Florida Journal of Educational Research
Fall 2000, Vol. 40, No. 1, 28 - 52.

Item Exposure Control in Computer-Adaptive Testing: The Use of Freezing to Augment Stratification

Cynthia Parshall
J. Christine Harmes
Jeffrey D. Kromrey
University of South Florida

Computerized adaptive tests are efficient because of their optimal item selection procedures that target maximally informative items at each estimated ability level. However, operational administration of these optimal CATs results in the administration of a relatively small subset of items with excessive frequency, while another portion of the item pool is almost unused. This situation both wastes a portion of the available items and is a security risk for testing programs that are available on more than a few scheduled test dates throughout the year. A number of exposure control methods have been developed to reduce this effect. In this study, we investigate the effectiveness of item "freezing" as a means of augmenting the Stratified-a method for exposure control. A second variation of the Stratified-a method investigated here concerns use of differing numbers of strata. Using Monte Carlo procedures, we examine these methods under varying conditions of freezing and number of strata. Results are reported in terms of pool usage and test precision, both unconditionally and conditionally on ability.

Computerized adaptive tests are efficient because they successively select items that provide optimal measurement at each examinee's estimated level of ability. However, when items are selected during a computerized adaptive test (CAT) based solely on their psychometric properties, certain items are found to be administered to nearly every

examinee, while other items remain almost unused. This both wastes a portion of the available items and, more importantly, it clearly presents a security risk for testing programs that are available on various occasions throughout the year. The concern is that frequently administered items will quickly become compromised and no longer provide valid measurement.

A number of exposure control methods have been developed to reduce this effect. The Sympson-Hetter method (Sympson & Hetter, 1985) was one of the earliest approaches to controlling item overexposure, and a number of adaptations of this method have been developed (Davey & Parshall, 1995; Nering, Davey & Thompson, 1998; Parshall, Davey, & Nering, 1998; Parshall, Kromrey, & Hogarty, 2000; Stocking & Lewis, 1995; Thomasson, 1995). In the Sympson-Hetter and related approaches to exposure control, a series of simulations is conducted to assign a unique *exposure parameter* to each item. This parameter is then used to probabilistically limit the frequency with which a selected item is administered. These methods have been found to be reasonably effective, but they can be cumbersome to implement. Furthermore, every time a change is made to the item pool (items are added or removed), the preparatory simulations must be conducted again.

A very different approach is taken in the Stratified-a method (Chang & Ying, 1997). No simulations or exposure parameters are used. Rather, the items in a pool are assigned to strata, based on their *a*-parameters (the Item Response Theory [IRT] estimate of the item's discriminatory power). The number of strata used, the *a*-value cut points that define the strata, and the number of test items drawn from each stratum must all be set in advance of

operational testing. Early in the test, items are administered from the stratum with the lowest *a*-parameters. As the test progresses, the strata with higher *a*-values are used. Within a stratum, the item that has the *b*-value (or IRT difficulty parameter) closest to the examinee's current estimate of theta is selected for administration. The rationale for this method is twofold. First, early in the test little information about the examinee's ability is available. It is most appropriate to use low-discrimination items at this point and items that are more highly discriminating later in the test in order to better pinpoint the examinee's ability. In addition, a maximum information item selection algorithm will typically lead to overexposure of the more highly discriminating items in the pool. This Stratified-a approach to item selection and exposure control is designed to yield much more balanced pool usage. While this method is logically appealing and simple to implement, extreme overuse of some items is still found under this method (Parshall, Kromrey, & Hogarty, 2000). An adaptation of the Stratified-*a* method that might address this problem is to temporarily render items unavailable for selection when they exceed a target administration rate – that is, to "freeze" these items in the selection algorithm until their administration rate drops below the target value.

Purpose

Although theoretically sound, the Sympson-Hetter is computationally complicated and logistically involved. The Stratified-a method, in contrast, is straightforward and easy to implement, but may provide exposure control to a lesser extent than the more complex methods. A variation of

the Stratified-*a* method that temporarily freezes items in the selection algorithm might address this weakness, while retaining the advantages of the method. Furthermore, there is little guidance in the literature on the effect of number of strata on the performance of the Stratified-a method. The purpose of the study was to empirically investigate controlled experimental variations of item freezing in conjunction with the Stratified-*a* method and number of strata, and to compare the levels of exposure control provided by the variations.

Method

The research was a Monte Carlo study in which adaptive testing was simulated under controlled conditions. In this study, the Stratifed-a method was modified in two specific ways. First, a *freeze* condition was investigated. Items that exceeded a target administration rate could be "frozen", or rendered temporarily unavailable for selection. As more tests are administered, this proportional administration rate for a frozen item could drop below the target rate again; at this point the frozen item would be "thawed", and once again be available for selection and use. There were two levels of this condition; one in which freezing was utilized and one in which it was not. In addition, the effects of item freezing were investigated across three levels of stratification of the item pool: four, six, and eight strata. Table 1 displays the number of strata used, and the number of items drawn from each stratum. The combinations of these variations resulted in six Stratified-*a* approaches.

