Balancing Security and Efficiency in Limited-Size Computer Adaptive Test Libraries

Cory Moclaire
KSH Solutions/Naval Aerospace Medical Institute

Eric Middleton
Naval Aerospace Medical Institute

Brennan D. Cox
Naval Aerospace Medical Institute

Chris Foster
Naval Aerospace Medical Institute

Thomas Prettyman
KSH Solutions

The majority of studies focusing on enhancing item bank security and measurement efficiency in computer adaptive tests (CATs) have featured large item banks consisting of thousands of items. For many practitioners of high-stakes tests, CATs are a viable alternative to static forms, but inadequate resources exist for developing expansive item libraries. Practitioners may need to consider alternative solutions for maintaining CAT integrity and proficiency. This study documents the effects of changing the size of minimum eligible item pools for selection on test length, maximum exposure frequency, and total item usage across four operational item banks of various size and quality.

INTRODUCTION

In the near future, the Navy Medicine Operational Training Center will make available variable length Computer Adaptive Test (CAT) versions of the Aviation Selection Test Battery (ASTB), the primary cognitive screening tool used to select student aviators and flight officers for the United States Navy, Marine Corps, and Coast Guard. The ASTB consists of the Math Skills Test (MST), Reading Comprehension Test (RCT), Mechanical Comprehension Test (MCT), and Aviation and Nautical Information Test (ANIT). In its current format, the ASTB consists of three static parallel forms. Each subtest is scored on a -4.0 to +4.0 metric using expected a posteriori (EAP) theta estimation in an item response theory framework (Phillips, 2004). The ASTB is administered to approximately 10,000 live aviation and officer applicants yearly (Moclaire, Middleton, & Phillips, 2011). The current study utilizes a methodology of selecting least-exposed eligible items based on Bayesian information calculated from three parameter logistics model (3PL) item parameters at the examinee's current estimated theta (ability) level using Owen's (1975) procedure. This paper focuses primarily on how the manipulation of the minimum number of items required in the CAT ASTB subtests' eligible item selection pools affected test length and exposure frequency using 10,000 computer simulations per subtest.

Three of the biggest obstacles to overcome before CAT ASTBs can be administered operationally include: maintaining maximal measurement accuracy and efficiency, maintaining the security of item banks for each subtest by limiting overall item exposure, and ensuring balanced coverage of content sub-domains between examinees within each test. These necessities can conflict with one another. For example, the items that are the most diagnostically valuable for assessing examinees at any given ability level, based on the item parameters (i.e., 'a,' or information, 'b,' or difficulty, and 'c,' or guessability), may be presented to nearly every examinee of a similar ability level (Parshall, Davey, & Nearing, 1998). This will lead to overexposure of items and may potentially compromise the test banks unless exposure controls are assigned to a CAT's item selection algorithm, which in turn has the possibility of jeopardizing the efficiency of measurement. In essence, prior studies have demonstrated a trade-off where measurement precision comes at the expense of reduced item exposure control (e.g., Way, 1998).

Several researchers have explored alternative methods of controlling item exposure frequency while simultaneously yielding accuracy in assessments, but no methodology has been without a unique set of shortcomings. Studies conducted by Revuelta and Ponsoda (1998) as well as van der Linden and Veldkamp (2004) have been successful at restricting overexposure for the most frequently exposed items within CAT libraries, but little attention was paid to increasing the exposure of the most underused items. Building and implementing extensive CAT item libraries is a time and cost intensive process, so underutilizing or simply not utilizing items that still provide valuable information for assessing examinees may be viewed as suboptimal use of resources. In addition, as Barrada, Veldkamp, and Olea (2008) emphasized, certain item selection techniques can lead to a reduction in item quality as the CAT progresses (e.g., administering only the most informative items without an element of randomization will lead to the availability of fewer maximally informative items toward the test's end). Recent studies have focused on limiting item exposures at certain difficulty ranges, targeting items with 'b' parameters near the test cut score (Li, Becker, & Gorham, 2009). The ASTB subtests do not have individual cut scores, as selection decisions are made based on composite scores yielded from weighted combinations of all subtests.

