Graduate Management Admission Council®

Limiting Item Exposure for Target Difficulty Ranges in a High-Stakes CAT

Xin Li, Kirk Becker, & Jerry Gorham
Pearson VUE

Ada Woo
National Council of State Boards of Nursing

Presented at the Item Exposure Paper Session, June 3, 2009

Abstract

Numerous studies have been conducted to evaluate the effectiveness of a variety of algorithms that modify the CAT selection process to control item exposure. No studies, however, have focused on exposure of items within a particular range, especially those items with difficulty level near the cut score on variable-length adaptive tests. The CAT algorithm may excessively administer these items under maximum item information selection. Overexposure of items might affect item parameter estimates and potentially the integrity of the test. This research investigated multiple methods for limiting exposure of items near the cut score and evaluated the results for measurement precision. Response data from a large-scale live CAT licensure exam were used to obtain the known item parameters for simulation. Four procedures were employed for controlling exposure of items near the cut score in a CAT, including the Kingsbury-Zara method, the "within-10-logits" method, a constrained beta method, and a constrained maximum exposure rate method. These methods were compared to a baseline condition with no exposure control. The performance of these procedures was evaluated for measurement precision by the standard error of measurement. Other variables associated with test security included exposure rates and utilization of the item bank.

Acknowledgment

Presentation of this paper at the 2009 Conference on Computerized Adaptive Testing was supported in part with funds from GMAC®.

Citation

Li, X., Becker, K., Gorham, J., & Woo, A. (2009). Limiting item exposure for target difficulty ranges in a high-stakes CAT. In D. J. Weiss (Ed.), *Proceedings of the 2009 GMAC Conference on Computerized Adaptive Testing.* Retrieved [date] from www.psych.umn.edu/psylabs/CATCentral/

Author Contact

Xin Li, 5601 Green Valley Drive, Bloomington, MN 55437, U.S.A.
Email: xin.li@pearson.com

Limiting Item Exposure for Target Difficulty Ranges in a High-Stakes CAT

Item exposure control has become a critical and practical issue since computerized adaptive testing was widely implemented in test administration. Overexposed items may be easily memorized and shared among examinees and become compromised, resulting in a decrease in item difficulty and even affecting ability estimation and test validity. However, incorporating exposure control procedures might restrict selecting the most informative items, which increases the measurement error of examinees' abilities. On the other hand, some items in the bank might be rarely selected by the CAT algorithm and become underexposed. It is economically inefficient to have unused items, given the cost of developing a large item bank and also functionally ineffective in terms of the diversity of items administered among examinees (Georgiadou, Triantafillou, & Economides, 2007).

Strategies for controlling item exposure have been developed to prevent overexposure of items while maintaining measurement precision and optimizing item bank utilization. Randomization and conditional selection are two major types of exposure control techniques (Way, 1998). Randomization procedures allow a random component for controlling item exposure. Kingsbury and Zara (1989) proposed the "randomesque" method that randomly selects one item out of a prespecified number of the most informative items. The number is arbitrary and typically remains the same throughout the testing. Another method designed by Lunz and Stahl (1998) randomly selects items within .10 logits of the target item difficulty. In this method, the number of items available in the range varies across selections.

Chang and Ying (1999) developed an *a*-stratified CAT to limit the exposure of items with high discrimination, by restricting their selection, and to leave them for potential administration later in the test when θ estimates become more stabilized. While CATs using the Rasch model do not have exposure issues due to the item discrimination parameter, there can be problems with exposure for certain ranges of item difficulty. A constrained beta method, a Rasch analogue of *b*-stratified adaptive testing, to control exposure in a target difficulty range was investigated in this paper.

Alternatively, conditional selection strategies impose an exposure control parameter for each item, given that it is selected for potential administration. The Sympson-Hetter method (Sympson & Hetter, 1985) applied a probabilistic model to constrain the maximum value of items being administered below the target rate. The exposure control parameter obtained via iterations for each item was used to determine whether the item can be administered or not. Detailed information about this method and modifications of this procedure were reviewed in Georgiadou, Triantafillou and Economides (2007). The most recent item exposure control method was presented by Barrada, Veldkamp, and Olea (2009); this multiple maximum exposure rate (r_{max}) method defined as many values of r_{max} as the number of items. A constrained maximum exposure rate design was also examined in the present study.