Table 1
Item Pool Characteristics by Stratum

Four Strata	Number of items drawn from each stratum	Number of items in each stratum
AB (cutpoint = 0.793)	10	120
CD (cutpoint = 1.0)	10	118
EF (cutpoint = 1.23)	10	120
GH (cutpoint = 3.0)	10	122
Six Strata		
A (cutpoint = 0.65)	5	60
B (cutpoint = 0.793)	5	60
CD (cutpoint = 1.0)	10	118
EF (cutpoint = 1.23)	10	120
G (cutpoint = 1.46)	5	61
H (cutpoint = 3.0)	5	61
Eight Strata		
A (cutpoint = 0.65)	5	60
B (cutpoint = 0.793)	5	60
C (cutpoint = 0.88)	5	57
D (cutpoint = 1.0)	5	61
E (cutpoint=1.1)	5	58
F (cutpoint = 1.23)	5	62
G (cutpoint = 1.46)	5	61
H (cutpoint= 3.0)	5	61

The effectiveness of these six variations of the Stratified-*a* exposure control method were compared to the Sympson-Hetter and two additional "baseline" conditions (no control and completely random item selection). All nine methods were investigated at target maximum exposure rates of .15 and .25, resulting in a total of 18 study conditions.

CAT Characteristics

An item pool consisting of 480 discrete items was used to generate fixed-length 40-item CATs. The *a*-parameters in this pool range from .27 to 2.35, with a median value of 1.01 and the *b*-parameters range from -3.5 to 3.4 , with a median of .43. Provisional ability estimates were computed by Owen's Bayes mode approximation (Owen, 1969, 1975), while final estimates were obtained using maximum likelihood estimation. No content constraints were imposed on the item selection procedures. Adaptive test administrations were simulated for 50,000 examinees in each study condition.

Item Selection

Item selection was managed differently depending upon the study condition. The no control method used maximum information (MI) item selection, with no exposure control. The Sympson-Hetter method also used MI, incorporating its own exposure control parameter as a limiting factor. Both of these methods began each test targeting an examinee ability of 0. The random method had no limitations on item selection; items were drawn randomly from the pool.

For the Stratified-*a* method, throughout most of the test, an item was selected based on how close its *b*-value was to the examinee's estimated theta, within the specified stratum. For the first five items, however, items were selected randomly from within the initial stratum. Since the simulated CAT began each test assuming an examinee's ability was 0, this modification was incorporated into the Stratified-*a* method to avoid all

examinees being presented with nearly identical items early in the test.

Source of the Information/Data Generation

The exposure control procedures detailed above were investigated in this study through simulated CATs. Simulated item responses were generated based on real data and a multidimensional item response theory (MIRT) model. This model included not only the major dimensions that provide basic structure, but also numerous minor dimensions that are characteristic of actual data. MIRT data generation provides simulated data that are more similar to real data than those produced by more typical unidimensional IRT models (Davey, Nering, & Thompson, 1997; Parshall, Kromrey, Chason, & Yi, 1997).

The scored responses of approximately 3500 actual examinees to each of eight separate ACT Mathematics tests were used to obtain the study's MIRT item parameters. These multidimensional item parameters were obtained for each test form using a modified version of the program Noharm (Fraser & McDonald, 1986) which calibrated item parameters in a 50-dimensional space (Reckase, Thompson, & Nering, 1997). A rotation procedure was then used to put the separate test forms on the same scale (Thompson, Nering, & Davey, 1997), resulting in a 480-item pool.

The set of MIRT item parameters were used along with simulated examinee abilities to generate data. Item responses were generated by determining the probability of a correct response on a given item, for a given examinee, and then comparing that probability to a random number sampled from a uniform (0,1) distribution. If the probability of a correct

response was greater than the random number then the response was scored correct; otherwise, the response was scored incorrect.

Simulations

The set of MIRT item parameters and simulated examinee abilities were used to generate data, both for determining exposure control parameters (in Phase 1, a preliminary simulation phase needed for the Sympson-Hetter method) and for administering the simulated "operational" tests (in Phase 2, for all methods). Item responses were generated by determining the probability of a correct response on a given item, for a given examinee, and then comparing that probability to a random number sampled from a uniform (0,1) distribution. If the probability of a correct response was greater than or equal to the random number, then the response was scored correct; otherwise, the response was scored incorrect.

Phase 1: Simulations to Obtain SH Exposure Parameters. For the Sympson-Hetter methods it was necessary to conduct a preliminary phase of simulations in order to obtain the exposure parameters. The exposure control parameters were initialized to values close to the target maximum exposure rates, and were allowed to either increment or decrement, depending upon the observed item administration rates. The final set of exposure parameters, to be used during Phase 2 for the Sympson-Hetter method, were based on several thousand adaptive test administrations. Phase 1 consisted of 600 simulation cycles, of 5000 examinees per cycle, to obtain operational exposure parameters.

Phase 2: Simulations of Operational Tests. Operational CATs were

simulated for 50,000 examinees in each of the study conditions detailed above.

Results

The results are reported in terms of pool usage and test precision. In addition, aspects of the freeze variation of the Stratified-a method are examined further. A variety of figures are used, in an effort to fully examine item exposure performance.

Pool Usage

Pool usage information is displayed in several figures. First, the entire distribution of marginal item administration rates is shown in Figures 1a and b for the target maximum exposure rates of .15 and .25, respectively. If an exposure control method allows an item to be administered more frequently than this target, the item may be considered to have been overexposed. A complementary goal in the use of the exposure control is to improve pool usage; thus, items may also potentially be underexposed. For this study, an item is classified as underexposed if it is administered less than half the times it would be given under completely random item administration. For a test length of 40 and a pool size of 480, an item with no restrictions might be administered roughly 8% of the time; half of that completely random administration would be approximately 4%. Thus, any item used on 4% of the exams or fewer is counted as underexposed. While criteria for underexposure are consistent for a given test length and pool size, the criteria for overexposure is dependent upon the target maximum exposure rate (e.g., .15, .25).