The aforementioned studies featured simulated and/or operational item banks that are much larger than those available for the ASTB subtests. Alternative solutions for operationally implementing CAT tests featuring item banks consisting of fewer than 400 items have not been thoroughly explored in existing literature. The comparative lack of depth in the ASTB subtest item banks limits the options available for successful CAT implementation while simultaneously placing the best items within the item banks at an elevated risk of overexposure. However, CATs may still be a viable option, even when organizational resources are limited. As a means of assessing potential item selection algorithm settings for the MST, RCT, MCT, and ANIT, four simulated scenarios involving manipulation of the minimum number of items in the eligible item pool were enacted in an attempt to yield the most pragmatic, functional combination of:

- Low exposure frequencies for the most-exposed items within each operational subtest bank.
- As many items as possible from each subtest being exposed at least one time.
- A low total number of items administered per subtest due to exams reaching the standard error of measurement (SEM) threshold.
- As few full-length, 30-item tests as possible failing to reach the SEM threshold. The current static forms range from 27 items (RCT) to 30 items (all other subtests).

Method

3PL parameters had previously been estimated on all items for the four subtests. Table 1 presents the number of items and content areas per subtest, the maximum and minimum number of items contained within any given subtest content area, and the distribution characteristics of the ‘a,’ ‘b,’ and ‘c’ parameters. Each subtest varies in total items and content areas. The MST has the deepest and most robust item bank, with the most items overall and the mean item providing more information than the other three subtests, as indicated by the elevated mean ‘a’ parameter. The RCT has the smallest item bank, and the average item yields less information than the average MST or ANIT item. The mean MCT item is more difficult than the mean item from the other three subtests, as indicated by the elevated mean ‘b’ parameter.

TABLE 1
ASTB ITEM, CONTENT AREA, AND PARAMETER CHARACTERISTICS

Subtest	N Items	N Content Areas	M 'a'	SD 'a'	M 'b'	SD 'b'	M 'c'	SD ‘c’
MST	394	8	0.99	0.32	0.05	1.08	0.24	0.06
RCT	220	5	0.79	0.27	-0.04	1.27	0.25	0.08
MCT	328	10	0.79	0.33	0.16	1.26	0.26	0.08
ANIT	255	5	0.85	0.37	0.06	1.31	0.25	0.08

Note. A ‘c,’ or pseudo-guessing parameter value of 0.25 for a given item would indicate that there would be 25% chance that the examinee could guess the correct answer response without knowing the answer.

10,000 simulated examinees were randomly assigned a uniformly distributed theta value for all four subtests ranging from -4.0 to + 4.0 (*M*= 0.07, *SD*= 1.94). A pseudo-random number was drawn from a uniform distribution of numbers ranging from 0.0 to 1.0 to simulate the examinees’ responses for each item. If the random number was less than the probability of a correct response for the item based on the item’s 3PL parameters at the examinee’s current estimated theta quadrature node, the item was scored as correct. A detailed explanation of this process may be found in Sympson and Hetter (1985). Quadrature nodes were divided evenly into 81 bins at the tenth of a decimal place, ranging from -4.0 to +4.0.

Establishing the SEM Stopping Threshold for Present Study

Prior to examining the effect of manipulating the minimum item selection pool, exam termination rules were evaluated based on the ability to accurately estimate simulated thetas. A SEM threshold for test stopping was established by running 10,000 simulations per test at the SEM threshold cut-off levels of 0.40, 0.35, 0.30 for the MCT and RCT. All simulations were run with a minimum eligible item pool of three items for theta SEM cut-off calibration. When the SEM threshold cut-off levels were set at 0.40, examinees’ mean absolute simulated theta fluctuated by nearly a full theta point (*M* = 0.92, *SD*= 0.67 on MCT, *M*= 1.02, *SD*= 0.69 on RCT) when examining the absolute value of change on both subtests. Considering that the operational static tests under investigation have theta standard deviations ranging from 0.69 on the RCT to 0.81 on the MST (Moclaire et. al, 2011), the potential impact on selection decisions would be unacceptable if this SEM were utilized. Using a SEM of 0.35 for the CAT stopping criteria, the observed mean absolute value of the theta change was appreciably improved (*M* = 0.28, *SD* = 0.22 for MCT, *M* = 0.28, *SD* = 0.22 for RCT) for both subtests. Using 0.30 as the CAT stopping criteria, the observed mean absolute value of the change was 0.25, with standard deviations of 0.20, for both subtests. The 0.30 SEM stopping criteria only provided a marginal improvement in comparison to the 0.35 scenario. With a priority given to reduction of item exposure frequencies and test administration time, all further simulations were conducted using 0.35 as the SEM for CAT stopping, as these scenarios led to approximately two fewer questions being administered to each examinee per subtest with little loss in measurement precision efficiency. Mean SEM on the current static ASTB forms range from 0.34 on the MST to 0.37 on the RCT.

