Graduate Management Admission Council®

A Gradual Maximum Information Ratio Approach to Item Selection in Computerized Adaptive Testing

Kyung T. Han

Graduate Management Admission Council®

Presented at the Item Selection Paper Session, June 2, 2009

Abstract

For long-term quality control of computerized adaptive test (CAT) programs, optimizing the usage of the item bank (i.e., controlling item exposure rate) is critical. Selecting the best item often conflicts with the procedure for item exposure control, however. In this study, a newly proposed approach to item selection in CAT, in which the efficiency of items is considered in the early stages of CAT administration, was compared with the partial randomization method and the traditional maximized Fisher information (MFI) method. The simulation study found that the new approach greatly improved item bank utilization, compared with the other methods, while minimizing the compromise of test precision. The findings of this study could help practitioners find the best possible balance between test information and item bank utilization.

Acknowledgment

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

Citation

Han, K. T. (2009). A gradual maximum information ratio approach to item selection in computerized adaptive testing. 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

Kyung (Chris) T. Han, Graduate Management Admission Council®, 1600 Tysons Blvd., Suite #1400, McLean, VA 22102, U.S.A. Email: khan@gmac.com

A Gradual Maximum Information Ratio Approach to Item Selection in Computerized Adaptive Testing

With the emergence of modern test theory, such as item response theory (IRT), and the rapid advancement of computer technology in the last few decades, computerized adaptive testing (CAT) has entered the mainstream of educational measurement. The most distinctive advantage of CAT is that a test can be altered to best fit each examinee's ability level (e.g., the test difficulty is matched to the examinee's expected ability). As a result, test accuracy and reliability can be substantially improved, while test length and time remain the same or can be reduced compared to paper-and-pencil-based tests (PBT). To achieve full advantage of CAT's capabilities, it is critical to have an item selection algorithm that maximizes test information for each examinee, while satisfying the other requirements such as content balancing. For long-term quality control of CAT programs, optimizing the usage of the item bank (i.e., controlling item exposure rate) can also be very important. Selecting the best item often conflicts with the procedure for item exposure control, however. Thus, the key to successful CAT implementation is finding the best possible balance between test information and item exposure.

One of the most widely used—and probably the oldest—item selection methods, in CAT involves selecting an item with maximized Fisher information (MFI) at the interim □ estimate based on test items previously administered to the examinee (i.e., finding item x maximizing $I_x[\hat{\theta}_{m-1}]$ for an examinee with the interim □ estimate $\hat{\theta}$ and m-1 as the number of items administered so far (Weiss, 1982). For example, with a typical case of a multiple-choice item bank, where item characteristics are defined by the three-parameter logistic model (3PLM; Birnbaum, 1968), the MFI method looks for item x that results in the largest value of