Purpose

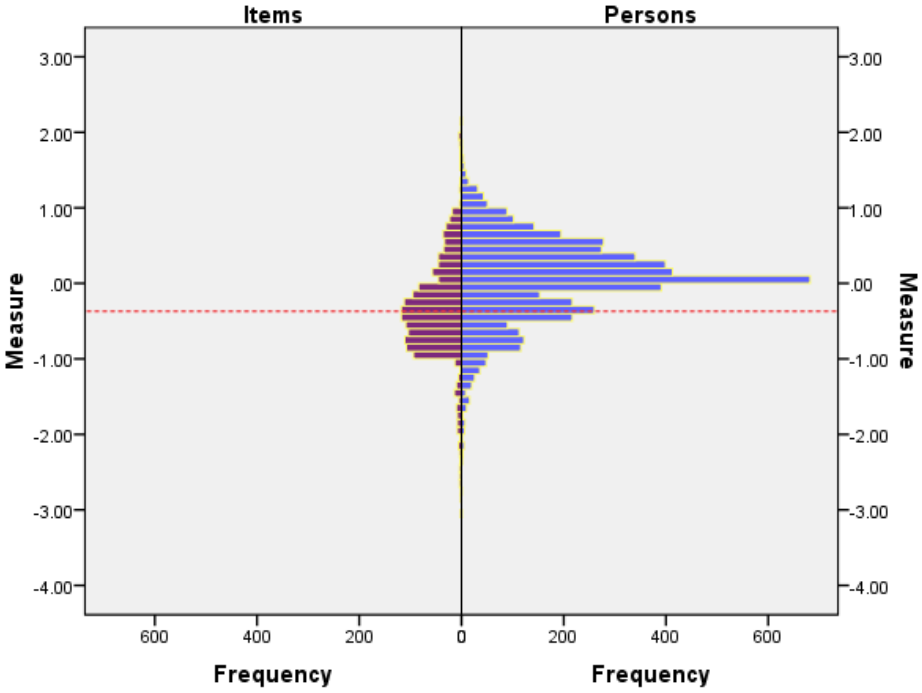
Numerous studies have been conducted to evaluate the effectiveness of a variety of algorithms that modify the CAT selection process to control item exposure. The strengths and weaknesses have been discussed for different models using dichotomous scoring, polytomous

scoring, and testlet-based CAT systems. However, no studies have focused on exposure of items within a particular difficulty range, especially those items with difficulty level near the cut score on variable-length CATs. The CAT algorithm tends to excessively administer these items under maximum item information selection. Overexposure of an item might affect the item parameter estimates and potentially the integrity of the test. This research investigated multiple methods for limiting exposure of items near the cut and evaluated the results for measurement precision.

Method

Response data from a large-scale live CAT licensure exam directed by The National Council of State Boards of Nursing (NCSBN) were used to obtain the known item parameters for simulation. θs of 5,000 simulees were randomly selected and had mean ability of .06 and standard deviation (SD) of .55. A total of 1,572 items were obtained from the live operational bank; their mean difficulty –.34, 1,420 (90%) of the items had difficulties between –1 and 1, and 754 items (48%) had difficulties between –.5 and .5. Figure 1 presents the distribution of item difficulty and examinee θ. The distribution of θ was representative of the population in which around 80% of candidates passed the examination. The cut-off value for passing was –.37 logits and is depicted by the red dotted line.

Figure 1. Distribution of Item Difficulty and Person Ability



Items were also selected to meet the target test plan for eight content areas, and their corresponding percentages were distributed as shown in Table 1. As a variable-length test, the minimum number of operational items was 60 and the maximum test length was 180 items. The Item calibrations and θ estimates were derived with the Rasch model.

Table 1. Target Test Plan for the Simulations

Content Area	I	II	III	IV	V	VI	VII	VIII
Percentage	15	11	10	11	14	12	13	14

The simulation study used four procedures to control item exposure:

1. The "randomesque" method (Kingsbury & Zara, 1989) that randomly selected one item from the most informative 15 or 25 items (represented as Random_15 and Random_25).
2. The second, proposed by Lunz and Stahl (1998), selected all items at random within the given range of logits. Three randomization intervals were deployed including .10, .20, and .30 logits (represented as Logit_ .1, Logit_ .2, and Logit_ .3).
3. The constrained beta method limited the exposure of items with difficulty within the target range at the early stage of the test when the θ estimates were not accurate so that they could be saved for use later at the most beneficial point in testing. In this case, items within the target difficulty range (between −.5 and .5) were blocked from the first 10, 30 or 60 items being administered (represented as Constrain_10, Constrain_30, and Constrain_60).
4. The constrained maximum exposure rate method which was designed to freeze items in the selection algorithm if their exposure rate exceeded the target maximum value. For this simulation, .1 and .15 were used as the maximum rate of item usage. The minimum sample size for computing the exposure rate for each item was 10. The exposure rates of items were evaluated for the first 10, 30, 60, and 180 items (represented as Max_ .1_10, for example). If they were above the tolerable rate, these items would be removed from administrations.

These methods were compared to a baseline condition with no exposure control. However, content balance was still implemented for the no control condition throughout the test to meet the target percentage for each content area. The performance of these procedures was assessed for measurement precision by the standard error of measurement (SEM). Other variables associated with test security included exposure rates and utilization of the item bank across test administrations.

Results

Table 2 presents the item exposure statistics by each design for those items with difficulties between −.5 and .5, which was considered as the target difficulty range in this case. The highest rate was .224 indicating that 22.4 % of candidates saw this highly exposed item. The average exposure rates for target range items using randomization strategies were higher than the other two methods; however, they had significantly lower maximum exposure rates compared to the constrained beta. In particular, the within-logit method had the lowest maximum exposure rate and the maximum rates were below .15 for all three randomization intervals. The constrained maximum exposure rate design was relatively low for both mean and maximum exposure rate.

The standard deviation (SD) of the exposure rate provides an indication of the uniformity of item exposure. A lower SD indicates a more uniform exposure of items in the pool whereas a

high SD means some items were exposed more than others. Both randomesque and constrained maximum exposure rate method observed relatively low SD in terms of the exposure rates.