CAT Specifications and Procedures for Present Study

Item selection was partially based on variations of the 4-3-2-1 procedure originally developed by McBride and Martin (1983). In their procedure, not only is the most informative item identified, but also the second, third, and fourth most informative items. The most informative item will be exposed to 40% of the examinees, with the second, third, and fourth best items presented 30%, 20%, and 10% of the time, respectively. Though this provides a reasonable alternative for somewhat limiting item exposure of the best items, given the risk associated with over exposing items it is important to further minimize the exposure rates of the most-informative items, using item eligibility methods similar to those established by van der Linden and Veldkamp (2007). Their item-eligibility method uses a top-down approach in which exposure control is achieved by limiting the percentage of examinees for whom a given item is eligible to be exposed (e.g., no item may be administered to more than 15% of examinees). This solution would be very beneficial for tests consisting of larger item banks. The present study used a bottom-up approach, in which a minimum number of suitable items were identified for a given examinee, and exposure control was achieved by randomly selecting from the least-exposed of these eligible items. This seemed to be a wiser approach to limiting maximum exposure frequencies when item banks are relatively small. The minimum number of items in the pool of eligible items was manipulated in each of the four simulation scenarios. The minimum eligible item pools for each simulation consisted of three, four, five, or six items.

In all scenarios, the simulation completed the following steps during the process of selecting and delivering an item. First, a content area within the subtest that contains multiple items that provide maximum information near the theta quadrature of 0.0 is randomly selected. Within this content area, the items presenting the most information at a theta quadrature of 0.0 (the examinee's estimated ability before having answered any questions) are identified. The number of initial items varies from three to six, depending on the simulation scenario. A randomly selected question from this group of eligible questions is delivered to the simulated examinee. A correct or incorrect response is recorded, based on Sympson and Hetter's (1985) aforementioned probability model, and the examinee's theta is recalculated based on a correct or incorrect response and the parameters of the item.

After the first question on a given subtest is selected, the process for selecting all remaining items is slightly altered. A new content area, always different from the last content area used within the subtest, is then randomly selected. Items from the same content area will never be administered back-to-back as a means of assuring adequate content coverage and balance. The item from this content area that provides the most information at the examinee's current estimated quadrature node is identified, then, all items that provide at least 75% as much information as this item are also placed into the eligible item selection pool, regardless of the size of the minimum eligible item pool. If there are not yet enough items to minimally populate the pool, the standard of 75% information value is reduced in increments of 10% until an adequate amount of items are made available in the eligible item pool. The size of the minimum eligible item selection pool was manipulated in these simulations. Next, the item from this eligible pool with the minimum exposure frequency is identified. All items with 1.1x or fewer exposures than this minimally exposed item remain eligible (e.g., if the minimally exposed items had 50 exposures, all items with 55 or fewer exposures would still be eligible for delivery to the examinee). If no other items fit this criterion, the minimally-exposed item is delivered. If more than one item is still eligible, one is randomly selected to be delivered to the examinee. These steps repeat until the SEM threshold of 0.35 is reached, so long as at least 15 questions or more were answered. If the SEM threshold is not reached, a maximum number of 30 questions will be administered on the given subtest.

RESULTS

Table 2 highlights characteristics of the exposure frequencies and test length characteristics observed, based on 10,000 simulations per subtest scenario. As a general rule, increasing the size of the minimum eligible item pool by one item increased the average test length by approximately one half of an item.

Similarly, the number of minimum-length tests decreased and the number of maximum-length tests increased as the minimum eligible item pool expanded.

