Running head: EXPOSURE CONTROL PROCEDURES

The Evaluation of Exposure Control Procedures for an Operational CAT

Brian F. French

Purdue University

&

Tony D. Thompson

ACT, Inc.

Poster presented at the annual meeting of the American Educational Research Association. April, 2003, Chicago, IL.

Address correspondence to either:

Brian F. French
Purdue University
1446 Liberal Arts and Education Building
West Lafayette, Indiana, 47907-1054
Email: frenchb@purdue.edu

Tony D. Thompson
ACT, Inc
2201 North Dodge Street
P.O. Box 168
Iowa City, IA 52243-0168
Email: thompsot@act.org

Abstract

Exposure control procedures were evaluated for a moderate stakes operational computerized adaptive test (CAT), and included (a) a new procedure, targeted exposure control, (b) the Sympson & Hetter, and (c) *a*-stratified with *b*-blocking. Procedures were applied to a variable length CAT to evaluate their effect on item pool use, test length, and test reliability. Performance was examined, conditionally and unconditionally, on several criteria, such as pool utilization, measurement precision, test overlap, and ease of implementation. All procedures preformed similarly in terms of reliability, bias, and root mean square error. The targeted exposure control procedure (TEC) did make better use of the item pool as judged by the percent of zeros, test overlap, the chi-square statistic, and maximum exposure rate. TEC also was able to use every item, unlike the other procedures. However, conditional results suggested that none of the procedures preformed adequately in the tails of the ability distribution.

The Evaluation of Exposure Control Procedures for an Operational CAT

CAT has received increased attention due to advantages compared to paper and pencil tests. For instance, CAT offers increased test efficiency and measurement precision, as well as instant score reports (Meijer & Nering, 1999). Through item selection algorithms, CAT administers items that provide the most information at each level of estimated ability. Test efficiency is increased due to the use of fewer and more informative items. However, the use of the best items can result in unbalanced utilization of the item pool. Thus, a substantial proportion of items are underexposed, while a small proportion of items are overexposed. This outcome compromises test security, test validity, and the cost efficiency of item development.

Test security and test validity are emphasized in a CAT program in which results are used for high stakes decisions (e.g., medical board certification). These issues require strict control over the frequency of item administration. Development of exposure control procedures has been driven by the need to provide a high level of control. However, less attention has been given to the requirements of procedures for a low or moderate stakes CAT program. The decisions from the results of a low to moderate stakes CAT (e.g., self-assessment, course placement) generally do not have serious ramifications compared to a high stakes CAT (e.g., fail a course vs. medical malpractice). While security may not be of the utmost concern for moderate stakes CATs, some users may occasionally use such tests to make high stakes decisions. Thus, overexposure of items and unbalanced item use remain problematic.

To achieve balanced item use, curtail overexposure, and address test security and validity issues, exposure control procedures are included in the item selection algorithm. Many procedures have been developed and modified (Chang & Ying, 1999; Davey & Parshall; 1995; Stocking & Lewis, 1995, 2000; Sympson & Hetter; 1985) in an attempt to achieve a desired level of control. However, comparisons have not revealed an optimal procedure for all situations (Chang & Twu, 1998; Parshall, Kromrey, Harmes, & Sentovich, 2001; Pastor, Dodd, & Chang, 2002; Revuelta & Ponsoda, 1998). The choice of procedure is complicated, as the decision depends of such factors as (a) purpose of the test, (b) security requirements, (c) item pool characteristics, and (d) test stakes (Deng & Chang, 2001). Regardless of the procedure selected, there will be a compromise among the levels of test efficiency, measurement precision, and test security. For instance, a more stringent exposure procedure may be selected for security concerns. However, tighter constraints result in the use of less informative items, which can lead to a decrease in test efficiency and measurement precision (Pastor et al., 2002; Stocking & Lewis, 2000). Based on the level of importance of these issues, a balance must be achieved that is specific to the CAT program.