Table 2. Statistics of Exposure Rate for Items With Difficulty Between −.5 and .5

Design	Mean	Maximum	SD
No Control	.055	.224	.024
Random_15	.067	.195	.018
Random_25	.069	.179	.016
Logit_ .1	.072	.142	.022
Logit_ .2	.070	.109	.015
Logit_ .3	.069	.100	.012
Constrain_10	.069	.184	.019
Constrain_30	.054	.209	.024
Constrain_60	.055	.222	.025
Max_ .1_10	.055	.188	.021
Max_ .1_30	.055	.160	.019
Max_ .1_60	.055	.125	.015
Max_ .1_180	.055	.101	.011
Max_ .15_10	.054	.191	.021
Max_ .15_30	.055	.151	.019
Max_ .15_60	.055	.123	.015
Max_ .15_180	.056	.106	.011

In general, the average exposure rates were close across different levels for each condition. However, the item exposure statistics had decreasing maximum exposure rates as the randomization parameter increased or the number of items allowed for constraint increased. There was also less variation in terms of item exposure rate at increasing randomization intervals.

Table 3 contains additional descriptive information about the overall exposure rate. The randomesque method was the best in terms of item usage, with most items having exposure rate between .05 and .1 but there were no items with zero exposure rate or exposure rates above .2. The within-logit method had the largest percentage of items within the range of .05 and .1 and all items had exposure rate below .15. It also had the greatest number of not-administered items. Both constrained beta and maximum exposure rates were effective in controlling items from being excessively selected, but they had one-third of items with exposure rates below .05, suggesting low efficiency in terms of item bank utilization.

Table 3. Distribution of Exposure Rates (ER) for Items With Difficulty Between –.5 and .5

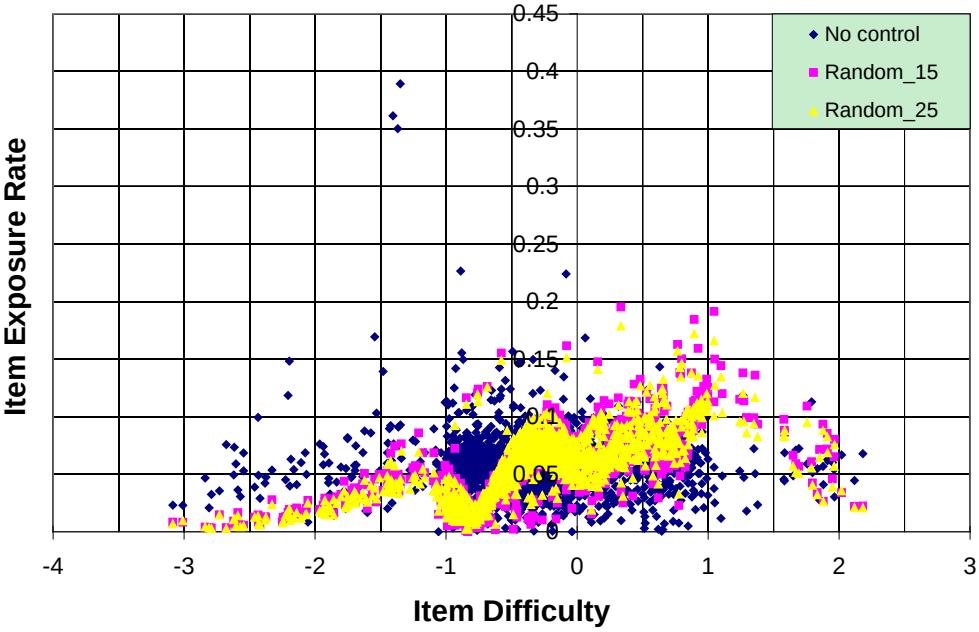
Design	ER=0	0<ER< .05	.05≤ER< .1	.1≤ER< .15	.15≤ER< .2	ER≥ .2
No Control	1	323	404	23	2	1
Random_15	0	92	635	25	2	0
Random_25	0	68	667	17	2	0
Logit_ .1	15	74	595	70	0	0
Logit_ .2	7	31	710	6	0	0
Logit_ .3	5	28	720	1	0	0
Constrain_10	0	71	656	21	6	0
Constrain_30	1	341	386	20	5	1
Constrain_60	2	335	387	24	5	1
Max_ .1_10	0	318	415	20	1	0
Max_ .1_30	0	303	438	12	1	0
Max_ .1_60	0	296	453	5	0	0
Max_ .1_180	0	229	524	1	0	0
Max_ .15_10	0	323	411	19	1	0
Max_ .15_30	0	312	430	11	1	0
Max_ .15_60	0	290	459	5	0	0
Max_ .15_180	0	223	530	1	0	0

Graphs presenting the distribution of item exposure rate against item difficulty for each design are shown in Figure 2. As expected, the no exposure control condition had extreme exposure rates as high as .39. Both randomesque and logit methods had evenly distributed item exposure rates between .05 and .1 for the target range of item difficulty. The exposure rates turned out to be reasonable even for items with moderately high or low difficulty. All items had exposure rates below .2 for the randomesque method and .25 for the within-logit method.

