

Impact of Altering Randomization Intervals  
on Precision of Measurement and Item Exposure

Timothy Muckle  
Betty Bergstrom, Ph.D.  
Kirk Becker  
John Stahi, Ph.D.

Promissor, Inc.

Correspondence concerning this article should be addressed to Timothy Muckle,  
Promissor, Inc., 1007 Church Street, Suite 314, Evanston, Illinois, 60201, e-mail:  
Timothy\_Muckle@Promissor.com.



## Impact of Altering Randomization Intervals on Precision of Measurement and Item Exposure

### Abstract

Item exposure and precision of measurement are critical aspects of computer adaptive testing. The purpose of this study is to evaluate the effects on bank usage (item exposure) and precision of measurement (standard error of measurement) when randomization intervals are relaxed. Recent studies suggest the efficacy of using an item selection procedure in which items are randomly selected from a group of items within a logit interval surrounding the estimated ability (also known as a *randomesque* procedure) for adaptive testing. Each subsequent item is chosen from all items within a certain specified logit distance of the target information value (Laughlin Davis, Dodd, et al., 2003). Presently, no study has been conducted to determine the optimum size of the logit interval from which items are selected. This project consists of studying the effects, specifically the impact on item exposure and measurement precision, of altering the width of this random selection interval. *Index terms: randomesque, item exposure, adaptive testing, simulation, measurement precision.*



## **Introduction**

Successful adaptive testing requires an item bank that is well targeted to the test-taker population. It also presents the challenge of balancing conflicting principles of maximizing measurement precision versus minimizing item exposure (Parshall, Davey and Nering, 1998) and optimizing item usage (Lewis, 2001). This study relates to the use of adaptive testing for criterion-referenced tests for certification and licensure. We report on an item exposure procedure in which each subsequent item is randomly selected from a group of  $n$  most informative items (also known as a *randomesque* procedure) (Featherman, Subhiyah & Hadadi, 1996; Davis, 2004; Dodd, 2003). Randomization strategies operate by randomly choosing the next item from a set of nearly optimal items instead of the single most informative item (Way, 1998). We present a simulation methodology to determine the optimum width of the interval from which items are selected and we report on the impact of relaxing the interval width on measurement precision and item exposure. We discuss the value of the simulation procedure for identifying the optimal randomization interval width given the particular distribution of the item bank.

Control of item exposure rates is essential to successful computer based, criterion referenced testing, and most testing programs monitor the extent to which active test questions are exposed to the testing population. An overexposed item may be compromised, potentially affecting the difficulty of the item and, if enough items are overexposed, the integrity of the test as a whole. Overexposure of an item bank can also translate into real item development costs to replace the compromised items.

Minimizing the Standard Error of Measurement (SEM) is also an important element of criterion referenced-testing. Too much error can distort measurement of test taker ability, potentially affecting the decision about whether a candidate has demonstrated a level of



competence sufficient to practice in their profession. The larger the error, the less precise the measurement.

In this paper, item exposure rates were studied for an adaptive test in which item calibrations and test taker ability were derived with the Rasch model (Wright & Stone, 1979; Gershon, 2005). Item exposure was regulated by a randomization algorithm set to choose an item within a specified logit range of the current estimate of test taker ability. As the test progresses, each item is selected randomly from a subset of items within a specified logit interval of the current estimate of the test taker's ability. This study evaluates the impact of widening this interval, thus increasing the size of the subset of items from which the subsequent item is drawn. We compare a .10 logit randomization interval previously used in adaptive tests for certification (Bergstrom & Lunz, 1999) with wider intervals. The .10 logit interval commonly used is an arbitrary number (as in other random procedures choosing from the best 5 or 10 is an arbitrary decision). We demonstrate that increased randomization may foster better bank usage and limit item exposure, with minimal sacrifices to information and precision of measurement.

### **Method**

The adaptive simulations in this study involved interactions between 1205 simulated test takers (Mean ability=1.93; SD=0.78; Mode=2.5) with an n=520 item pool (mean difficulty= 1.39; SD=1.14). Figures 1 and 2 show the distributions of the item difficulty and the test taker ability. Both the item distribution and the test taker distribution were chosen because they are actual client distributions. The adaptive test program that they were derived from has been administered since 1996. The test taker distribution is representative of certification test taker distributions in which many of the candidates pass the examination and the distribution is negatively skewed. The item distribution represents an item bank that has been sculpted to



provide a well-targeted bank for adaptive testing for this test taker population. The pass point on the actual examination is 1.27 logits.

