A Comparison of Item Exposure Control Procedures Using a CAT System Based on the Generalized Partial Credit Model

Winona Burt
Soo-jin Kim
University of Texas at Austin

Laurie Laughlin Davis
Pearson Educational Measurement

Barbara G. Dodd
University of Texas at Austin

Abstract

Security of test items is paramount in computerized adaptive testing (CAT) where tests are shorter and given on demand and therefore more susceptible to over exposure than traditional paper-and-pencil tests. A number of algorithms have been developed to control exposure rates of items during CAT. The purpose of the present study was to assess the accuracy of three exposure control procedures using a CAT system based on the generalized partial credit model. The Sympson-Hetter, modified within .10 logits, and randomesque procedures were investigated. For the modified within .10 logits and randomesque procedures, either 3 or 6 items were included in the item group from which the next item to be administered was randomly selected. A no exposure control condition was also included as a baseline for comparison of the exposure control procedures. The results revealed that the Sympson-Hetter procedure performed better than the other exposure control procedures in terms of fewer non-convergent cases and good control of maximum exposure rates.

Introduction

Computerized adaptive testing (CAT) achieves maximum measurement precision when the most informative items are selected for the current ability level estimate. As a result, the more informative and more discriminating items are administered more frequently. A consequence of over exposed items is a less valid test. If examinees have knowledge of what items to expect on a test their test scores become artificially inflated and scores no longer reflect an accurate estimate of the examinee's true ability level.

In order to maintain test security, constraints must be implemented in CAT to control the exposure rates of the items. Several factors contribute to the overexposure of a subset of items—the continuous nature of CAT, the small number of items in the CAT pool, and the nature of item selection in CAT. Selecting items using maximum information item selection can lead to a discrepancy between the available item pool and those items that are actually administered (Davis, 2002). Wainer and Eignor (2000) found that in some circumstances as few as 12% of the available item pool could account for as much as 50% of the functional item pool (i.e., those items actually administered).

Exposure control techniques are used to protect the security of the item pool by limiting the degree to which items may be exposed. There are several approaches to exposure control, and each produces different testing algorithms. Way (1998) discusses two types of exposure control strategies – randomization and conditional selection. Rather than choosing the single most informative item, randomization strategies randomly choose the next item to be administered from a subgroup of the most informative items. Conditional selection strategies limit item administration directly based on a predefined criterion, for example, the expected frequency of item usage.

Two examples of the randomization approach that can be implemented in CAT to control item exposure rates are the randomesque method developed by Kingsbury and Zara (1989) and within .10 logits procedure proposed by Lunz and Stall (1998). With the randomesque technique, a subgroup of the most informative items for a given trait level is assembled and the next item to be administered is randomly chosen from this group. For dichotomously-scored items, the within .10 logits procedure finds all items with a difficulty within .10 logits of the most recent theta estimate and the next item to be administered is randomly chosen from this group. Davis and Dodd (2001) modified the within .10 logits procedure for use with polytomously-scored items. A specified number of the most informative items are selected for each of three points along the theta metric: estimated theta, estimated theta minus 0.10 and estimated theta plus 0.10. The next item to be administered is then randomly selected from this subset of items.

One of the most well known conditional procedures is the Sympson-Hetter (Sympson & Hetter, 1985) algorithm that controls the item exposure rate through the use of item parameters that have been determined probabilistically. Iterative CAT simulations are conducted with a large representative sample of the population to determine the exposure control parameter (K_i) for each item in the item pool. These exposure control parameters are then used in the actual CATs to constrain the probability of administering each item. When an item is selected for administration, the K_i is compared to a random number drawn from a uniform distribution. If K_i for the selected item exceeds the random

number, the item is administered. Otherwise, the item is blocked from administration and the next most informative item is selected to be considered for administration.

While much research has been conducted which examines the utility of exposure control with dichotomous item pools, only recently have researchers begun to investigate these issues in relation to item pools in which polytomous (partial credit) scoring is used. Polytomously-scored items yield a higher modal level of information across a wider range of the theta scale than do dichotomously-scored items (Koch & Dodd, 1989). Since polytomously-scored items have properties unique and separate from dichotomously-scored items they must be separately studied under conditions of constrained item selection.

Pastor, Chiang, Dodd, and Yockey (1999), examined the performance of the Sympson-Hetter exposure control algorithm for polytomous items with two item pool sizes (60 and 120). Pastor et al. (1999) concluded that the Sympson-Hetter provided some protection against item exposure with a minimum loss of measurement precision when used with the partial credit model. Davis, Pastor, Dodd, Chiang, and Fitzpatrick (in press) also examined the performance of the Sympson-Hetter with the partial credit model with the inclusion of content balancing, an additional item pool size (240 items), and two levels of test dimensionality. While the Davis et al. (in press) study replicated the results of Pastor et al. (1999) with regard to measurement precision, they found that the Sympson-Hetter was ineffective in constraining item exposure rates to the desired target value and that the added benefits in exposure rate, item overlap, and pool utilization of the Sympson-Hetter were modest. Davis (2002) later concluded that when compared head to head with other options for exposure control, randomization procedures demonstrated comparable or better performance on almost all outcome measures for CATs based on polytomous item response theory (IRT) models.