TABLE 2
ASTB ITEM EXPOSURE FREQUENCIES AND TEST LENGTH CHARACTERISTICS

	N	M	Med.	Max	N 0	M	SD	Med.	N 30	N 15
	Items	EF	EF	EF	EF Items	ID	ID	ID	IT	IT
Min. 3 IP										
MST	394	469.7	58	3862	115 (29.2%)	17.5	3.8	16	386	4488
RCT	220	1028.1	284	6545	25 (11.4%)	21.6	4.1	21	965	96
MCT	328	713.9	203	3926	56(17.1%)	22.4	5.1	22	1564	1030
ANI	255	770.0	164	5192	47(18.4%)	18.6	4.2	17	478	3098
Min. 4 IP										
MST	394	482.0	99	3513	96 (24.4%)	18.0	4.0	16	473	3761
RCT	220	1066.8	406	5796	14 (6.4%)	22.5	4.1	22	1184	49
MCT	328	742.0	333.5	3423	41 (12.5%)	23.3	5.0	23	2065	496
ANI	255	792.8	244	4564	39 (15.3%)	19.2	4.4	18	589	2478
Min. 5 IP										
MST	394	493.4	161.5	2899	85 (21.6%)	18.4	4.1	17	537	3091
RCT	220	1103.1	509.5	5484	6 (2.7%)	23.3	4.0	22	1519	13
MCT	328	770.7	399	3158	25 (7.6%)	24.3	4.7	24	2477	250
ANI	255	811.5	297	4268	29 (11.4%)	19.7	4.5	19	713	2068
Min. 6 IP										
MST	394	499.8	200	2755	77 (19.5%)	18.6	4.2	17	556	2723
RCT	220	1114.7	565	5152	5 (2.2%)	23.5	4.0	23	1613	16
MCT	328	783.8	455	2813	13 (4.0%)	24.7	4.6	25	2735	174
ANI	255	832.4	356	4195	23 (9.0%)	20.2	4.6	19	814	1675

Note: IP = Item Pool. EF= Exposure Frequency. ID = Items delivered per test. IT = Item tests. Med. = Median. Based on 10,000 simulations per subtest and scenario.

Median exposure frequencies more than doubled as the minimum eligible item pool expanded from three to six items, while mean exposure frequencies increased only slightly. This, along with the decrease in unexposed items, provides evidence that the exposure frequencies for the most-exposed items were reduced, and more underutilized items were being selected as the minimum eligible item pools expanded.

Results also indicated the overall quality of items within any given subtest item bank influenced the length of the tests. For example, on the MST which has the highest mean (0.99) ‘a’ or information parameter and a larger item bank overall, test length ranged from 17.51 items with a minimum eligible item pool or 3 to 18.61 items with a minimum eligible item pool of 6. The MCT, which has a mean ‘a’ parameter of 0.79, had test lengths which grew from 22.42 to 24.68 items as the minimum eligible item pool increased. This indicates that more robust subtests are more resistant to decrements in measurement efficiency as the minimum eligible item pool expands.

DISCUSSION, LIMITATIONS, AND FUTURE DIRECTION

The present study confirms Way’s (1998) assessment of the paradox of CATs: measurement efficiency will come at the cost of exposure control. By loosening the restrictions on exposure control, the most-used items were exposed less and more items were used overall, however the mean increase in test length by 1.10 items on the MST to 2.26 items on the MCT when comparing the three minimum item pool scenario to the six minimum item pool scenario. Still, even where the minimum eligible item selection pools were the largest, all mean subtests lengths are much shorter than the current static forms.

Previous CAT research has typically utilized item banks consisting of thousands of items, whereas the operational ASTB CATs only have several hundred items available per subtest. Accordingly, there have been more options available as to how to best limit item exposure (e.g., rotating sub-banks within each master test bank) for other researchers and practitioners. However, even with limited options, all CAT simulation scenarios presented in this research still yielded improvements over the current static forms. For example, if each of the three 30-item per subtest static forms were administered an equal number of times, we would see 3,333 exposures per item for every item. In the simulation featuring a minimum eligible item pool of five items, there were only 23 RCT and 9 ANI items that were exposed this many times or more. Proportionally, only a very small portion of the items would be administered to as many examinees as all items are on the static forms currently are. In addition, there were not any MST or MCT items for which such a high exposure frequency rate was observed. Even though a handful of items were administered to several thousand simulated examinees, many items had desirable exposure rates, where roughly one-third of the items within each of the given subtest item banks were exposed to between 1% and 10% of examinees.

Though not covered in the Results section of this research, the initial SEM calibration presented in the Methods suggests that practitioners should carefully evaluate theta recovery before deciding on SEM for stopping thresholds. If the results observed in this study generalize, there may be an optimal point at which adequate theta recovery can be obtained without sacrificing reduced test length, though not setting the SEM criteria for test termination at a low enough level can lead to unreliable tests with poor predictive value, and hence, yield poor selection decisions. Practitioners utilizing fixed-length CATs should also consider these implications, and may want to consider lengthening tests should theta recovery be low.