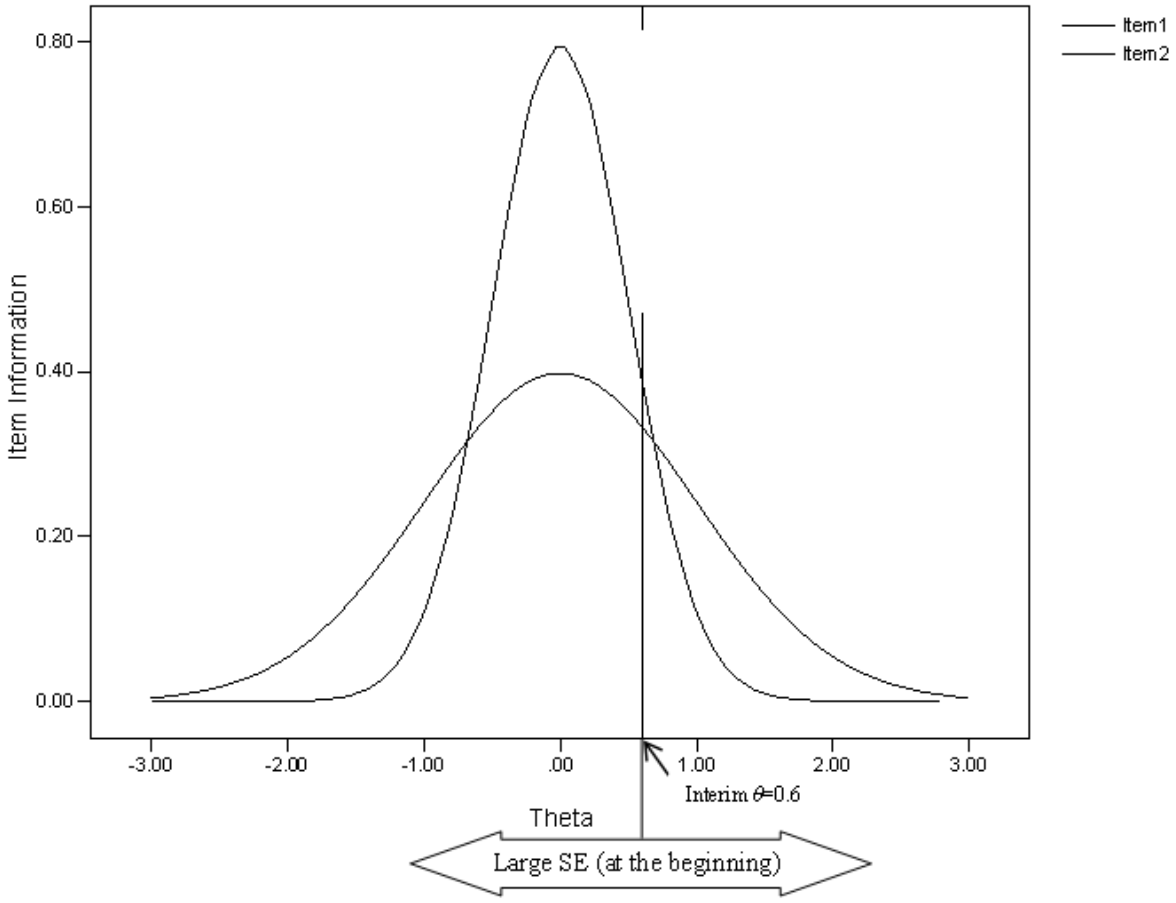
$$I_x[\hat{\theta}_{m-1}] = \frac{(Da_x)^2(1-c_x)}{\left\{c_x + \exp\left[Da_x\left(\hat{\theta}_{m-1} - b_x\right)\right]\right\}\left\{c_x + \exp\left[-Da_x\left(\hat{\theta}_{m-1} - b_x\right)\right]\right\}^2}. \quad (1)$$

where a_x, b_x, and c_x are the discrimination, difficulty, and pseudo-guessing parameters of the 3PLM, respectively, and D is a scaling constant equal to 1.702. The MFI approach has been very popular because it is a simple, straightforward, and effective means of administering CAT that results in maximized test information for each individual; however, it has two significant drawbacks:

1. The MFI approach itself is not capable of controlling item exposure rate, and as a result, a portion of the items in the item bank might be used excessively while the rest of items might be used rarely. This problem can be easily solved by incorporating one of the various item exposure control strategies (Georgiadou, Triantafillow, & Economides, 2007) such as randomization (McBride & Martin, 1983; Kingsbury & Zara, 1989; Revuelta & Ponsoda, 1998), conditional selection (Sympson & Hetter, 1985; Stocking & Lewis, 1995, van der Linden & Veldkamp, 2005), and multiple-stage testing (Luecht, 2003).

2. The more challenging problem to solve is that the interim θ estimates at the beginning of a test (e.g., before at least five items are administered) are rarely accurate, so applying the MFI method at the start of testing might not be very efficient, and might cause excessive exposure of those items with greater information. For example, as shown in Figure 1, if one of two eligible items needed to be selected with an interim of $\hat{\theta} = 0.6$, Item 2 would always be preferred over Item 1 with the MFI method. If that item selection happened with an examinee in the early stage of CAT administration, however (within the first five items administered, for example), the standard error (SE) for the interim $\hat{\theta}$ of 0.6 would be very large (often between 1.0 and 4.0 before the fifth item administration). As a result, the actual information gained from Item 2 at the true θ could be far less than what was expected when $\hat{\theta} = 0.6$. In fact, Item 1 might have a better chance to provide more information if the SE of the interim $\hat{\theta}$ were larger than 1.5.

