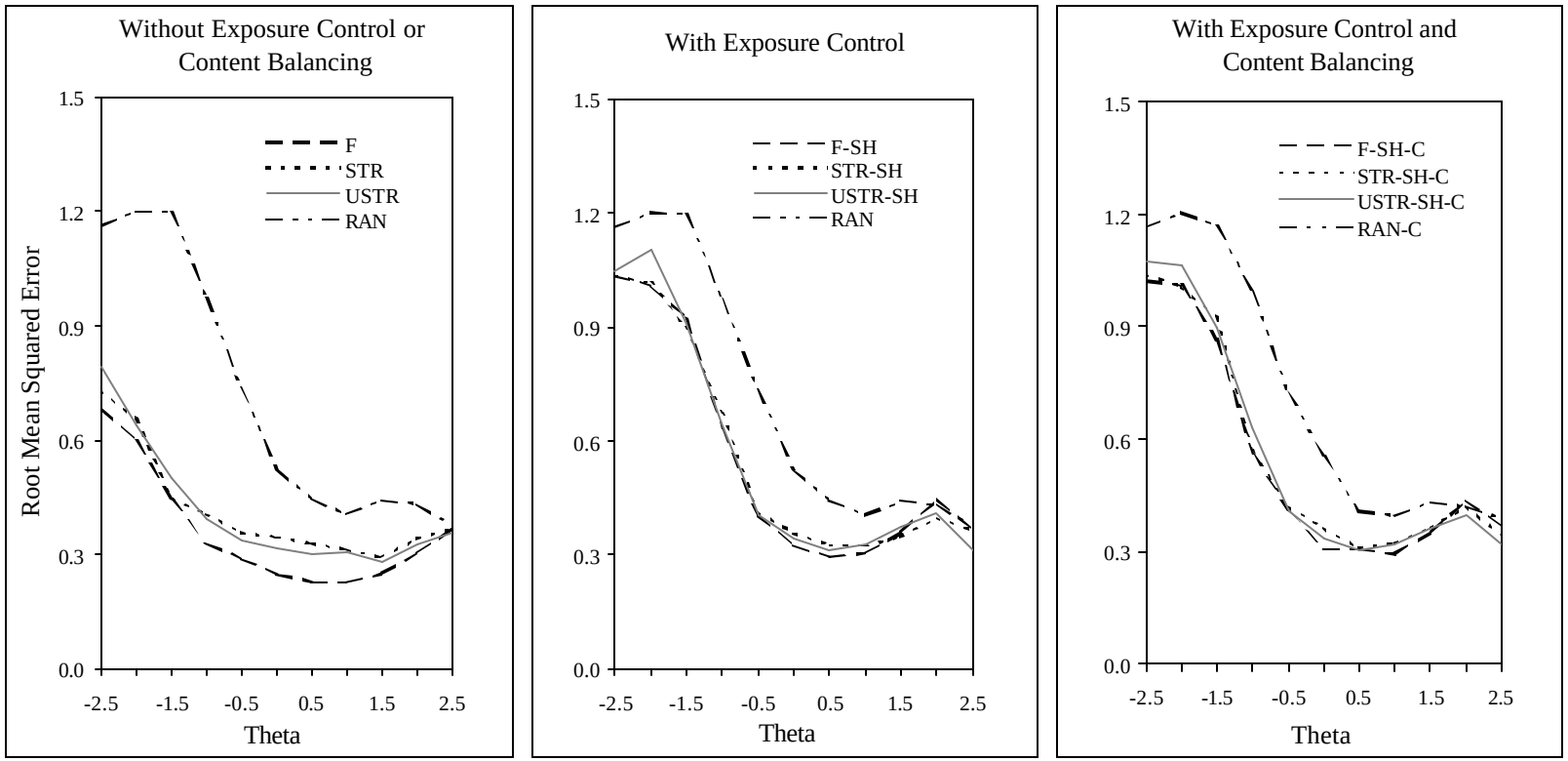

Figure 6. Conditional Root Mean Squared Errors for Variouos Methods under Case 3 Conditions (n =20, r =.125)