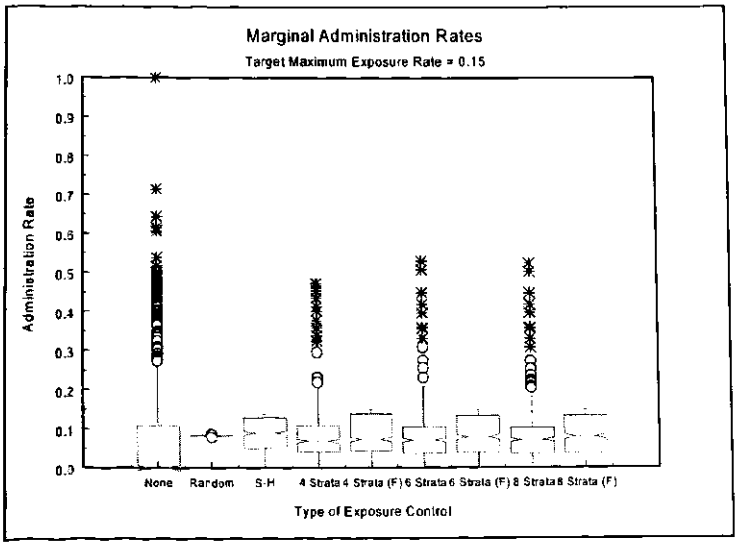


Figure 1a. *Marginal Administration Rates by Type of Exposure Control*

The pattern of results for the nine exposure control conditions is similar across the two target maximum exposure rates. Note that the random method shows ideal pool usage, without problems of either overexposure or underexposure, while the no control condition shows severe problems with both. The results also clearly show that the inclusion of freezing in the Stratified-*a* method is both necessary and effective in dealing with overexposure, regardless of the number of strata levels used. Freezing also appears to help address underexposure. Finally, the data suggest that finer distinctions in number of strata may improve the overall distribution of pool usage within the Stratified-a method (note the more

central location of the notched lines for the condition of eight strata as opposed to four strata).

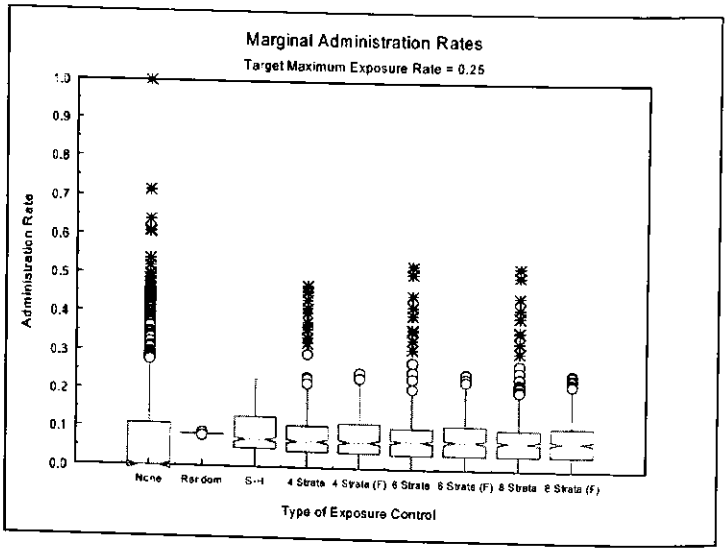


Figure 1b. *Marginal Administration Rates by Type of Exposure Control*

Figures 2a and b also displays results for pool usage. In these figures, the proportion of items over- and underexposed is displayed, for each exposure control method. Note that no control shows the worst performance, and random the best. For both target maximum exposure rates, the Sympson-Hetter method displays no problem with overexposure, and only a relatively modest problem with underexposure. For the remaining exposure control methods, overexposure is less of a problem for the less stringent target maximum of .25 than the target of .15. Under both

targets, the Stratified-*a* variations with freezing show better performance than those without. The inclusion of freezing removes any overexposure problem, and reduces the underexposure problem for this method, particularly under .15.

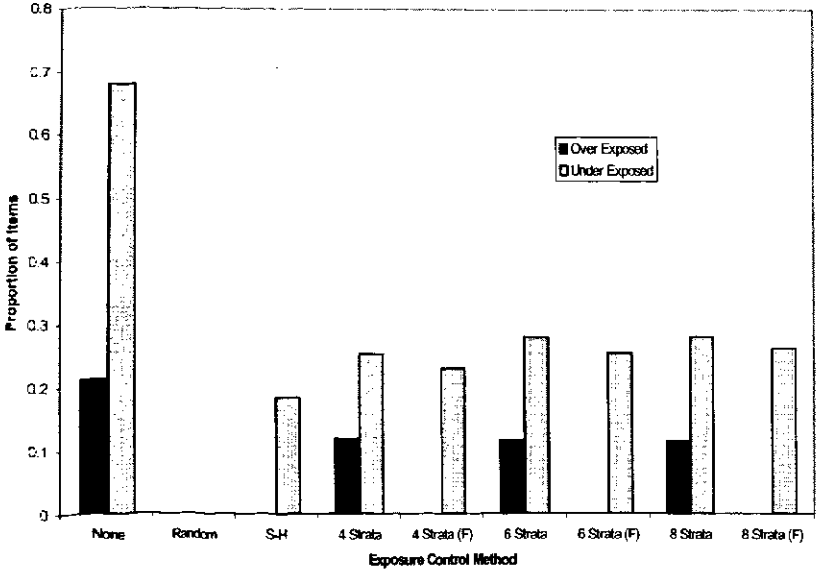


Figure 2a. *Proportion of Items Over and Under Exposed (Target Rate = 0.15) by Exposure Control Method*