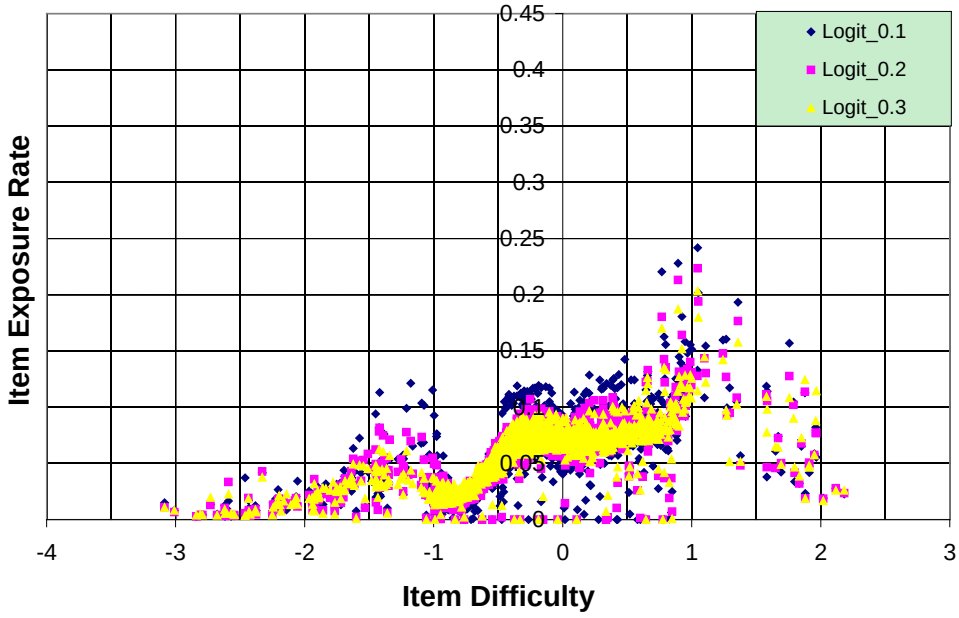
The constrained beta and constrained maximum exposure rate methods showed similar patterns in terms of the distribution of item exposure rates. With regard to those items with difficulty between –.5 and .5, most of their exposure rates were between zero and .1. However, there were a number of items with significantly high exposure rates around .4 indicating that almost two-fifths of the examinees had seen these items.

Figure 2. Scatterplots of Item Exposure Rate by Item Difficulty

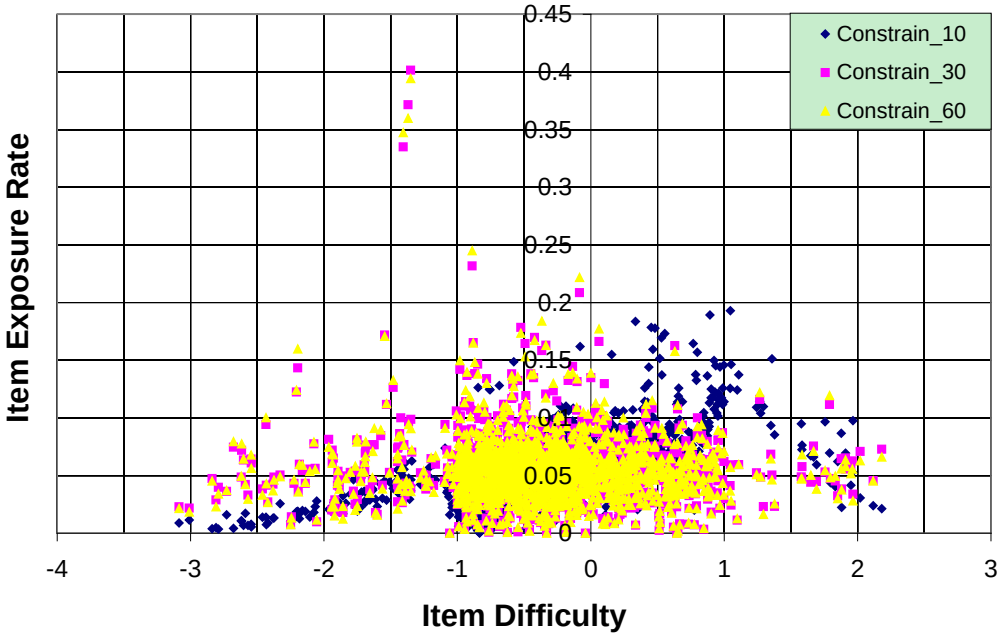
a. Randomesque Strategy and No Exposure Control