In the simulated tests, 70 dichotomously scored items were administered in a fixed-length adaptive test. Each test was balanced across five content areas. The content areas and their percentages are represented in Table 1. Simulations of the 1205 test takers were repeated for increasing randomization intervals (0.10, 0.50, 1.0, 1.5, 2.0, 2.5, 3.0 logits), for a total of 8435 simulated response strings. The first item was selected from items within the randomization interval of the passing standard (1.27 logits). The difficulty of the first ten items was constrained to within .10 logits of the previously administered item difficulty plus or minus the randomization interval. As the simulation proceeds, all unused items within the specified logit distance of the most informative (or target) item difficulty are available for selection, and the next item to be administered is randomly chosen from among them.

Two measures of the efficacy of this procedure were reviewed: 1) Item percent exposure and 2) Test taker SEM. For each item in the pool, the number of times it was administered to a simulee test taker was recorded. Percent exposure was calculated by dividing the number of test takers who saw an item by the total number of test takers, rendering a proportion of the population who saw an item. We chose 25% or less as tolerable item exposure (van der Linden, 2004).

Standard Error of Measurement (SEM) was used as an indicator of measurement precision. SEM was calculated as the root inverse of test information:

$$\frac{1}{\sqrt{\sum_{j=1}^J p_{ij}(1-p_{ij})}} \quad (1)$$

where  $p$  is the model-defined probability of a correct response by simulee, given by:

$$p_{ij} = \frac{e^{(\theta_i - \delta_j)}}{1 + e^{(\theta_i - \delta_j)}} \quad (2)$$



where  $\hat{\theta}_i$  is the estimated theta for simulee  $i$  and  $\delta_j$  is the difficulty of item  $j$ . The SEM is calculated from the final estimate of theta using a maximum likelihood algorithm (Wright & Stone, 1978) applied to the 70 items administered to a simulee.

## **Results**

### **Item Exposure**

Table 1 shows the item exposure statistics by item selection intervals including the average, maximum, and standard deviation percent exposure, for each condition. The average exposure rate remains invariant (13.5%). Chen, Ankenmann, and Spray (1999) explained that the average exposure rate for any fixed-length test would always be constant and equal to the ratio of test length to pool size ( $70/520 = .135$ ). Because test length and the item pool were the same for all conditions studied, the observed average exposure rates did not differ across conditions. The maximum exposure rate (reflecting the most highly exposed item) decreases with every increase in the randomization interval. For the 3.0 randomization interval, the maximum exposure rate is 20% (20% of the testing population saw the most highly exposed item), compared to 53% at the 0.1 randomization interval. Hence, widening the randomization bandwidth resulted in a 33% reduction in exposure for the most highly exposed items. Also, note that the number of overexposed items (items with an exposure rate above 25%) decreases dramatically with a randomization interval greater than 1.0 logit. 95 items were overexposed in the first run of the experiment (0.1 logit bandwidth) while 0 items were overexposed when the randomization interval is greater than or equal to 2.0.

Figure 3 below depicts a series of graphs in which item difficulty is plotted against percent exposure for each test condition. When the randomization interval is small, the exposure rates peak at the item difficulty corresponding to the mode of the ability distribution



(refer to Figure 2) indicating that item exposure is greatest at the ability level where the most test takers are performing. Economy of item bank usage is optimal when the peaks in the item exposure plots are minimized. In Figure 3, the peaks in exposure decrease as the randomization interval is expanded. Items near the middle of the distribution are used less often while items at the extremes of the distribution are used more often, resulting in a smoother distribution of item exposure.

### **Precision of Measurement**

Table 2 shows a summary of SEM at the pass point (1.27 logits) and at the least able and most able extremes of theta. Figures 4 and 5 depict mean and max SEM vs.  $\hat{\theta}$  where  $\hat{\theta}$  is aggregated at .1 intervals. These figures show that for all conditions, standard error was lowest for the middle range of the  $\hat{\theta}$  distribution, with degradation to measurement precision increasing as  $\hat{\theta}$  became more extreme. Since the goal of this examination is to make a pass/fail, criterion referenced decisions, the item bank is targeted to provide optimal precision at the pass point. The lower and higher end of the item bank are less well targeted to test taker ability, resulting in larger standard errors. Table 2 and figures 4 and 5 show that, for test takers near the pass point, SEM degrades from 0.24 to .0.27 as the randomization interval is widened from 0.1 to 3.0., an increase in error of only .03 logits. SEM is more affected at the extremes of the ability distribution. At the lower extreme of  $\hat{\theta}$ , the mean SEM varies from .29 to .44 (increase of .15 logits). At the high end of the  $\hat{\theta}$  distribution, mean SEM increases from .25 to .32 (increase of .07 logits).