Figure 1. Example of CAT Item Selection at the Earlier Stages of Testing



To avoid such a wasteful selection of those items with large *a* parameter values in the early stage of CAT administration, Chang and Ying (1999) suggested stratifying the items in the bank

by *a* values and using those item strata with lower *a* values in the early stage of CAT. The *a*-stratified strategy is a practical and effective means of controlling item exposure rate; however, this strategy yields overall test information for individuals that tends to be somewhat lower than MFI item selection. This approach also can be problematic when the *a* and *b* parameters are correlated. Subsequently, Chang and van der Linden (2003) proposed using the 0-1 linear programming optimization method to stratify item banks to overcome the situation where there was a correlational relationship between *a* and *b* parameters. This method clearly improved the item exposure control even when the *a* and *b* parameters were correlated. Several problems inherited from the stratification of an item bank still persist, however. For example, determining the number of item strata could be ambiguous, and stratification can cause or increase the chance of facing infeasible solutions when there are a number of nonstatistical constraints and the total number of items is too small.

A Gradual Maximum Information Ratio Approach

Fisher information (Equation 1) can be seen as a measure of the effectiveness of an item at a certain point on the θ scale. The efficiency of an item can be evaluated by the ratio of Fisher information at a certain θ value to the maximum information across the θ scale. Thus, the efficiency of an item can be expressed as

$$\frac{I_x[\hat{\theta}_{m-1}]}{I_x[\theta^*]}, \tag{2}$$

where θ^* is a certain θ point where the information is maximized. When the *c* parameter is equal to zero (i.e., when the one- or two-parameter model is used), θ^* is equal to b_x. If $c_x \neq 0$, θ^* can be easily computed using Birnbaum's (1969) solution:

$$\theta_x^* = b_x + \frac{1}{Da_x}\log\left(\frac{1+\sqrt{1+8c_x}}{2}\right). \tag{3}$$

This paper proposes a new approach, in which the ratio of expected information with an interim $\hat{\theta}$ to the potential maximum information (i.e., the item efficiency) is used as an item selection criterion in the earlier stages of CAT administration. As the CAT administration progresses toward the end and the SE of the interim $\hat{\theta}$ gets smaller, however, the new approach considers item effectiveness (i.e., MFI) as the more important criterion. The new approach, hereafter referred to as the gradual maximum information ratio (GMIR) approach, looks for an item that maximizes

$$\frac{I_x\left[\hat{\theta}_{m-1}\right]}{I_x\left[\theta^*\right]}\left(1-\frac{m}{M}\right)+I_x\left[\hat{\theta}_{m-1}\right]\frac{m}{M}, \tag{4}$$

where *M* is the test length, and *m* is 1 plus the number of items administered thus far. The first part of Equation 4 is the item efficiency term (Equation 2), and second part is the Fisher information term (Equation 1). Each part of Equation 4 is inversely weighted by the progress of the CAT administration. Equation 4 can be factored by the Fisher information term, as follows:

$$I_x\left[\hat{\theta}_{m-1}\right]\frac{M-m\left(1-I_x\left[\theta^*\right]\right)}{I_x\left[\theta^*\right]M}. \quad (5)$$

Simulation Study

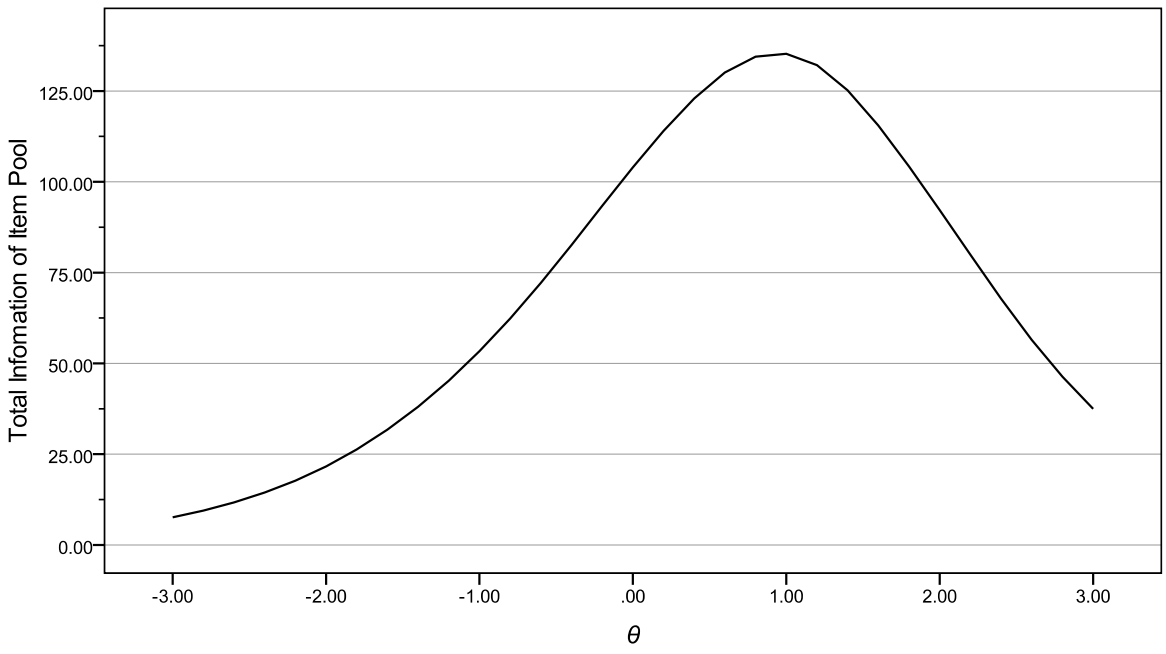
A series of simulation studies were conducted to evaluate the effectiveness of the GMIR approach. The simulation studies mimicked one month (20 administration days) of an existing CAT program for higher education (with simplified content balancing) and used the evaluation criteria of item exposure rate, test information, item bank usage, and θ estimation bias and errors.