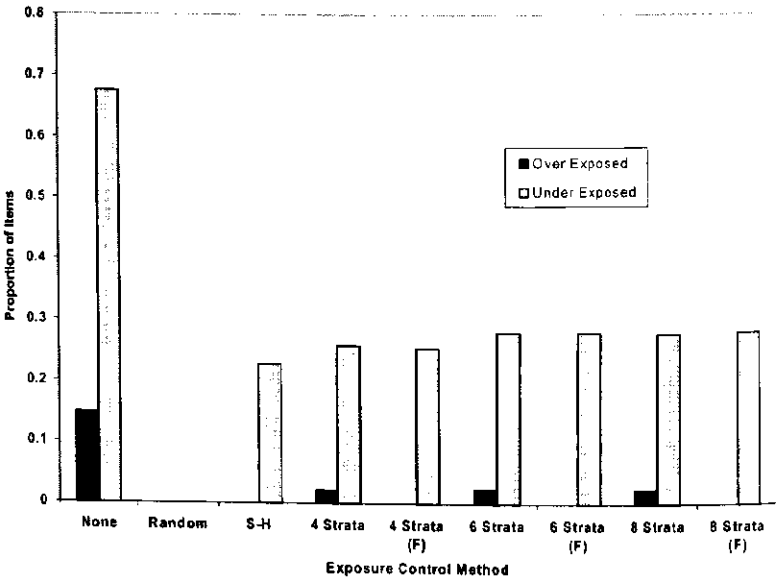


Figure 2b. *Proportion of Items Over and Under Exposed (Target Rate = 0.15) by Exposure Control Method*

Yet another view of pool usage in displayed in Figures 3a and b. In these figures conditional pool usage is examined. Conditional usage was obtained in a three-step process. First, for each of the 51 levels of true ability, the distribution of item administration rates for each of the 480 items was examined (that is, the item administration rates for examinees with the same true ability). Second, the 95th percentile of each of these conditional administration rate distributions was calculated (i.e., the rate below which 95% of the items were administered). Finally, the 95th percentile values were plotted in the figures as a function of the true

ability. In conditional usage, the random method has the lowest item administration rates across ability, as would be expected. On the other extreme, the no control method shows the highest item administration rates; however, at the high end of the ability scale, it under performs the set of Stratified-*a* methods. The variations of the Stratified-*a* method perform similarly throughout the upper portion of the scale. At the lower end of ability, however, those methods that incorporate freezing show better performance than those that do not. This difference is particularly evident in the demanding .15 condition. The number of strata levels does not appear to have an effect on conditional maximum pool usage.

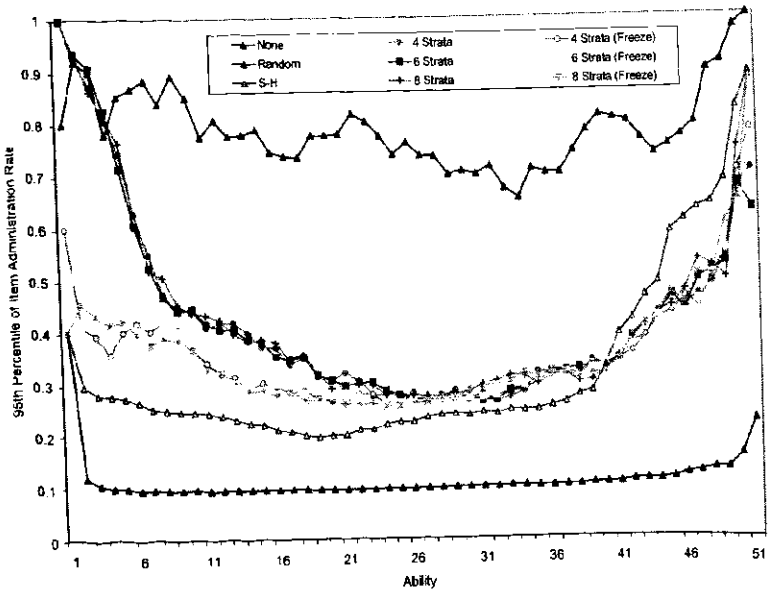


Figure 3a. *95th Percentiles of Conditional Exposure Distributions (Target Maximum Exposure Rate = 0.15)*