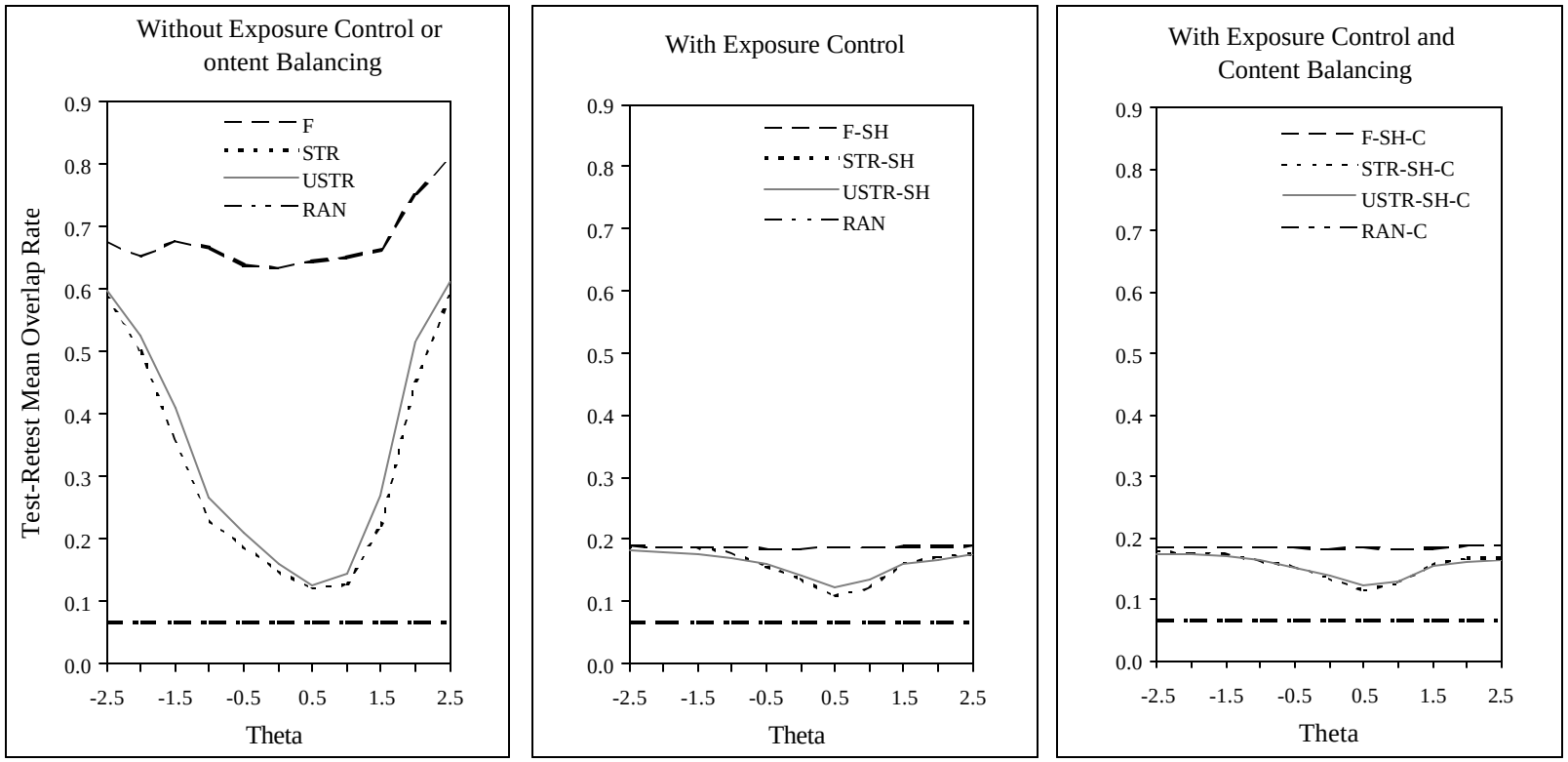

Figure 7. Conditional Overlap Rates for Various Methods under Case 1 Conditions (n =20, r =.20)