Data

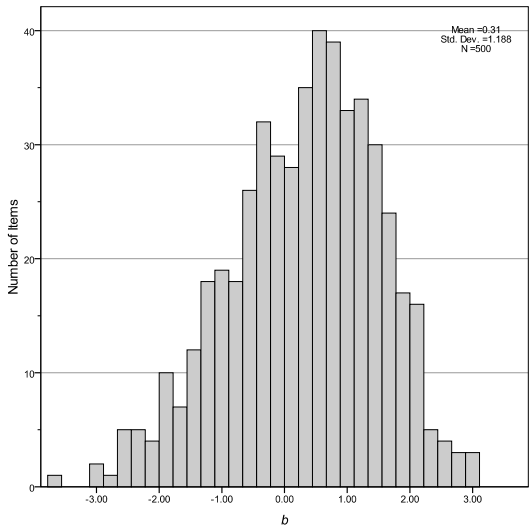
To construct the item bank, 500 multiple-choice math items were drawn from the GMAT® item bank (the size of the item bank in this study—n = 500—was not the actual size of the operational GMAT® item bank). The aggregated total information of the item bank showed the peak around $\theta = 1$ (Figure 1a), not only because there was a large number of difficult items (Figure 1b) but also because the difficult items tended to be more discriminating (Figure 1c). To simplify the study and to increase the generalizability of the results, constraints on content balancing were not applied.

Figure 2. Item Bank Characteristics

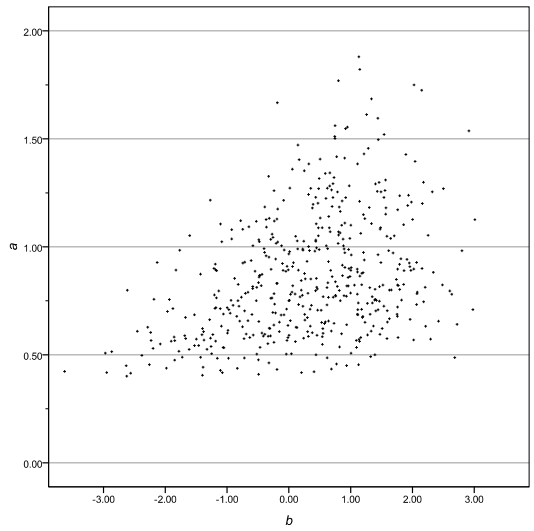
a. Bank Information Function



b. Item Difficulty Distribution



b. Correlational Relationship Between *a* and *b* Parameters



One month of the CAT administration was simulated with 10,000 examinees that were drawn from the standard normal distribution [~N(0,1)]. Each examinee was administered 40 items. Five hundred examinees were administered each day for the 20 days, and each day had two time slots. Thus, the 250 examinees were simultaneously administered CAT for each testing time slot, and the item usage information in the item bank server was updated after each time slot.