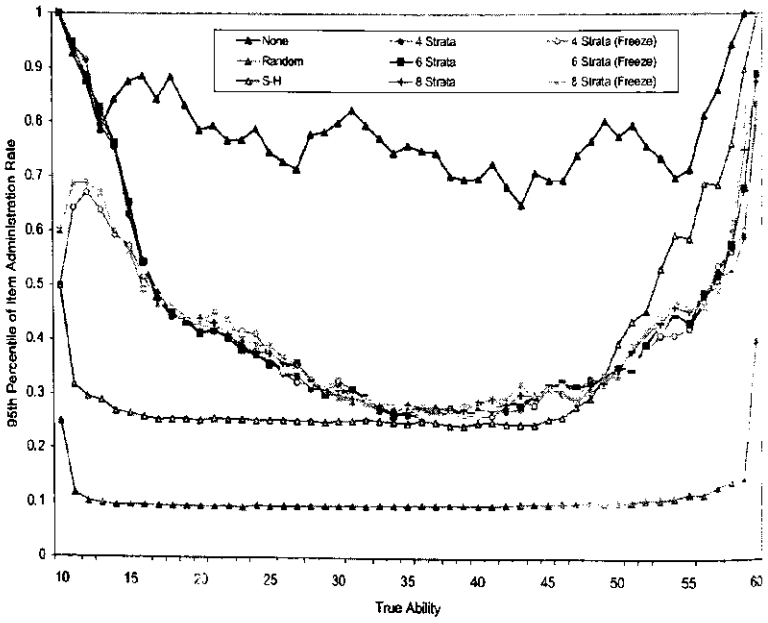


Figure 3b. *95th Percentiles of Conditional Exposure Distributions (Target Maximum Exposure Rate = 0.25)*

Test Precision

Test precision was investigated in this study by an examination of the asymptotic standard errors of the ability estimates (Hambleton & Swaminathan, 1985). These standard errors, conditional on true ability, are provided for all study conditions in Figures 4a and b. All of the methods display greater error in the tails of the ability distribution, where less information is available in the item pool. The smallest marginal error is found, as expected, for the no control condition (where no limitations are

placed on item selection), and the largest marginal error is found for the random method (where no targeting of the test to the examinee occurs). For both the .15 and .25 target maximum conditions, the six variations of the Stratified-a method and the Sympson-Hetter method all fall between these two extremes. At the high end of the ability scale, the Sympson-Hetter method performs slightly better than the set of Stratified-a methods, which perform very similarly to one another. At the low end of the ability scale, however, the three Stratified-*a* methods that incorporate freezing display slightly greater standard errors than those that do not. While the Stratified-*a* methods with freezing show relatively poorer performance than the standard Stratified-*a* variations, they perform at least as well as the Sympson-Hetter.

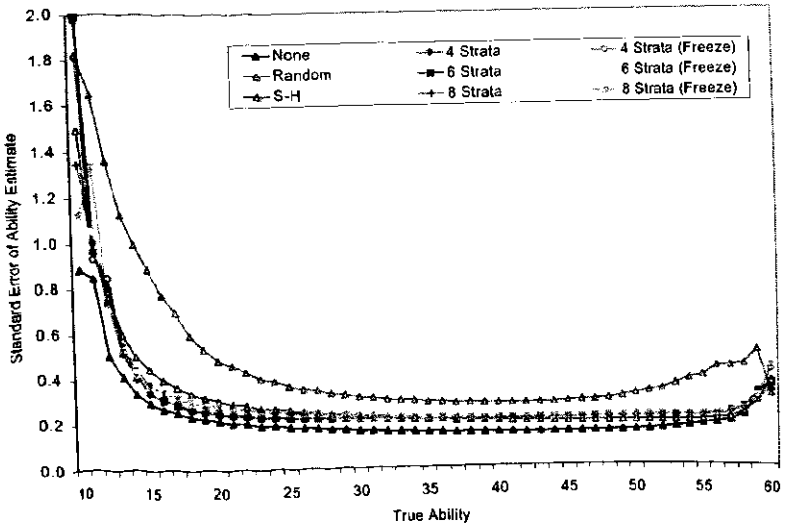


Figure 4a. *Standard Error of Ability Estimates (Target Maximum Exposure Rate = 0.15).*

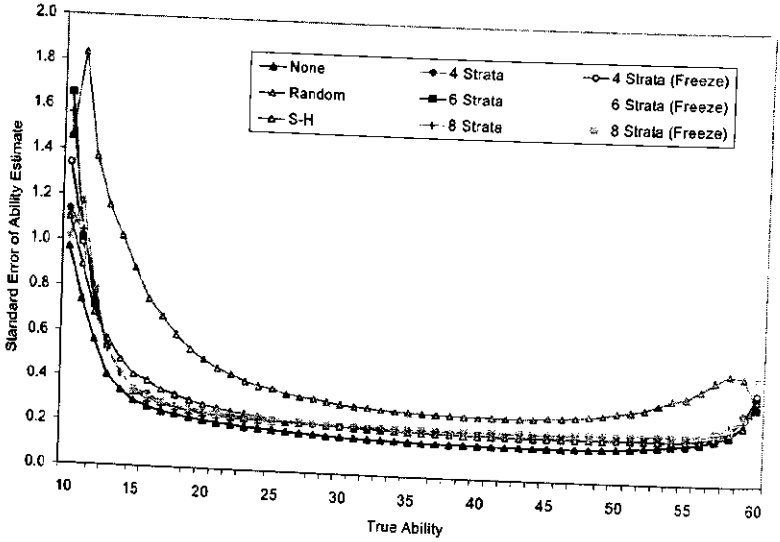


Figure 4b. *Standard Error of Ability Estimates (Target Maximum Exposure Rate = 0.25).*