Three exposure control procedures were examined in this study. The Sympson-Hetter procedure (SH, Sympson & Hetter; 1985) was included because it currently is used in the operational CAT examined in this study. The procedure requires an exposure control parameter (i.e., a value between 0 and 1 obtained through simulation) for each item. To select an item with SH, first an item is provisionally selected based on information and content constraints. Then, the exposure parameter for the selected item is compared to a random number from a uniform distribution. An item is administered if the exposure parameter is greater than the random number. Otherwise, a new item is provisionally selected and a new random number is generated and compared to the item’s exposure parameter. This process is repeated until an item passes exposure control and is administered. The procedure functions well for controlling the overexposure of items. However, a large portion of the item pool can go unused. For instance, at least 60% of the item pool examined in this study is not used.

The *a*-stratified with *b*-blocking procedure (BASTR, Chang, Qian, & Ying, 2001) has been suggested for use with moderate stakes CAT (Chang et al., 2001; Pastor et al., 2002) because it is easy to implement and exposure parameters are not required. BASTR also was included because of a lack of research on the procedure's performance with a variable length CAT. The procedure was implemented as described by Chang et al. Items are split into blocks in ascending order of *b*-parameter values. Then, each block is sorted by *a*-parameter values. The lowest *a*-item from each block is placed in a stratum, the next lowest *a*- item is placed in a second stratum. This continues until the desired number of strata are achieved, resulting in mini-tests from which a few items are administered from each stratum. Item administration occurs when an item's *b*-value is closest to the current ability estimate. This procedure forces the use of low *a*-items early in the test and the use of high *a*-items later in the test. Thus, the high *a*-items are reserved for use when ability estimates are more accurate.

The targeted exposure control procedure (TEC; Thompson, 2002) attempts to increase the administration probability of unused items. The focus of this procedure, in contrast to other methods that mainly focus on controlling overexposure, is to ensure examinees are administered informative items while making good use of the item pool. To select an item with TEC, items must first meet measurement and content constraints, (i.e., upper and lower bound of intermediate information target, Davey & Fan, 2000). These items form an acceptable set of items, any one of which would be considered appropriate for administration. Once an acceptable set of items is formed, an item's probability of administration is inversely related to its administration rate. Therefore, items that were used less frequently have a higher probability of being administered. Exposure parameters are obtained through simulation as in the SH procedure. This procedure has not been compared to other procedures and was included for empirical evaluation.

Method

Conditions

A variable test length with a maximum of twelve and a minimum of seven items was used. Exposure parameters were generated separately for the SH and TEC procedures through simulations with 4000 simulees from a N (0, 1) distribution. Each simulation completed 125 iterations to ensure stability of results.

Data

Item parameters for 269 items from an operational mathematics course placement CAT were used. Item parameter estimates were obtained via marginal maximum likelihood estimation in BILOG (Mislevy & Bock, 1990) based on the three-parameter logistic item response theory model. The mean and standard deviation of the discrimination (*a),* difficulty (*b*), and pseudo-guessing (*c*) parameters were 1.13 (.25), .25 (.85), and .15 (.06), respectively. Simulated tests were administered to a sample of 10,000 simulees with true abilities randomly generated from a N (0, 1) distribution. Simulations were completed for 20,000, 50,000, 60,000 and 100,000 simulees with minor differences in results. Thus, a sample size of 10,000 was considered sufficient. Maximum information selection (MI) was used with the SH. TEC, however, required the use of targeted information selection (Davey & Fan, 2000; Thompson, 2002). Simulees' provisional ability estimates and final ability estimates were obtained by Owen's Bayes estimation (Owen, 1975) and maximum likelihood estimation, respectively.