Item Selection Methods

Five different item selection methods were implemented and compared. In the first method, the item selection algorithm looked for five items that included *b* parameter values closest to the interim $\hat{\theta}$, and randomly selected one of those five items. This method can be seen as a combination of the randomesque strategy (Kingsbury & Zara, 1989) and the simplified version of the *a*-stratified strategy (Chang & Ying, 1999) where there was only one item stratum. The item exposure rate was constrained to be less than 0.20, and those items exceeding the constraint were temporarily kept from the selection. The item exposure rate was computed based on the latest item usage information in the item bank server.

The second method was the typical MFI approach (Equation 1). The item exposure rate was constrained to be less than 0.20 as in the first method.

The third method also used the MFI approach, but the item selection algorithm integrated a different item exposure control mechanism and looked for an item that maximized

$$I_x\left[\hat{\theta}_{m-1}\right]\frac{C_x-\left(U_x/N_x\right)}{C_x}, \tag{6}$$

where C_x was the constraint of the item exposure rate 0.20 in this study, U_x was the item usage for the life of item x, and N_x was the number of CAT administrations while item x was in the item bank. With this method, those rarely used items were expected to be promoted more strongly, whereas those excessively used items were likely to "fade away" from the item selection (this method will be referred to hereafter as the fade-away method).

The fourth method was the GMIR approach (Equation 5) with the exposure rate constraint of 0.20 as in the first and second methods.

Finally, the fifth method involved the GMIR approach, which uses the fade-away item exposure control method utilized in the third method. Thus, the fifth method looked for an item that maximized

$$\frac{\left(U_x/N_x\right)}{C_x}I_x\left[\hat{\theta}_{m-1}\right]\frac{M-m\left(1-I_x\left[\theta^*\right]\right)}{I_x\left[\theta^*\right]M}. \tag{7}$$

Procedure

A modified version of the computer software *WinGen* (Han, 2007) was used to simulate the CAT administration using the five item selection methods. The first item for each examinee was randomly chosen among those items whose *b*-parameter value was between –0.5 and 0.5, and the interim $\hat{\theta}$ was estimated using the *expected a posteriori* (EAP) method after each item administration. The θ estimation algorithm limited the change of the interim $\hat{\theta}$ from the previous estimate to ± 0.5.

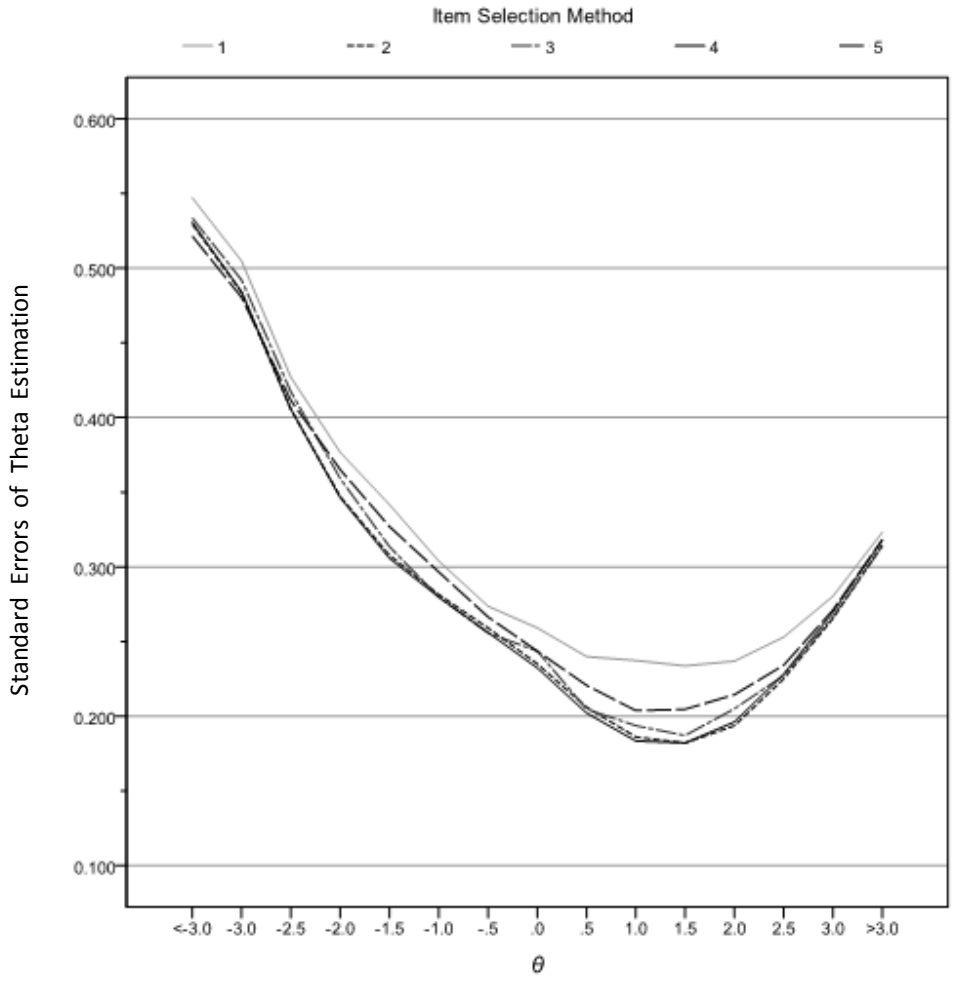
In the simulation, each client terminal was assumed to communicate with the item bank server only before and after each individual's test administration. Therefore, the item exposure control was based on the item usage information that was updated up to the previous time slot.