### **Discussion**

This paper demonstrates the value of simulation for determining the optimal conditions for adaptive test parameters. Adaptive testing is always modulated by the interaction between the distribution of the test taker population and the distribution of the item bank. In this case, the



purpose of the adaptive test is to make accurate pass fail decisions with a criterion referenced item bank. Because the item bank has been continuously updated and is well targeted about the pass point, the randomization interval can be widened without seriously impacting the error of measure for test takers whose ability estimate is near the pass point. Poor performing test takers will be less accurately measured and while their fail decision will remain the same, diagnostic feedback on their performance will be less accurate. Widening the randomization interval provides more even item exposure across items in the bank. Simulation allows the test developer to make reasonable decisions to maximize the use of the item bank while still maintaining test reliability.



## References

- Bergstrom, B. A., & Lunz, M. E. (1999). CAT for certification and licensure. In F. Drasgow & J. B. Olson-Buchanan (Eds.), *Innovations in Computerized Assessment* (pp. 67-92). Mahwah, NJ: Lawrence Erlbaum Associates Publishers.
- Chen, S.-Y., Ankenmann, R. D., & Spray, J. A. (1999). Exploring the relationship between item exposure rate and test overlap rate in computerized adaptive testing. National Council on Measurement in Education annual meeting. Montreal, Canada.
- Davis, Laurie Laughlin (May, 2004). Strategies for Controlling Item Exposure in Computerized Adaptive Testing With the Generalized Partial Credit Model. *Applied Psychological Measurement*, 28(3), 165-185.
- Dodd, Barbara G., et al. (2003). An Examination of Exposure Control and Content Balancing Restrictions on Item Selection in CATs using the Partial Credit Model. *Journal of Applied Measurement*, 4(1), 24-42.
- Featherman-Morrison, C. A., Subhiyah, R. G., & Hadadi, A. (1996, April). The effects of item pool size on CAT proficiency estimation and classification accuracy. National Council on Measurement in Education annual meeting. New York.
- Gershon, R. (2005). Understanding Rasch Measurement: Computer Adaptive Testing. *Journal of Applied Measurement*, 6(1), 109-127.
- Lewis, D. M. (2001, April). Standard setting challenges to state assessments: Synthesis, consistency, balance, comparability. National Council on Measurement in Education annual meeting. Seattle, WA.
- Parshall, C. G., Davey, T. A., & Nering, M. L. (1998, April). Test development exposure control for adaptive testing. National Council on Measurement in Education annual meeting. San Diego, CA.
- Way, W. D. (1998, Winter). Protecting the integrity of computerized testing item pools. *Educational Measurement: Issues and Practice*, pp. 17-27.
- Wright, B. D., & Stone, M. H. (1979). *Best test design*. Chicago, IL: MESA Press.



Table 1. Item Exposure Statistics by Item Selection Interval

Randomization interval	Maximum	Mean	Std. Deviation	# overexposed items
0.1	52.50%	13.46%	11.20%	95
0.5	50.20%	13.46%	11.14%	94
1.0	29.80%	13.46%	8.08%	61
1.5	25.70%	13.46%	6.36%	3
2.0	22.20%	13.46%	5.01%	0
2.5	20.00%	13.46%	4.03%	0
3.0	20.00%	13.46%	3.55%	0



Table 2. Summary of SEM at Pass Point and Extremes of  $\hat{\theta}$

Randomization interval	$\hat{\theta}$ low extreme		Pass Point		$\hat{\theta}$ high extreme	
	Mean	Max	Mean	Max	Mean	Max
0.1	0.29	0.29	0.24	0.25	0.25	0.26
0.5	0.29	0.29	0.24	0.24	0.25	0.26
1.0	0.31	0.31	0.25	0.25	0.26	0.26
1.5	0.31	0.31	0.26	0.26	0.27	0.28
2.0	0.31	0.31	0.26	0.27	0.29	0.3
2.5	0.35	0.35	0.26	0.27	0.3	0.34
3.0	0.44	0.44	0.27	0.27	0.32	0.37