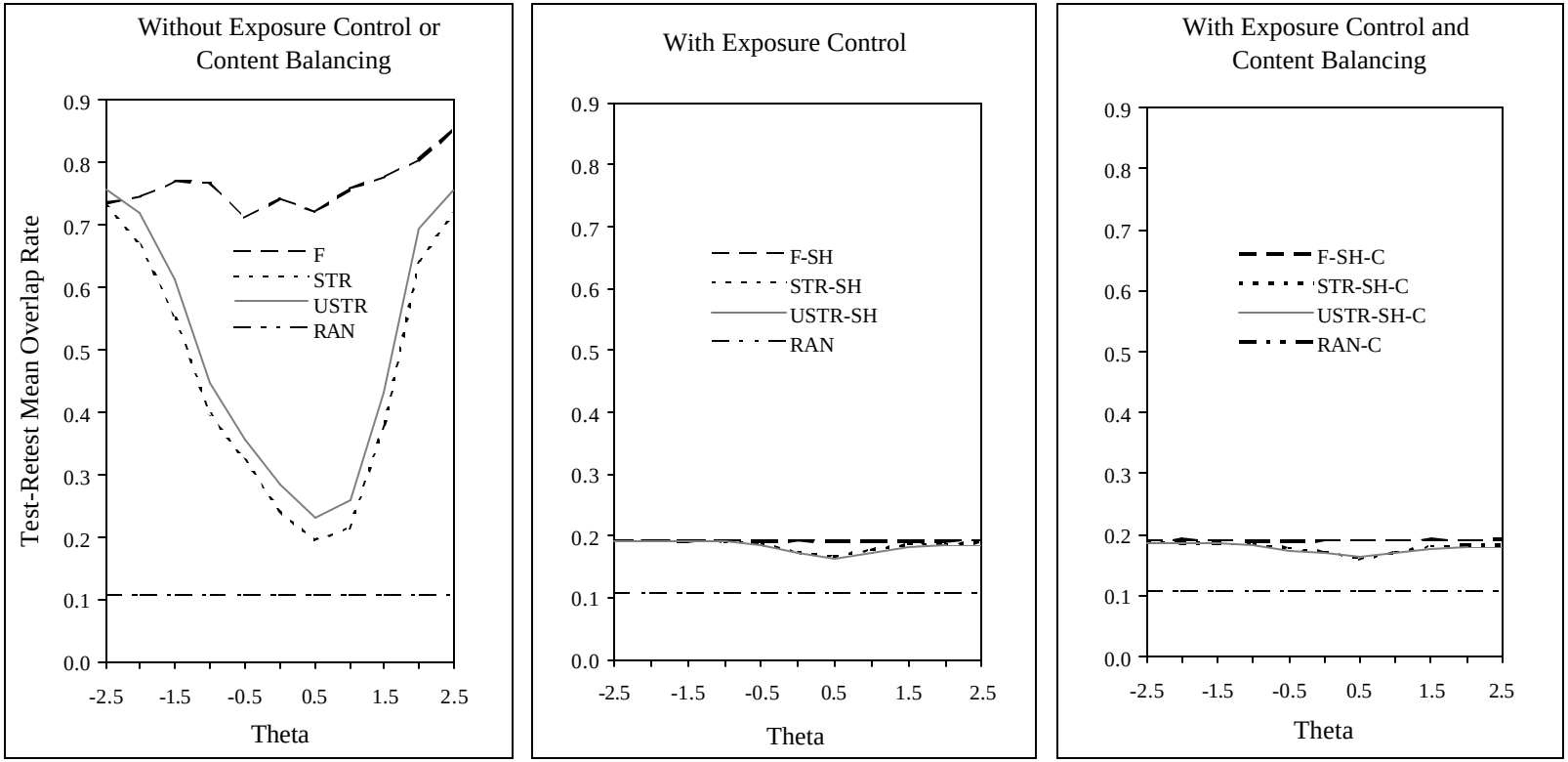

Figure 8. Conditional Overlap Rates for Various Methods under Case 2 Conditions (n =32, r =.20)