Effects of Freezing

The three Stratified-*a* variations that increase freezing were investigated further to provide a more detailed examination of the effect of freezing. As can be seen, Figures 5a and b displays the proportion of times items were frozen, by number of strata levels. The majority of the items are never frozen; for the .15 condition almost 80% of the pool remains unfrozen, while for the more relaxed .25 target, over 95% of the items are never frozen. Under both targets, a small number of items are frozen fairly frequently. For the .15 condition, a few items are frozen approximately 70% of the time and in the .25 target condition a few items are frozen approximately half the time. Under both targets, a relatively small

proportion of items in the pool tend to be selected over-frequently, and thus need to be frozen regularly. For the more restrictive target maximum of .15 this effect is increased. Note that the proportion of times items are frozen is very similar across the three Stratified-*a* methods that use freezing.

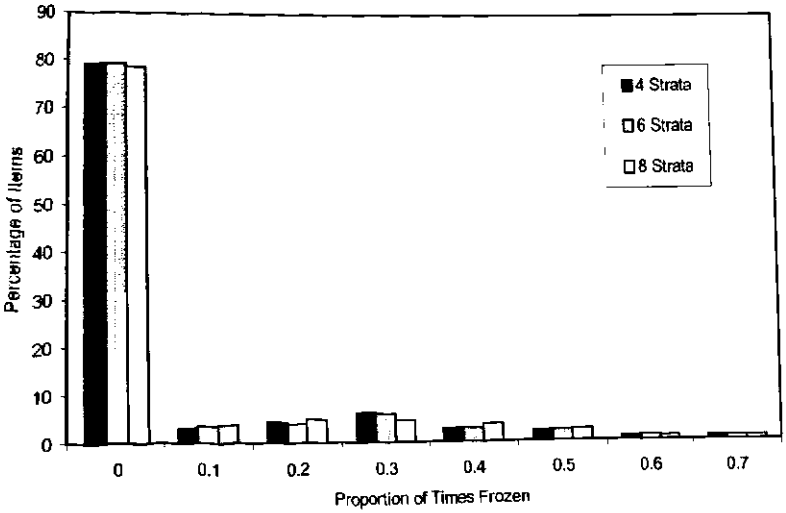


Figure 5a. *Proportion of Times Items are Frozen By Number of Strata (Target Maximum Exposure Rate = 0.15)*

In order to determine characteristics of the items that are selected too often, further plots were produced. Figures 6, 7, and 8 a and b provide plots of item freeze rate, by *a*-parameter and *b*-parameter, for the three strata level conditions and two target maximum conditions. Every item in the pool is plotted as a circle in these figures; the more frequently an item

was frozen, the larger the size of that item's circle. It is evident that items with b-values in the range of -1 to 0, and with a-values over 1.0, tended to be frozen more frequently. These middle-difficulty, high-discrimination items were apparently in great demand, resulting in their tendency to be frozen at higher rates. Items were frozen more frequently under the target maximum exposure rate of .15 than .25; however, the number of strata levels used does not appear to have had a great effect.

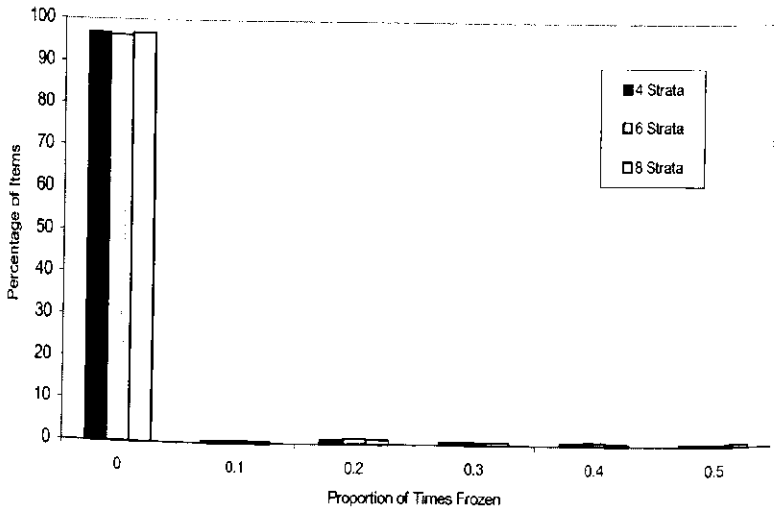


Figure 5b. *Proportion of Times Items are Frozen By Number of Strata (Target Maximum Exposure Rate = 0.25).*