The current study investigated two variations of randomization procedures for controlling item exposure—the randomesque (Kingsbury & Zara, 1989) and the modified within .10 logits (Lunz & Stahl, 1998; Davis & Dodd, 2001). Either 3 or 6 items defined the item group size from which the item to be administered was selected. The Sympson-Hetter procedure and a maximum information item selection procedure were also studied. The maximum information procedure served as a no exposure control baseline to which the other exposure control procedures were compared. Each exposure control procedure was evaluated in terms of its ability to successfully control item exposure and item overlap, use the available item pool, and reproduce known ability values in the context of a CAT system based on the generalized partial credit model (Muraki, 1992).

Method

Exposure Control Procedures:

- Randomesque with an item group size of 3
- Randomesque with an item group size of 6
- Modified Within .10 Logits with an item group size of 3
- Modified Within .10 Logits with an item group size of 6
- Sympson-Hetter (.29 exposure control target)
- No exposure control (maximum information)

Item Pool:

- 210 items from the NAEP 1996 Science Assessment
- Number of categories:
  - 83% three category items
  - 17% four category items
- Content Areas:
  - 30% physical science items
  - 35% earth science items
  - 35% life science

Data Generation:

Using the published NAEP Science item parameter estimates with 0.40 added to the discrimination parameter to mirror a high-stakes test, responses were generated using conventional IRT data generation procedures for two samples. The first data set consisted of 8,000 simulees and was used to establish the exposure control parameters for the Sympson-Hetter method. For the second sample, 1,000 simulees were generated for use in the CAT simulations.

CAT Simulations:

- A SAS program for the CAT simulations based on the generalized partial credit model was used (Davis, 2002).
- The initial theta was set equal to 0.0, the mean of the known theta values.
- A variable step size was used prior to MLE of theta.
- Kingsbury and Zara's (1989) content balancing procedure was used to ensure each CAT mirrored the percentages of each content area and number of categories per item that exist in the item pool.
- A fixed-length stopping rule of 20 items was used.

Data Analyses:

- Listwise deletion of non-convergent cases
- Descriptive statistics (mean and SD of theta and standard error; number of non-convergent cases)
- Correlations (known and estimated thetas)
- Accuracy indices (RMSE and bias)
- Pool utilization
- Exposure rates (distribution and descriptive statistics)
- Item overlap rates

Results

Table 1
Means (and Standard Deviations) for Estimated Theta, Standard Error, and Number of Non-convergent Cases for the Exposure Control Conditions (N = 848)

Exposure Control Condition	Theta Estimate*	Standard Error	Non-convergent Cases
No Exposure Control	0.019 (1.124)	0.256 (0.044)	17
Randomesque-3	0.022 (1.124)	0.264 (0.048)	39
Randomesque-6	0.006 (1.141)	0.282 (0.057)	59
Within .10 Logits-3	0.022 (1.130)	0.265 (0.048)	43
Within .10 Logits-6	0.000 (1.136)	0.281 (0.057)	77
Sympson-Hetter	0.009 (1.132)	0.274 (0.050)	21

* Note: Mean and SD for Known Thetas were Mean = .04, SD = 1.02.

Table 2
Correlation Coefficients between Known and Estimated Theta, Bias and RMSE for the Exposure Control Conditions (N = 848)

Exposure Control Condition	Correlation	Bias	RMSE
No Exposure Control	0.95	0.02	0.34
Randomesque-3	0.95	0.02	0.36
Randomesque-6	0.94	0.03	0.38
Within .10 Logits-3	0.95	0.02	0.36
Within .10 Logits-6	0.93	0.04	0.41
Sympson-Hetter	0.94	0.03	0.39

Table 3
Pool Utilization and Exposure Rates the Exposure Control Conditions (N = 848)

Exposure Control Condition						
Exposure Rate	None	R-3	R-6	W-3	W-6	SH
.81 - 1.0	0	0	0	0	0	0
.71 - .80	2	0	0	0	0	0
.61 - .70	3	1	0	1	0	0
.51 - .60	4	7	1	5	1	0
.41 - .50	8	2	1	6	1	0
.31 - .40	13	14	15	14	12	9
.21 - .30	9	14	24	13	29	43
.11 - .20	15	24	40	26	39	24
.01 - .10	44	77	77	73	79	51
0.0	102	71	52	72	49	83
Mean	0.09	0.09	0.09	0.09	0.09	0.09
SD	0.17	0.14	0.11	0.14	0.11	0.12
Maximum	0.73	0.61	0.52	0.63	0.50	0.33
% Not Admin.	48.6%	33.8%	24.8%	34.3%	23.3%	39.5%