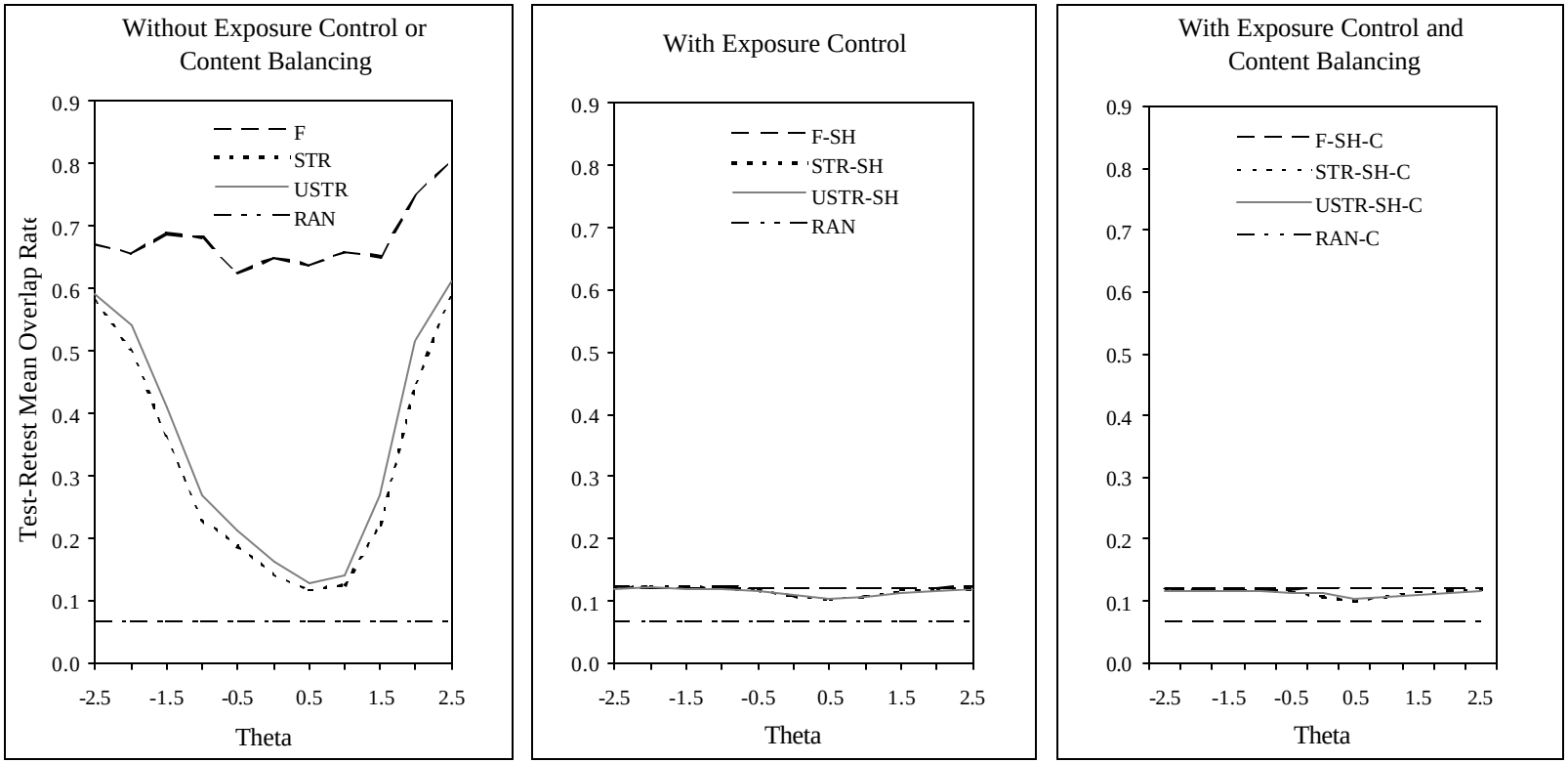

Figure 9. Conditional Overlap Rates for Various Methods under Case 3 Conditions (n =20, r =.125)

References

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. Psychometrika, 46, 443-459.

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

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

Chang, H., Qian, J., & Ying, Z. (2001). *a*-stratified multistage CAT with *b*-blocking. Applied Psychological Measurement, 25(4), 333-341.

Chen, S. Y., & Ankenmann, R. D. (1999). Effects of practical constraints on item selection rules at the early stages of computerized adaptive testing. Paper presented at the annual meeting of the American Educational Research Association, Montreal, Canada.

Davey, T., & Parshall, C. G. (1995). New algorithms for item selection and exposure control with computerized adaptive testing. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, USA.

Deng, H., & Chang, H. (2001). *a*-stratified computerized adaptive testing with unequal item exposure across strata. Paper presented at the annual meeting of the National Council of Measurement in Education, Seattle, USA.

Hau. K., & Chang, H. (1998). Item selection in computerized adaptive testing: Should more discriminating items be used first? Paper presented at the annual meeting of the American Educational Research Association, San Diego, USA.

Leung, C., Chang, H., & Hau, K. (1999). Item selection in computerized adaptive testing: improving the a-stratified design with the Sympson-Hetter algorithm. Paper presented at the annual meeting of the American Educational Research Association, Montreal, CA.

Leung, C., Chang, H., & Hau, K. (2001). 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 Research Association, Seattle, USA.

Lord , F. M. (1980). Application of item response theory to practical testing problems. Hillsdale, NJ: Erlbaum.

Lord, F. M. & Novick, M. R. (1968). Statistical theories of mental test scores. Reading, MA: Addison-Wesley.

Stocking M. L. (1994). Three practical issues for modern adaptive testing item pools (ETS Research Report No. 94-5). Princeton, NJ: Educational Testing Service.

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

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

Wainer, H. (1990). Computerized adaptive testing: A primer. Hillsdale, NJ: Lawrence Erlbaum Associates.

Way, W. D. (1998). Protecting the integrity of computerized testing item pools. Educational Measurement: Issues and practices, 17, 17-27.

Way, W. D., Steffen, M., & Anderson, G. S. (1998). Developing, maintaining, and renewing the intern inventory to support computer-based testing. Paper presented at the colloquium, Computer-based testing: building the foundation for future assessment, Philadelphia, PA.

Weiss, D. J., & Kingsbury, G. G. (1984). Application of computerized adaptive testing to educational problems. Journal of Educational Measurement, 21(4), 361-375.

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