A potential limitation to this item selection algorithm is that random selection of item content areas does not assure uniform content representation across examinees, which may call into question test fairness. When parameter estimates were obtained for the items within each test's item bank, the items were administered on forms consisting of multiple items from all content areas. Any items not fitting the 3PL model, either in terms of low bi-serial correlation or poor model-data fit, for the subtest construct of interest during parameter estimation in BILOG-MG 3 (du Toit, 2003) were removed from further analyses and not included in the CAT item banks. Though items may come from various content areas, the tests as a whole are acceptably unidimensional (Stark, Chernyshenko, Choah, Lee, & Wadington, 2008). Given the relatively limited amount of content areas per subtest, never administering items from the same content area consecutively will ensure that in the majority of cases, examinee test should be comparable in terms of content representation. Though unlikely, the potential for extreme cases exists (e.g., alternating between only two content codes for the duration of an entire test). A potential alternative would be to dictate the order in which content areas would appear on a given test, but this would come at the expense of risking more overlap between any two given tests, particularly at the beginning of tests, when all examinees will have similar estimations of theta.

Several additional features have also been built into the capabilities of the CAT item selection algorithm to further limit items from becoming over-exposed and to enhance test security. For instance, items can be temporarily suspended, or even removed outright from the operational test bank. Several of the MST content areas feature upwards of 100 items, leading to many items that have relatively high information values across the ability continuum to never emerge in the eligible item pools. For several of these content areas, items with the highest exposure frequencies can be temporarily "turned off," which will have the effect of rotating under-exposed and non-exposed items into the eligible item selection pool.

Operationally, it may be beneficial to implement slightly different scenario specifications for each subtest. Based on the observed results, it may be most beneficial to utilize larger (e.g., five item) minimum item selection pools for all scenarios. This will only slightly increase the length of the average test for all subtests, but will make noticeable improvements on all tests in terms of number of items being utilized and keeping the highest exposure frequencies manageable. However, where item banks are large and contain more informative items (i.e., MST), it may be beneficial to regularly rotate out items from the largest content areas as a means of giving more items the potential to be utilized on the test, or perhaps even make the minimum eligible item pool even larger for content areas with many items. This likely would not increase the length of the average test in a noticeable way. However, where item banks are generally small and contain fewer informative items (i.e., RCT), regularly rotating out the most informative/most exposed items would likely lead to long test lengths, as any given content area may not contain highly informative items across ability levels. With nearly all test items already being utilized, there would be little gain from adopting this strategy.

Follow Up Study with Live Examinees

The CAT ASTB subtests were administered to 305 aviation students using a 5-item minimum eligible-item selection pool. Table 3 displays descriptive characteristics of exam lengths and exposure frequencies in comparison to those observed for the simulated sample previously described. Where appropriate, data from the 305 live cases have scaled to estimate what the expected values for 10,000 cases would be (e.g., exposure frequencies).

TABLE 3
COMPARISON BETWEEN LIVE EXAMINEE AND SIMULATED SAMPLES

	N	M	Med.	Max	N 0 EF	M	SD	Med.	N Max	N Min
	Items	EF	EF	EF	Items	ID	ID	ID	IT	IT
MST Sim.	394	447	230	1738	113 (28.7%)	16.6	3.5	16	98	1705
MST Live	394	493	162	2899	85 (21.6%)	18.4	4.1	17	537	3091
RCT Sim.	220	1103	510	5484	6 (2.7%)	23.3	4.0	22	1519	13
RCT Live	220	859	426	2656	35 (15.9%)	19.9	0.5	20	9443	0
MCT Sim.	328	744	607	1902	29 (8.8%)	24.6	4.4	25	2262	131
MCT Live	328	771	399	3158	25 (7.6%)	24.3	4.7	24	2477	250
ANIT Sim.	255	722	328	2262	45 (17.6%)	19.6	4.2	19	689	1705
ANIT Live	255	811	297	4268	29 (11.4%)	19.7	4.5	19	713	2068

Note: Note: EF= Exposure Frequencies. ID = Items delivered per test. IT = Item tests. Med. = Median. A 5-item minimum-eligible item pool was utilized for both the simulated and live sample. Where applicable, the values for the 305 live examinee sample have been adjusted to a sample of 10,000 examinees, assuming that the results from this sample extrapolate perfectly. These values were calculated for the mean exposure frequency, median exposure frequency, max exposure frequency, maximum item tests, and minimum item tests variables. Time expired on 136 of 305 MST before they had been completed. The mean number of questions answered would be higher otherwise. The MST time limit has since been expanded for operational use. The maximum number of RCT items delivered was set to 20 for the live sample, leading to fewer items delivered overall and a high observed frequency of maximum-length tests. All other subtests were set to a maximum of 30 items.