Analysis and Results

Several criteria were used to evaluate performance of the procedures. Bias and the root mean square error (RMSE) of ability estimates on the scale scores were computed to examine accuracy. Reliability was computed as the squared correlation of the estimated scores with the true scores (Lord, 1980). The percent of zeros, or items never exposed, test overlap, administration rates, maximum exposure of a single item, a chi-square statistic and an *F*-ratio based on the chi-square were calculated to examine pool usage. The chi-square statistic is a measure of skewness in the administration rate distribution, and the *F*-ratio assesses the reduction in skewness of one procedure relative to another procedure (Chang et al., 2001; Chang & Ying, 1999).

As seen in Table 1, reliability, bias, and RMSE were consistent across methods. The average test length differed with SH having lower lengths compared to the other procedures. The TEC procedure was closest to the target exposure rate followed by BASTR, as seen in Figure 1. The SH procedure used the least number of items and had the highest exposure rates. Comparison of the *F*-ratio revealed that the TEC procedure had the greatest improvement relative to the other procedures. For instance, 64% and 30% of administration distribution skewness in the TEC procedure was reduced in comparison to the SH and BASTR procedures, respectively. The BASTR had a 49% reduction in skewness relative to the SH. The TEC procedure had the lowest (a) maximum exposure rate of a single item, (b) test overlap rate, and (c) percent of zeros, as seen in Figure 2. In addition, the TEC procedure was able to use every item.

Conditional results from a range of –5 to 5 on the ability scale followed the trends previously mentioned. However, all procedures were problematic with extreme low and high ability levels. For instance, test overlap across procedures had a range of 8% to 24% in the middle of the ability distribution and a range of 29% to 66% in the tails of the ability distribution, as seen in Figure 3. The TEC procedure used all items regardless of ability, whereas the SH procedure left a consistently high percentage of items unexposed. (See Figure 4.) The BASTR procedure did show somewhat of an improvement over the SH procedure, but remained problematic, especially at extreme ability levels. Similar results were observed with maximum exposure rates and test length, with all methods approaching a 100% exposure rate for a single item and an increase in test length in the high ability range.

Conclusion

The primary purposes of the study were to identify an appropriate exposure control procedure for a relatively short, moderate stakes, variable length CAT and empirically evaluate the TEC procedure. The three procedures examined are quite distinct in their approach to controlling exposure, and thus displayed somewhat different strengths and weaknesses.

All procedures preformed similarly in terms of reliability, bias, and root mean square error. The SH procedure had the lowest average test length. The TEC procedure made better use of the item pool in terms of the percent of zeros, test overlap, the chi-square statistic, the maximum exposure rate, and was able to use every item, unlike the other procedures. This procedure appears most favorable based on these results. However, more time was spent making adjustments to the TEC procedure to best fit this particular CAT. Substantial improvements may be seen with the other methods with comparable modifications. For example, the performance of BASTR was adequate with a variable length CAT. However, the optimal number of strata and number of items used from each stratum remain unclear (Chang et al., 2001; Pastor et al., 2002). Also, conditional results revealed that none of the procedures were adequate in the tails of the distribution. This issue must be addressed by any high stakes CAT testing program. For a low to moderate stakes CAT, this may be less of a concern. In addition, the proportion of the pool size to test length was large and content constraints were not considered. The evaluation of the methods under more constrained conditions is a necessary next step.

References

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

Chang, S. W., & Twu, B. Y. (1998). *A comparative study of item exposure control methods in computerized adaptive testing.* (Research Report 98-3). Iowa City, IA: ACT.

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

Davey, T., & Fan, M. (2000, April). *Specific information item selection for adaptive testing.* Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

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.

Deng, H. & Chang, H. H. (2001, April). *A-stratified computerized adaptive testing with unequal item exposure across strata.* Paper presented at the Annual Meeting of the American Educational Research Association, Seattle, WA.

Fan, M., Thompson, T., & Davey, T. (1999, April). *Constructing adaptive tests to parallel conventional programs.* Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal, Canada.

Leung, C. K., Chang, H.-H., & Hau, K. T. (2001, April). *Integrating stratification and information approaches for multiple constrained CAT.* Paper presented at the annual meeting of the National Council on Measurement in Education, Seattle, WA.