Figure 1. Item Difficulty Distribution

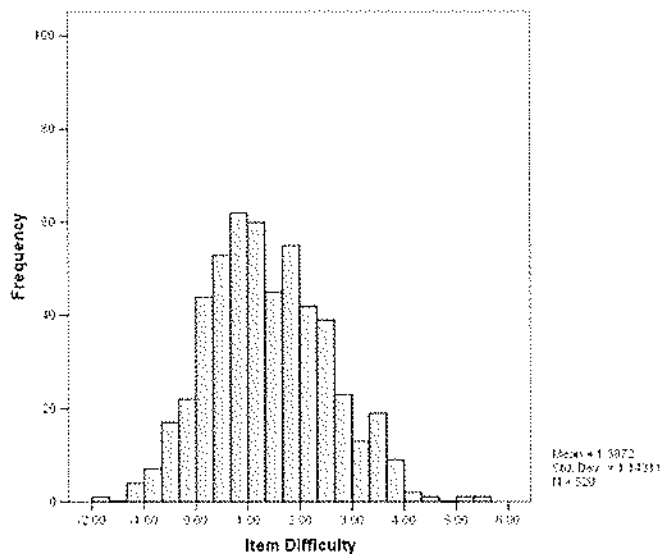




Figure 2. Test Taker Distribution

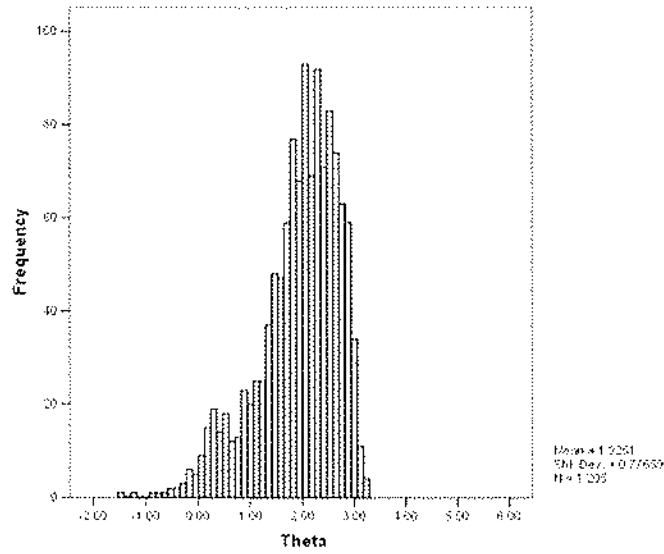




Figure 3. Item Exposure Scatterplots by randomization intervals

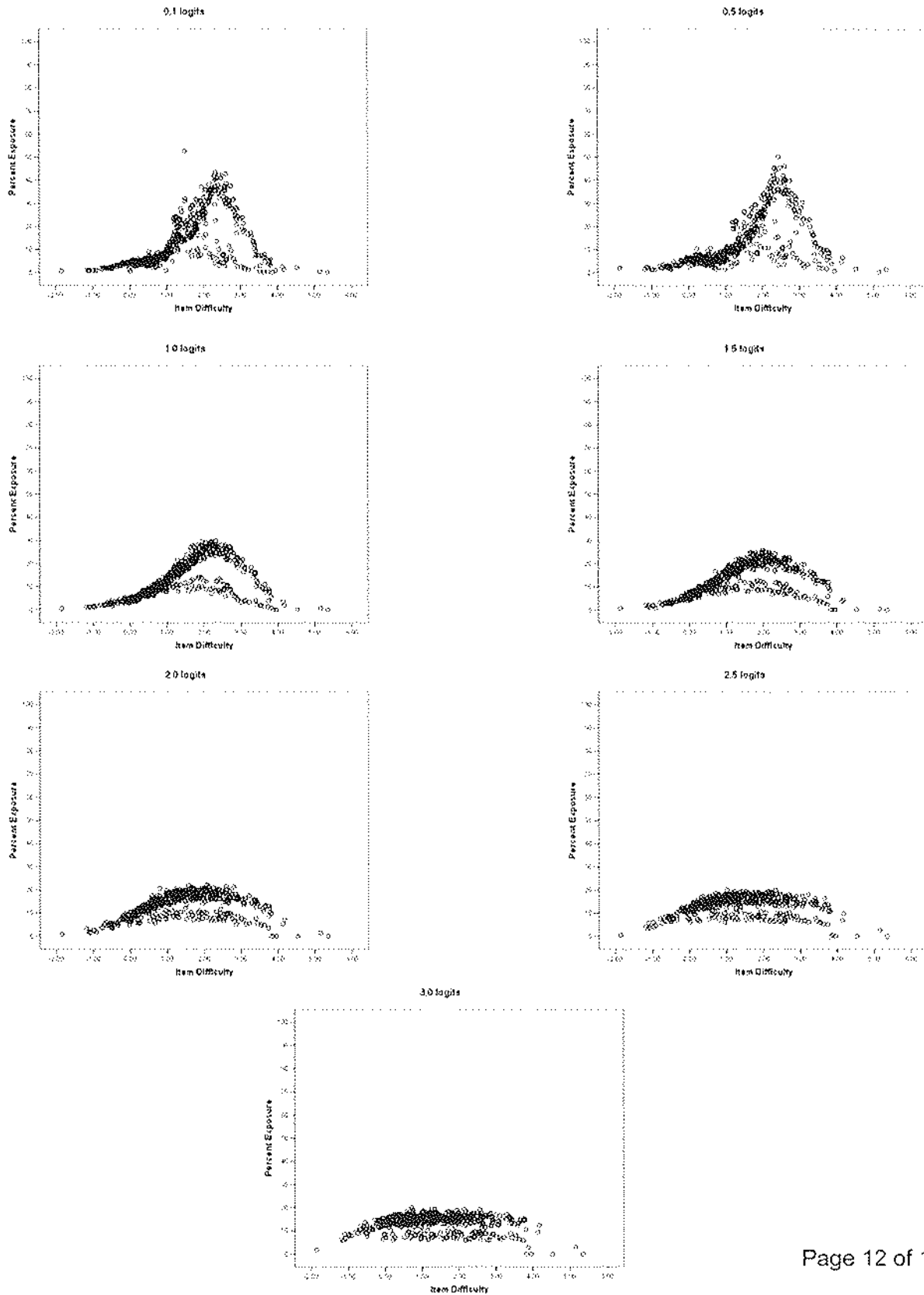




Figure 4. Plot of Mean SEM vs.  $\hat{\theta}$ , by randomization interval

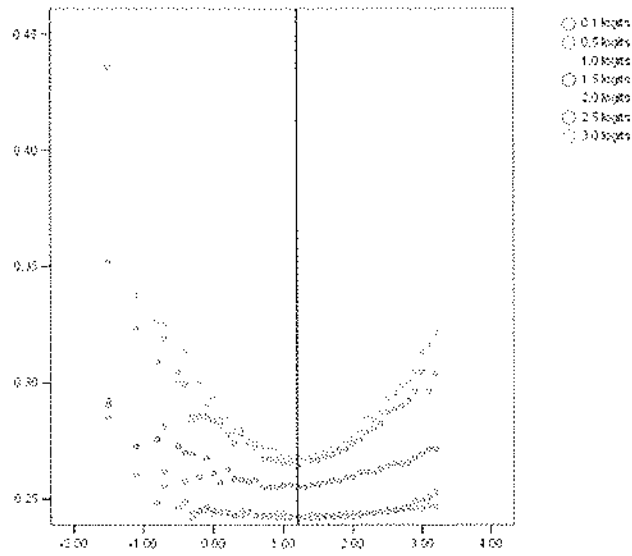
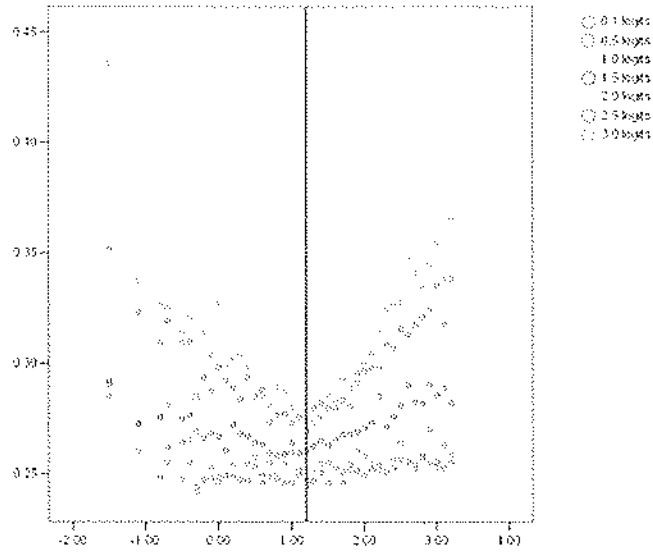




Figure 5. Plot of Max SEM vs.  $\hat{\theta}$ , by randomization interval





### Acknowledgements

Thanks to James Masters, David Hoadley, Deb Schnipke for their helpful comments on earlier versions of this paper. Thanks also to Kirk Becker for designing and coding the computer simulation program. Special thanks are due to the Council on Certification of Nurse Anesthetists, whose certification testing program provided the basis for the questions addressed in this study. This paper was presented at the 2005 annual meeting of the American Educational Research Association, Montreal, Canada.



Author's Address

Correspondence concerning this article should be addressed to Timothy Muckle, Promissor, Inc., 1007 Church Street, Suite 314, Evanston, Illinois, 60201, e-mail: [Timothy\\_Muckle@Promissor.com](mailto:Timothy_Muckle@Promissor.com).