These findings hold promise for the viability of the CAT ASTB operationally. Some of highlights of this study include:

- Results closely mirrored those observed in the simulation studies. Mean items delivered on the MCT and ANIT were almost identical for the live and simulated examinees.
- The average item was presented to less than 9% of examinees on all 4 subtests.
- The most frequently exposed items were presented to a smaller proportion of live examinees than simulated examinees (e.g., the most-exposed ANIT item was presented to 22.6% of live examinees and 42.3% of simulated examinees).
- Considering the relatively small size of the live examinee sample, the majority of items were exposed at least once.

CONCLUSION

Previous CAT research has focused on simulated and operational CAT tests featuring extensive item libraries. In practice, many organizations that regularly test could benefit from implementing CAT tests, but do not have the resources available or a true need to develop an item library consisting of thousands of items per subtest. The method examined herein, where item exposure is limited for the most-exposed items by presenting examinees with the least-exposed item that still yields high levels of information at their current estimated ability level, sacrifices little in the way of efficiency for reaching a desirable level of measurement accuracy. There are small trade-offs in efficiency between obtaining desired measurement accuracy and overexposing the most-exposed items. Practitioners are advised to consider the characteristics of their item libraries when deciding on the size of a minimum eligible item pool before finalizing the specifications of CAT algorithms.

Results yielded from all four simulated CAT scenarios of ASTB subtests would greatly reduce item exposure while simultaneously maintaining the statistical precision necessary for high-stakes selection testing when compared to the current static versions of the test. Additionally, with the majority of examinees answering fewer questions than they would on a static test, we expect to see a reduction in average exam administration time, saving the money and time associated with proctoring the ASTB. An initial study on live examinees yielded results comparable to those found in the simulations, holding great promise for the operational implementation of the CAT ASTB subtests.

REFERENCES

Barrada, J. R., Veldkamp, B. P., & Olea, J. (2008). Multiple Maximum Exposure Rates in Computerized Adaptive Testing. *Applied Psychological Measurement 2009, 33*, 58-73.

du Toit, M. (Ed.). (2003). *IRT from SSI*. Lincolnwood, IL: Scientific Software International, Inc.

Li, X., Becker, K., Gorham, J., & Woo, A. (June, 2009). *Limiting Item Exposure for Target Difficulty Ranges in a High-Stakes CAT*. Paper presented at the GMAC Conference on Computerized Adaptive Testing, Minneapolis, MN.

Moclaire, C., Middleton, E., and Phillips, H. (2011). *The Aviation Selection Test Battery: Descriptive Statistics and Predictive Relationships for Fiscal Years 2005-2010*. Project technical report. Pensacola, FL: Naval Aerospace Medical Institute.

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 Annual Meeting of the National Council of Measurement in Education. San Diego, CA.

Phillips, H. L. (2004). *ASTB Forms 3, 4, and 5 Development*. Project technical report. Pensacola, FL: Naval Aerospace Medical Institute.

Revuelta, J., & Ponsoda, V. (1998). A Comparison of Item Exposure Control Methods in Computerized Adaptive Testing. *Journal of Educational Measurement, 35,* 311-327.

Stark, S., Chernyshenko, S., Chuah, D., Lee, W., & Wadlington, P. (2008). University of Illinois Item Response Theory Laboratory, University of Illinois, Champagne-Urbana. http://io.psych.uiuc.edu/irt/default.asp.

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.

Van der Linden, W.J., & Veldkamp, B. P. (2004). Constraining Item Exposure in Computerized Adaptive Testing with Shadow Tests. *Journal of educational and Behavioral Statistics, 29,* 273-291.

Van der Linden, W. J., & Veldkamp, B. P. (2007). Conditional Item-Exposure Control in Adaptive Testing Using Item-Ineligibility Probabilities. *Journal of Educational and Behavioral Statistics, 32*, 398-418.

Way, W.D. (1998). Protecting the Integrity of Computerized Testing Item Pools. *Educational Measurement: Issues and Practice, 17*, 17-27.