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

Meijer, R. R., & Nering, M. L. (1999). Computerized adaptive testing: Overview and introduction. *Applied Psychological Measurement, 23*, 187-194.

Mislevy, R. J., & Bock, R. D. (1990). Item analysis and test scoring with binary logistic models. *Bilog 3.* Chicago: Scientific Software International, Inc.

Parshall, C. G., Kromrey, J. D., Harmes, J. C., & Sentovich, C. (2001, April). *Nearest Neighbors, Simple Strata, and Probabilistic Parameters: An Empirical Comparison of Methods for Item Exposure Control in CATs.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, Seattle, WA.

Pastor, D.A., Dodd, B. G., & Chang, H. 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.

Revuelta, J., & Ponsoda, V. (1998). A Comparison of Item Exposure Control Methods in Computerized Adaptive Testing. *Journal of Educational Measurement, 35*, 311-27.

Stocking, M. L. & Lewis, C. (2000). Methods for controlling the exposure of items in CAT. In W. J. van der Linden & C. A. W. Glass (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 163-182). Dordrecht, The Netherlands: Kluwer Academic.

Stocking, M. L. & Lewis, C. (1995). *A New Method of Controlling Item Exposure in Computerized Adaptive Testing* (Research Report 95-25). Princeton, NJ: Educational Testing Service.

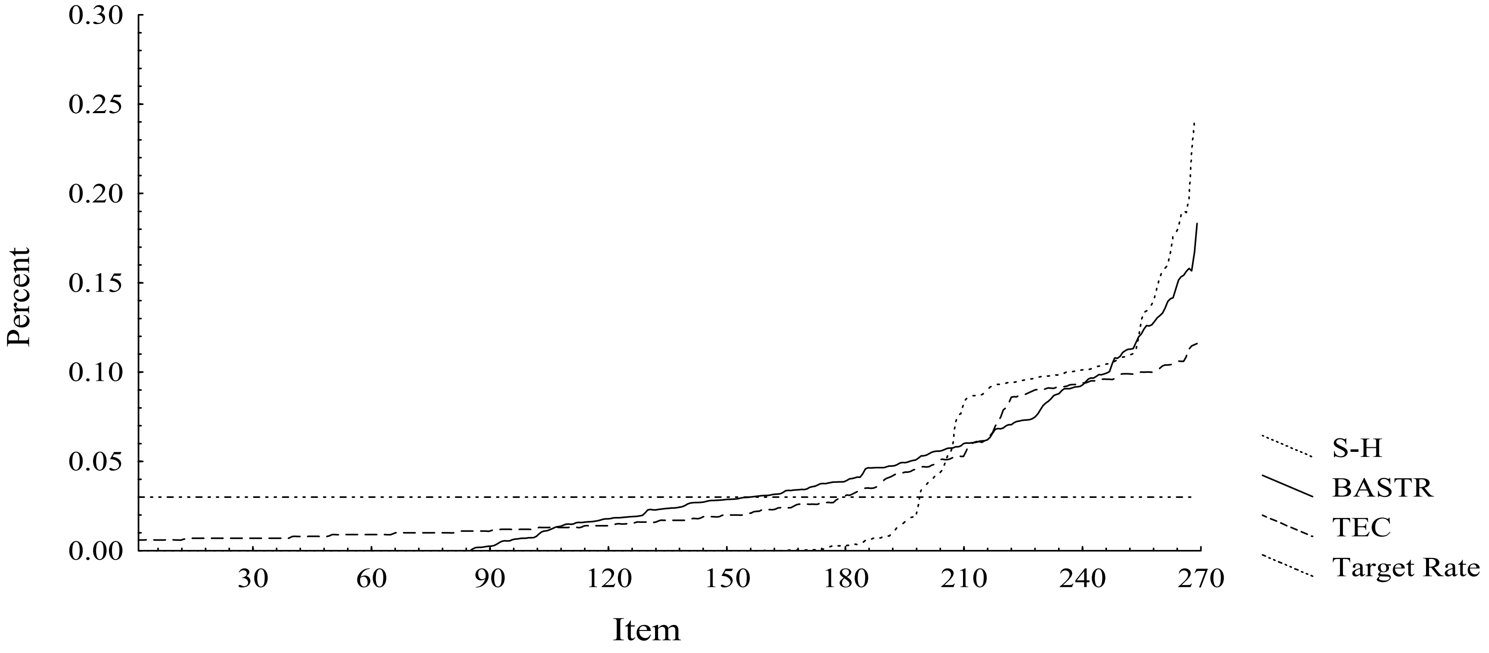
Sympson, J. B., & Hetter, R. D. (1985). Controlling item-exposure rates in computerized adaptive testing. *Proceedings of the 27th annual meeting of the Military Testing Association* (pp.973-977). San Diego, CA: Navy Personnel Research and Development Center.