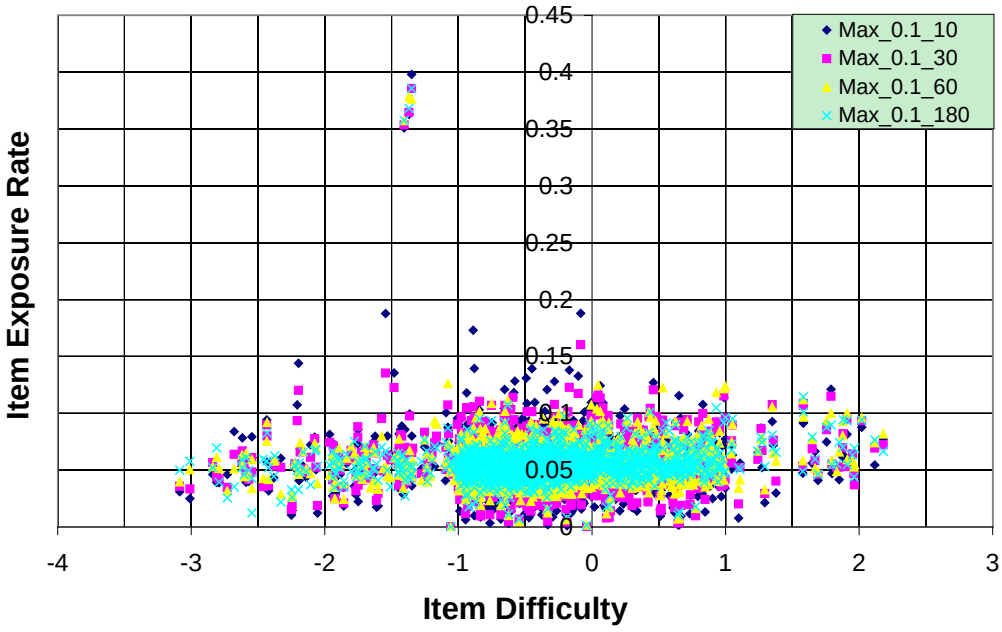
b.Within-Logit Strategy



c. Constrained Beta Strategy



d. Constrained Maximum Exposure Rate of .1



e. Constrained Maximum Exposure Rate of .15

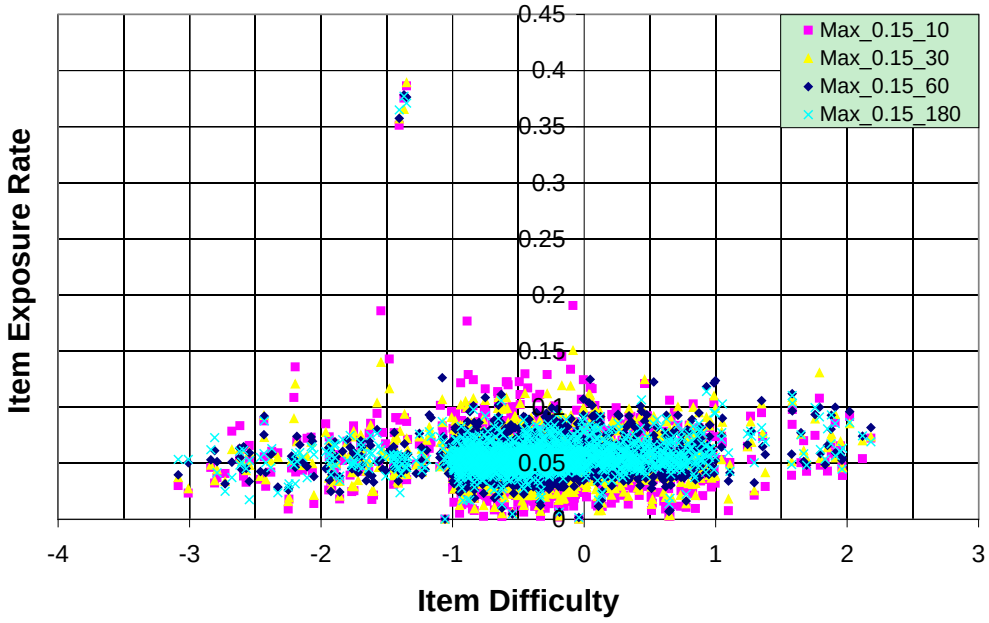


Table 4 summarizes the overall measurement precision including bias, mean square error (MSE) and SEM. Reliability was computed as the correlation between the estimated θ and the true θ values. Better item selection methods are reflected in higher reliabilities. In general, all designs yielded similar reliability, which were all above .9. Both randomization methods had the highest reliabilities, suggesting the θ estimates were most highly associated with their true values.

Bias and MSE provided more performance characterizations in terms of θ estimation. They were also comparable across all designs, but those from the randomization methods were relatively lower. The statistics of SEM showed a similar pattern. In particular, those obtained from the randomization method were lower in terms of minimum and average SEM, and exceptionally low in terms of maximum SEM.