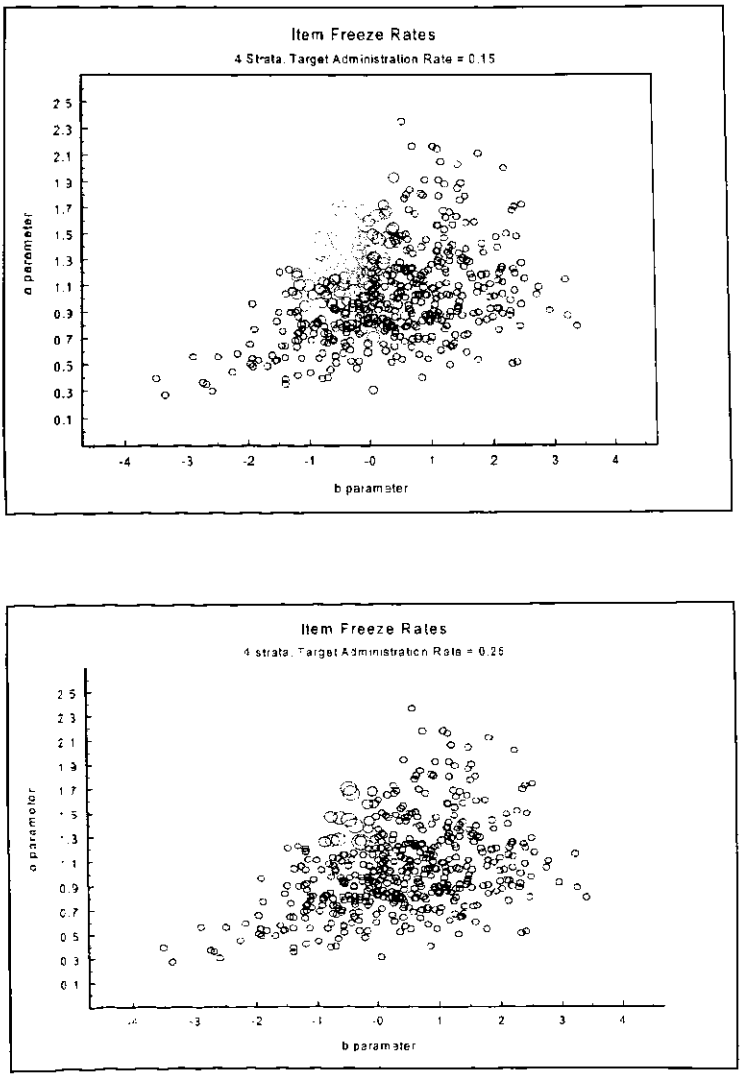


Figure 6 a and b. *Item Freeze Rates with Four Strata.*

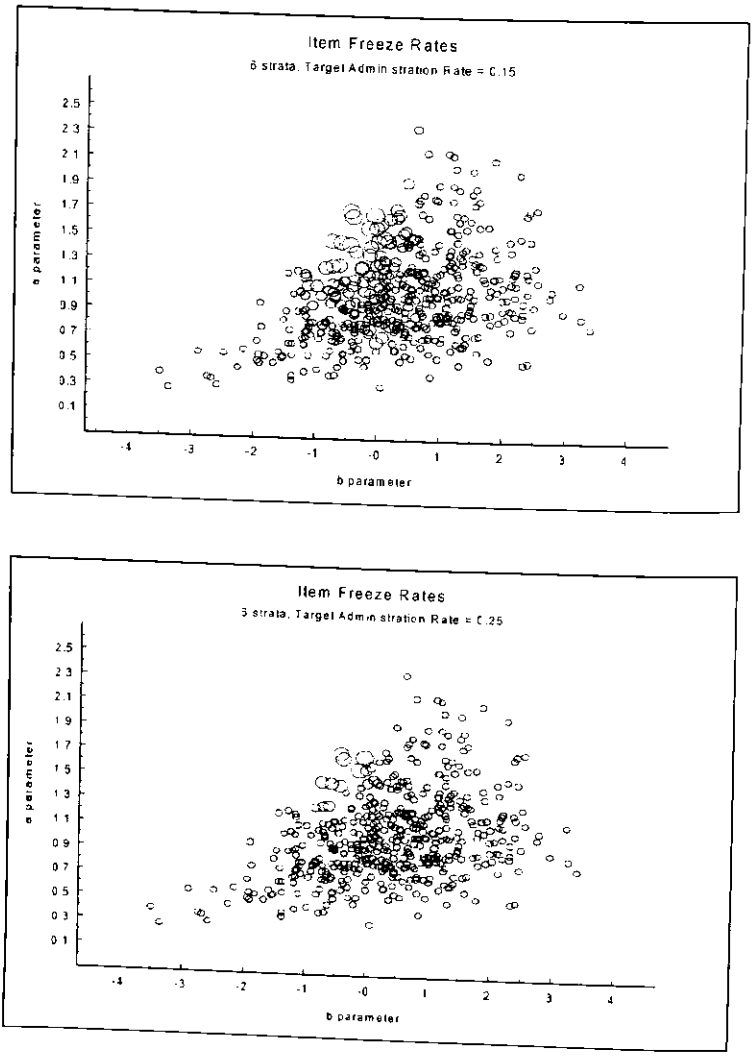
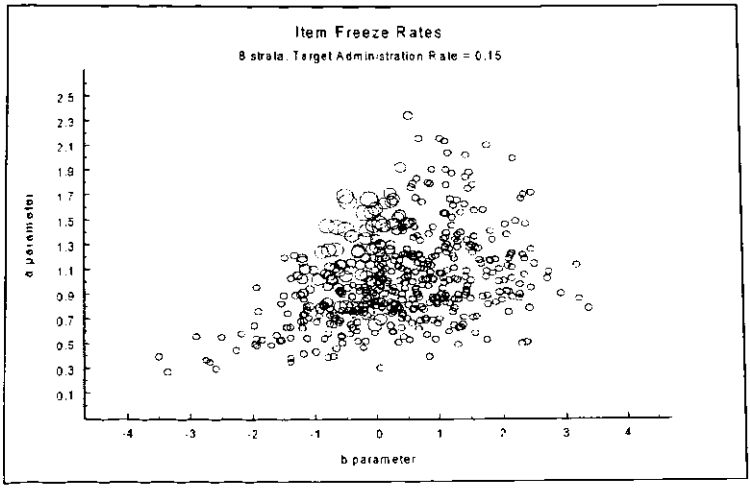