Thompson, T. (2002, April). *Employing new ideas in CAT to a simulated reading test.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, New Orleans, LA.

Table 1
Evaluation criteria for exposure control procedures

Condition	**Reliability**	***M* test length**	**Bias**	**RMSE**
BASTR	.828	9.45	-.0073	.092
S-H	.828	7.72	-.006	.091
TEC	.826	9.05	-.0079	.094

Figure 1. Exposure Rates for Items



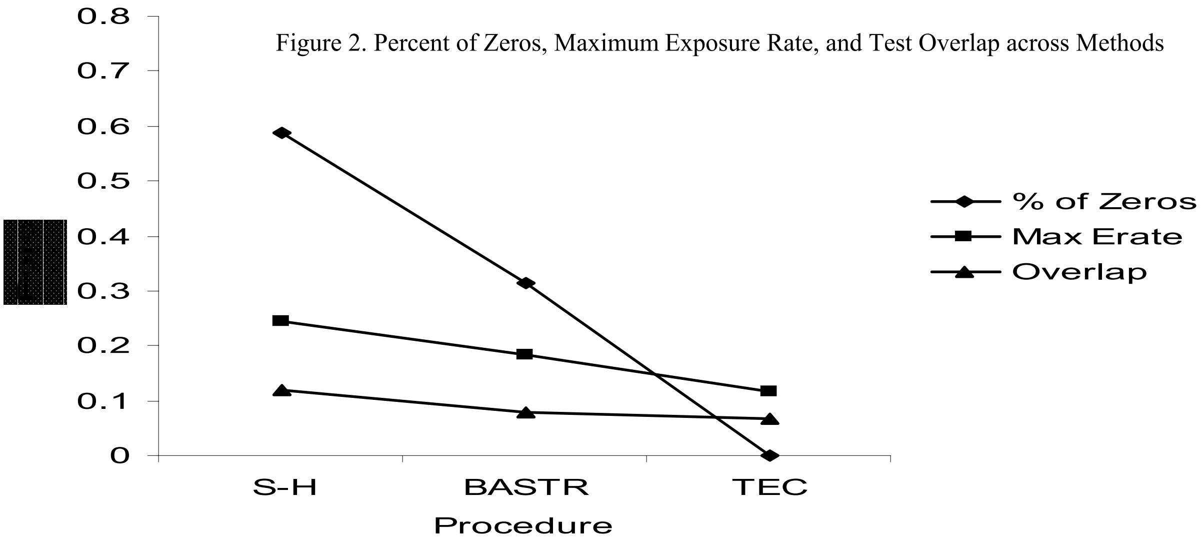


Figure 2. Percent of Zeros, Maximum Exposure Rate, and Test Overlap across Methods