Table 4. Overall Reliability, Bias, MSE and Descriptive Statistics for the SEM

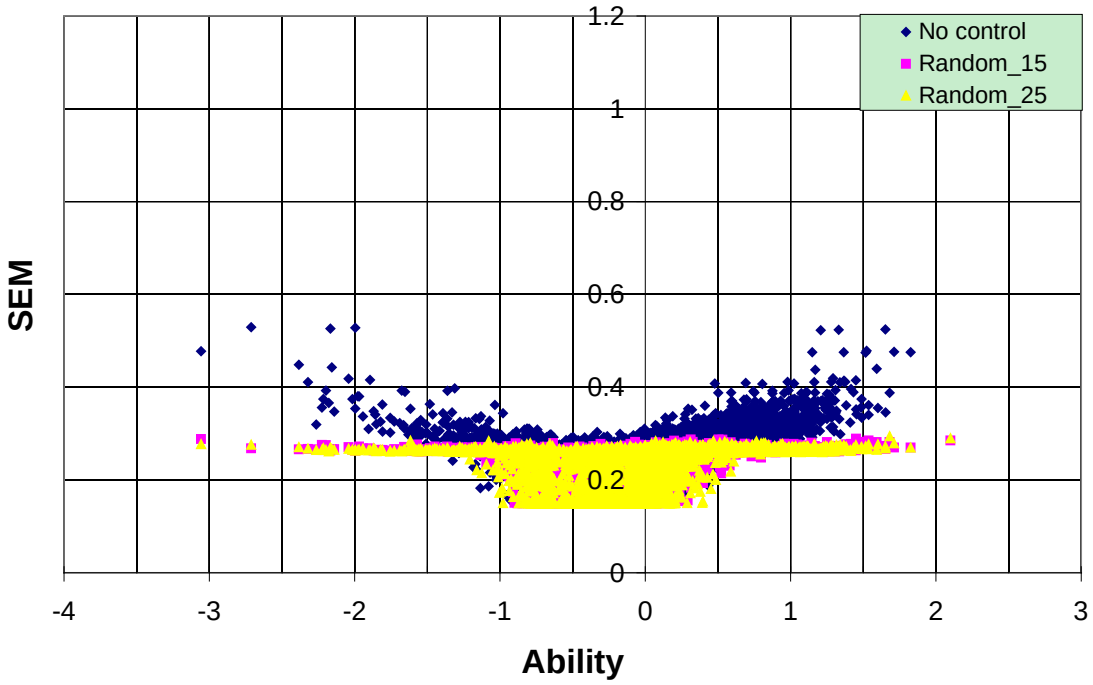
Method	Reliability	Bias	MSE	Min SEM	Mean SEM	Max SEM
No Control	.908	.026	.070	.154	.258	1.013
Random_15	.919	.016	.059	.149	.238	.289
Random_25	.920	.017	.056	.149	.236	.294
Logit_ .1	.923	.022	.056	.149	.237	.312
Logit_ .2	.922	.022	.056	.149	.237	.366
Logit_ .3	.919	.022	.057	.150	.237	.329
Constrain_10	.922	.019	.057	.149	.237	.301
Constrain_30	.906	.025	.073	.154	.259	1.011
Constrain_60	.910	.024	.069	.154	.259	.601
Max_ .15_10	.903	.027	.075	.154	.259	1.015
Max_ .15_30	.909	.028	.070	.154	.257	.608
Max_ .15_60	.911	.028	.067	.154	.256	.609
Max_ .15_180	.909	.030	.069	.154	.255	.735
Max_ .1_10	.908	.019	.068	.154	.257	.604
Max_ .1_30	.909	.027	.068	.154	.257	.606
Max_ .1_60	.910	.020	.068	.154	.256	.735
Max_ .1_180	.905	.027	.071	.154	.256	.733

Figure 3 shows the distributions of SEM versus true values of examinee θs. Both randomization methods provided optimal precision for candidates with true θs close to the passing point. Even for candidates with the highest or the lowest θs, the SEM was quite low. The standard errors appeared low correspondingly for candidates with medium level of θ for the other two methods. However, the SEM increased dramatically as the θs approached either the upper or lower end. There existed cases with substantial SEMs that were as large as 1.

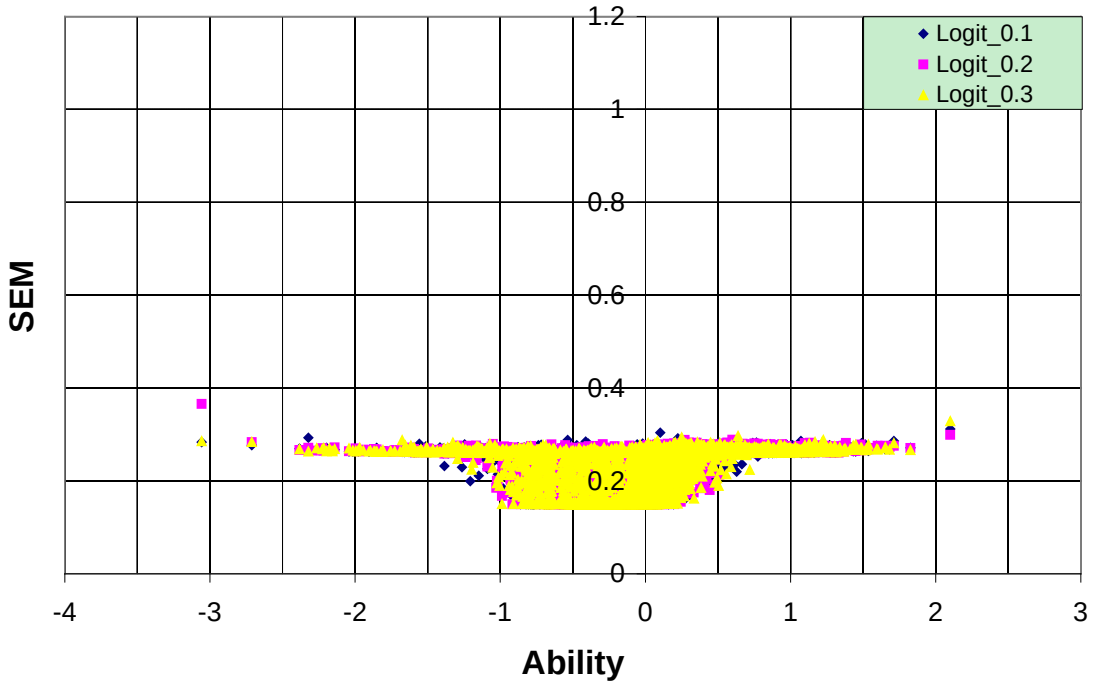
For each design, the scatterplots nearly overlapped, indicating no evidence of the distinction in terms of SEM as the randomization parameter varied. The only exception was the constrained beta for the first ten items, which resulted in consistently low SEMs for candidates with θs across all levels.