Figure 7 a and b. *Item Freeze Rates with Six Strata.*



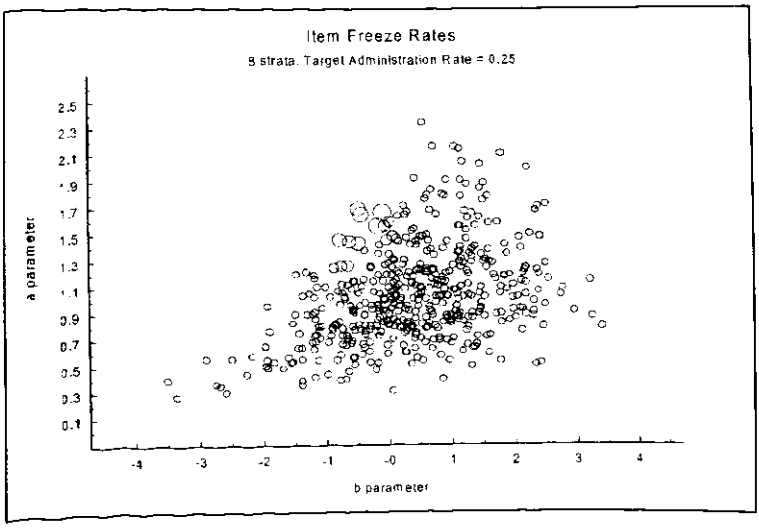


Figure 8 a and b. *Item Freeze Rates with Eight Strata.*

Summary

Computerized adaptive tests are efficient. They allow for the selection of items that provides optimal measurement at each examinee's estimated level of ability, thereby maximizing efficiency and accuracy. However, this efficiency results in very uneven item pool usage. In addition to the economic concern of items that are used too rarely, frequently administered items can become compromised, at which point they no longer provide valid measurement. The need for exposure control is clear. Whereas many exposure control procedures have been developed, none has been demonstrated to have generally superior performance, and additional work is needed.

The results of this exposure control study are very promising. While any CAT program must be a compromise between competing goals, the Stratified-*a* method with freezing appears to do remarkably well at constraining item administration rates to their target maximum goals, without degrading test precision unacceptably. Further research is needed, to confirm these findings under a greater variety of conditions. For example, the relative effectiveness of these variations under content specifications must be investigated. In this study, the number of strata levels had very little impact. However, as the number of strata levels was increased, similar *a*-value cut-points were used. Different divisions of the pool may have resulted in more of an effect for number of levels. Finally, only a single, specific item pool was used; research using other pools, with differing item characteristics, might find different results.

References

Chang, H. & Ying, Z. (1997). *A-stratified multistage computerized adaptive testing*. Applied Psychological Measurement.

Davey, T., Nering, M. L., & Thompson, T. (1997, June). *Realistic simulation of item response data*. Paper presented at the annual meeting of the Psychometric Society, Gatlinburg, TN.

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

Fraser, C. (1986). *NOHARM: An IBM PC computer program for fitting both unidimensional and multidimensional normal ogive models of latent trait theory* [computer program]. Center for Behavioral Studies, The University of New England, Armidale, New South Wales, Australia.

Fraser, C., & McDonald, R. P. (1988). NOHARM: Least squares item factor analysis. *Multivariate Behavioral Research, 23*, 267-269.

Nering, M., Davey, T., & Thompson, T. (1997, June). *Simulation of realistic ability vectors*. Paper presented at the annual meeting of the Psychometric Society, Gatlinburg, TN.

Owen, R. J. (1969). *A Bayesian approach to tailored testing* (Research Report 69-92). Princeton, NJ: Educational Testing Service.

Owen, R.J. (1975). A Bayesian sequential procedure for quantal response in the context of adaptive mental testing. *Journal of the American Statistical Association, 70*, 351-356.

Parshall, C. G., Davey, T., & Nering, M. L. (1998). *Test development exposure control for adaptive testing*. Paper presented at the symposium Adaptive Testing Research at ACT at the annual meeting of the National Council on Measurement in Education, San Diego.

Parshall, C. G., Kromrey, J. D., Chason, W. & Yi, Q. (1997) *Evaluation of parameter estimation under modified IRT models and small samples*. Paper presented at the Annual meeting of the Psychometric Society, Gattlinburg, TN.

Reckase, M. D., Thompson, T., Nering, M. L. (1997, June). *Identifying similar item content clusters on multiple test forms.* Paper presented at the annual meeting of the Psychometric Society, Gatlinburg, TN.

Stocking, M. L., & Lewis, C. (1995). *Controlling item exposure conditional on ability in computerized adaptive testing.* (Research Report 95-24). Princeton, NJ: Educational Testing Service.

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.

Thomasson, G. L. (1995). *New item exposure control algorithms for computerized adaptive testing.* Paper presented at the annual meeting of the Psychometric Society, Minneapolis.

Thompson, T., Nering, M., & Davey, T. (1997, June). *Multidimensional IRT scale linking without common items or common examinees.* Paper presented at the annual meeting of the Psychometric Society, Gatlinburg, TN.