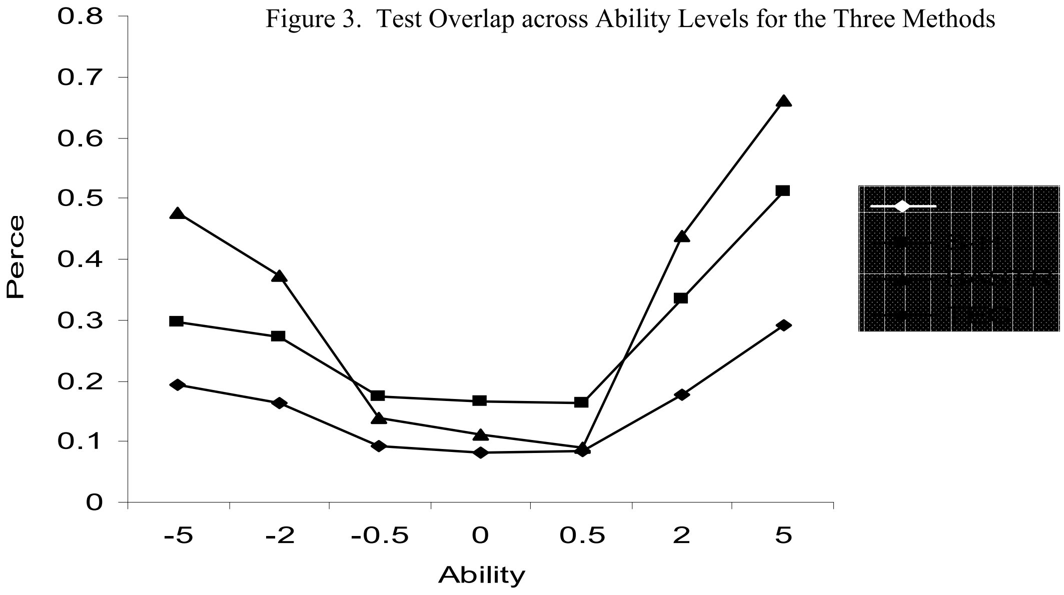
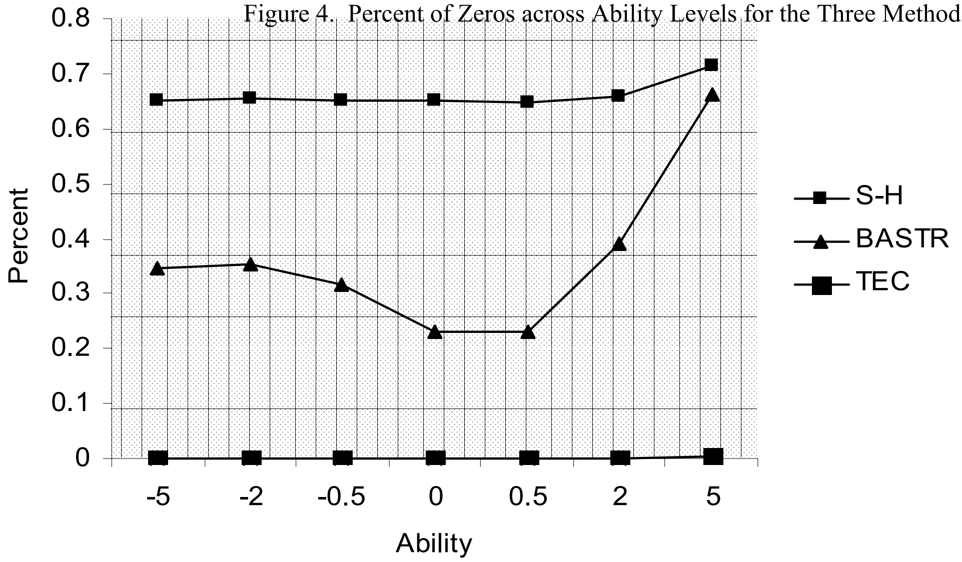


Figure 3. Test Overlap across Ability Levels for the Three Methods

Figure 4. Percent of Zeros across Ability Levels for the Three Methods


Percent
Ability
S-H
BASTR
TEC