3. Scatterplots of SEM by θ

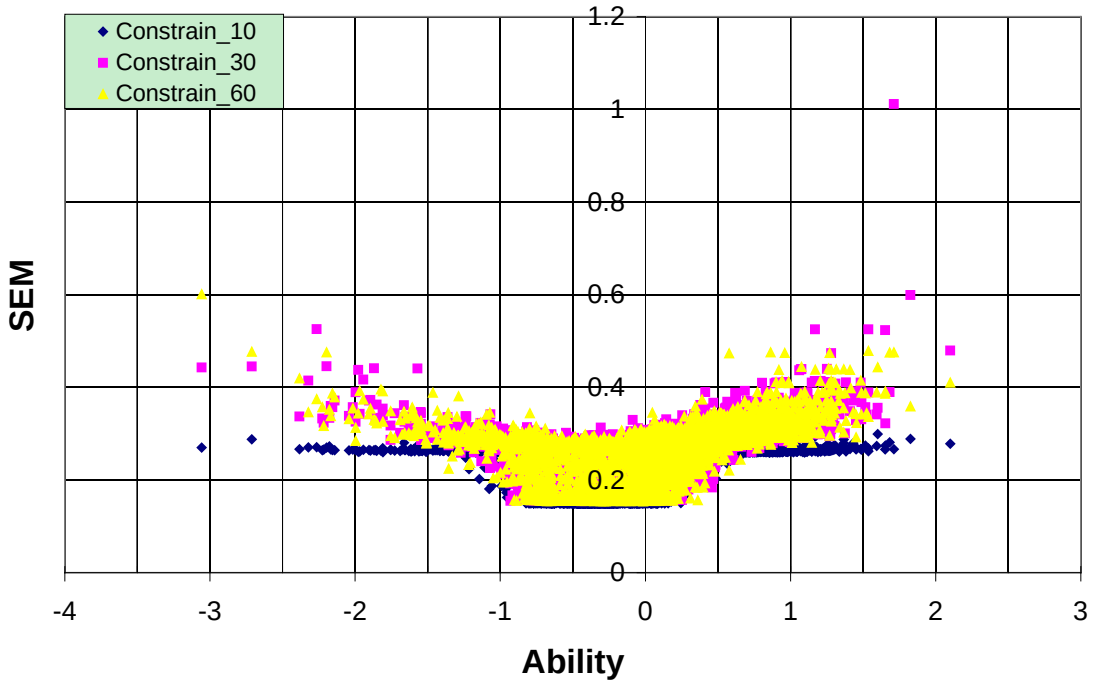
a. Randomesque Strategy and No Exposure Control