Table 4
Mean (and Percentage of Test Length) Item Overlap Rates for the Exposure Control Conditions (N = 848)

Exposure Control Condition	Overall Average Overlap	Different Abilities Average Overlap	Similar Abilities Average Overlap
No Exposure Control	7.64 (38.2%)	1.70 (8.5%)	9.00 (45.0%)
Randomesque-3	6.11 (30.6%)	1.59 (7.9%)	7.16 (35.8%)
Randomesque-6	4.56 (22.8%)	1.43 (7.2%)	5.28 (26.4%)
Within .10 Logits-3	6.11 (30.5%)	1.59 (8.0%)	7.15 (35.7%)
Within .10 Logits-6	4.56 (22.8%)	1.45 (7.2%)	5.27 (26.4%)
Sympson-Hetter	4.73 (23.7%)	1.03 (5.1%)	5.58 (27.9%)

Noteworthy Findings:

- The number of non-convergent cases was unacceptably high for the randomesque and within .10 logits procedures. Only the Sympson-Hetter procedure yielded comparable numbers of non-convergent cases relative to the no exposure control condition.
- The largest mean standard error associated with the ability estimates was found for the randomization procedures that used an item group size of 6.
- The maximum exposure rate for the Sympson-Hetter procedure was considerably smaller than that found for the other exposure control procedures.
- The randomization procedures with an item group size of 6 used more of the item pool than the other procedures.
- The smallest item overlap rates for all simulees and for simulees with similar abilities were obtained for the randomization procedures that used an item group size of 6 and the Sympson-Hetter procedure. For simulees with different abilities, the Sympson-Hetter procedure yielded the smallest overlap rate.

Discussion

- The item pool that was used in the current study was less than optimal for a CAT system based on a polytomous IRT model that employs content balancing procedures. Very few of the items within each content area were scored with four categories (11 – 12 items). As a consequence, the randomization procedures that were based on a six-item group set, produced the most cases of non-convergence because inappropriate items were being randomly selected for administration. The randomization procedures with a three-item group set were only a bit better in terms of non-convergence problems. Only the Sympson-Hetter procedure faired well in terms of the non-convergence problems relative to the no exposure control condition.
- Examination of the maximum exposure rate findings revealed that only the Sympson-Hetter procedure was able to adequately control the exposure rate (.33 or less). The other procedures yielded maximum exposure rates of .5 or higher. This finding is in conflict with the results of the Davis (2002) study where the randomization procedures out performed the Sympson-Hetter procedure. The discrepancy may be due to differences in the distribution of items for content balancing purposes.
- Collectively, the results reveal that the Sympson-Hetter procedure was able to control the maximum exposure rate and utilized 60 percent of the item pool. Further research is warranted to establish guidelines for the use of exposure control procedures in relationship to the characteristics of the item pool.

References

Davis, L.L. (2002). *Strategies for Controlling Item Exposure in Computerized Adaptive Testing with Polytomously Scored Items.* Unpublished doctoral dissertation, the University of Texas at Austin.

Davis, L.L., & Dodd, B.G. (2001). *An examination of testlet scoring and item exposure constraints in the verbal reasoning section of the MCAT*. MCAT Monograph Series: Association of American Medical Colleges.

Davis, L.L. Pastor, D.A., Dodd, B.G., Chiang, C., & Fitzpatrick, S. (in press). An examination of exposure control and content balancing restrictions on item selection in CATs using the partial credit model. *Journal of Applied Measurement.*

Kingsbury, G.G., & Zara, A.R. (1989). Procedures for selecting items for computerized adaptive tests. *Applied Measurement in Education, 2*, 359-375.

Koch, W.R., & Dodd, B.G. (1989). An investigation of procedures for computerized adaptive testing using partial credit scoring. *Applied Measurement in Education, 2*, 335-337,

Lunz, M.E., & Stahl, J.A. (1998). *Patterns of item exposure using a randomized CAT algorithm*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego, CA.

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

Pastor, D.A., Chiang, C., Dodd, B.G., & Yockey, R., (1999, April). *Performance of the Sympson-Hetter exposure control algorithm with a polytomous item bank.* Paper presented at the annual meeting of American Educational Research Association, Montreal, Canada.

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.

Sympson, J. B. & Hetter, R. D. (1985, October). *Controlling item exposure rates in computerized adaptive testing*. Paper presented at the annual meeting of Military Testing Association. San Diego, CA: Navy Personnel Research and Development Center.

Wainer, H., & Eignor, D. (2000). Caveats, pitfalls, and unexpected consequences of implementing large-scale computerized testing. In Wainer, Howard (Ed). *Computerized adaptive testing: A primer (2nd ed.)*. pp. 271-299. Mahwah, NJ Lawrence Erlbaum Associates.

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