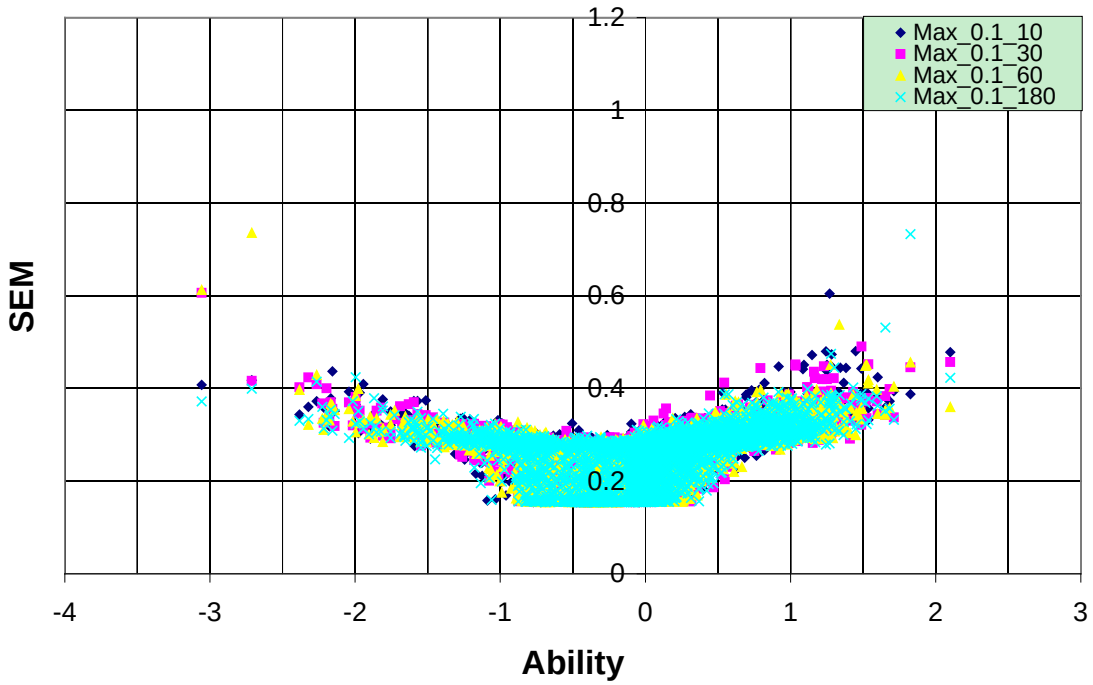
b. Within-Logit Strategy



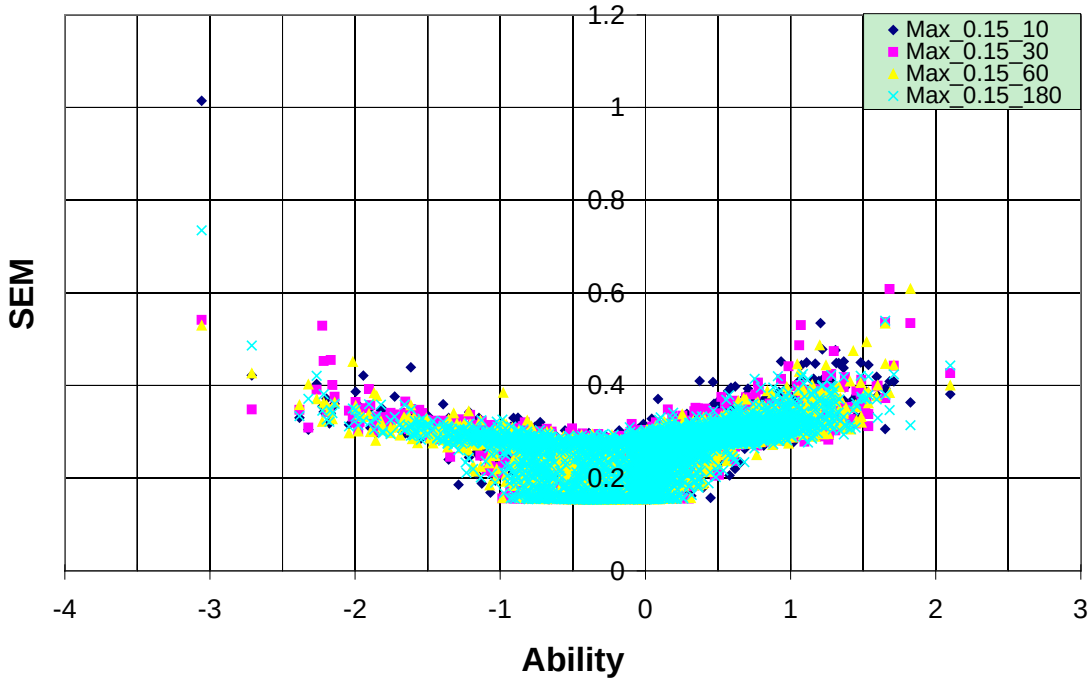
c. Constrained Beta Strategy



d. Constrained Maximum Exposure Rate of .1



e. Constrained Maximum Exposure Rate of .15



Discussion and Conclusions

This simulation study focused on comparing four methods for controlling exposure for items within a target range of difficulty while balancing measurement precision. All items were selected carefully according to the content in order to exactly match the target test plan. All methods performed well for controlling exposure of target difficulty range items. Both randomization strategies constrained exposure rate between .05 and 01 with the most optimal usage of item pools, while the within-logit technique had the most unexposed items. Constrained beta and maximum exposure rate methods had one-third of items exposed to less than 5 percent of candidates and a few items to more than 15 percent of candidates.

Estimates of θ for all methods were highly correlated to the true values. Randomization methods maintained measurement precision with comparatively low SEM even for extreme levels of θ and the overall bias and mean square error were the lowest of the methods for estimated θs. Constrained beta and maximum exposure rate strategies showed low standard errors for the middle range of θ but precision decreased substantially as θ approached the extremes and there were outliers with considerably high SEM.

The randomesque method, which is used in practice, proved to be effective for controlling exposure of items around the cut score, while achieving high measurement precision. A random selection from the 25 most informative items provided a better solution than that selected from 15 most informative items. The within-logit method had equivalent patterns but optimized the measurement precision with fewer items. A well-targeted item bank played an important role and might compromise the impact of item exposure constraints upon measurement precision across different designs. Test developers should make decisions in selecting the most rational strategy for controlling exposure. Future research can explore other exposure control techniques and the impact upon pass/fail decisions given overexposed items around cut scores.

References

Barrada, J. R., Veldkamp, B. P., & Olea, J. (2009). Multiple maximum exposure rates in computerized adaptive testing. *Applied Psychological Measurement, 33*, 58-73.

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

Georgiadou, E., Triantafillou, E., & Economides, A. (2007). A review of item exposure control strategies for computerized adaptive testing developed from 1983 to 2005. *Journal of Technology, Learning, and Assessment, 5*(8).

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

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.

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

Way, W. D. (1998). Protecting the integrity of computerized testing item pools. *Educational Measurement: Issues and Practice, 17* (4), 17-27.