The evaluation of the five item selection methods focused on two major points: (1) performance of θ estimation, and (2) item bank usage. First, θ estimation was evaluated by the standard errors of the θ estimates (SEE) across the θ scale. The bias and mean absolute error of the θ estimates were also computed. To see if the quality of the CAT administration held during the whole month, the change in SEE across administration days was investigated as well. Second, the item exposure rate was analyzed at the item level to determine which item selection method resulted in the most optimal item bank usage. The entire procedure was replicated 100 times and median values were reported.

Results

The SEEs are plotted in Figure 3. The item selection in Methods 2 and 4 showed the smallest SEE across the θ scale, whereas Method 1 resulted in the largest SEE. This was the expected result: the MFI and GMIR approaches select items to maximize test information either during the whole CAT administration (MFI) or during the later part of CAT administration (GMIR). When the MFI or GMIR approaches teamed up with the fade-away method to control item exposure more rigorously (Methods 3 and 5), the SEE was slightly increased across the θ scale.

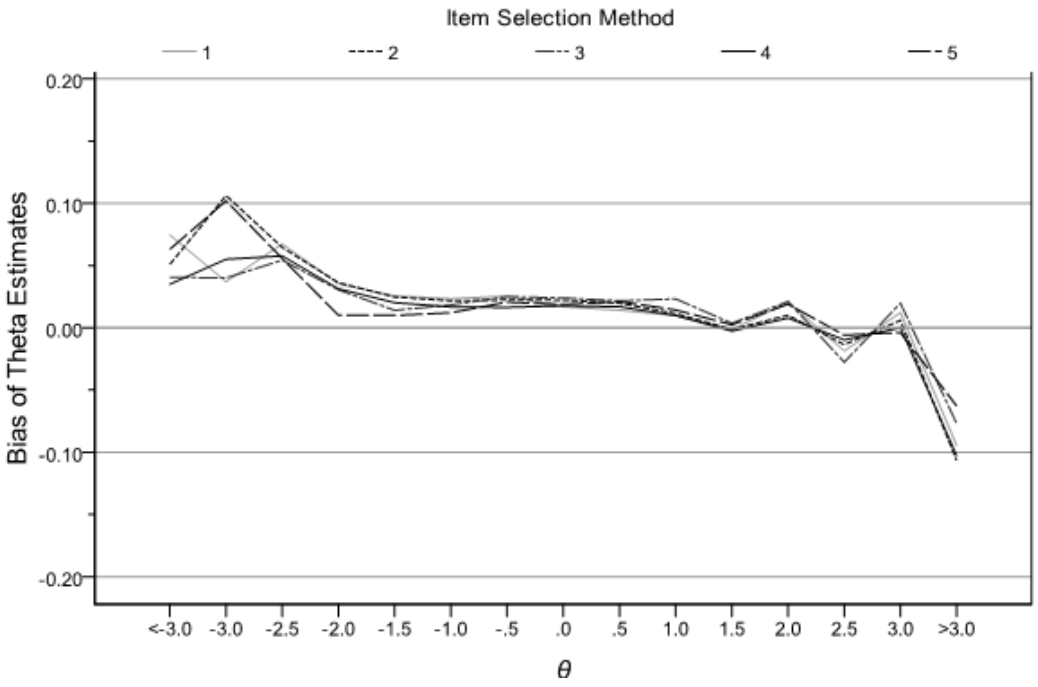
Figure 3. Standard Error of θ Estimates



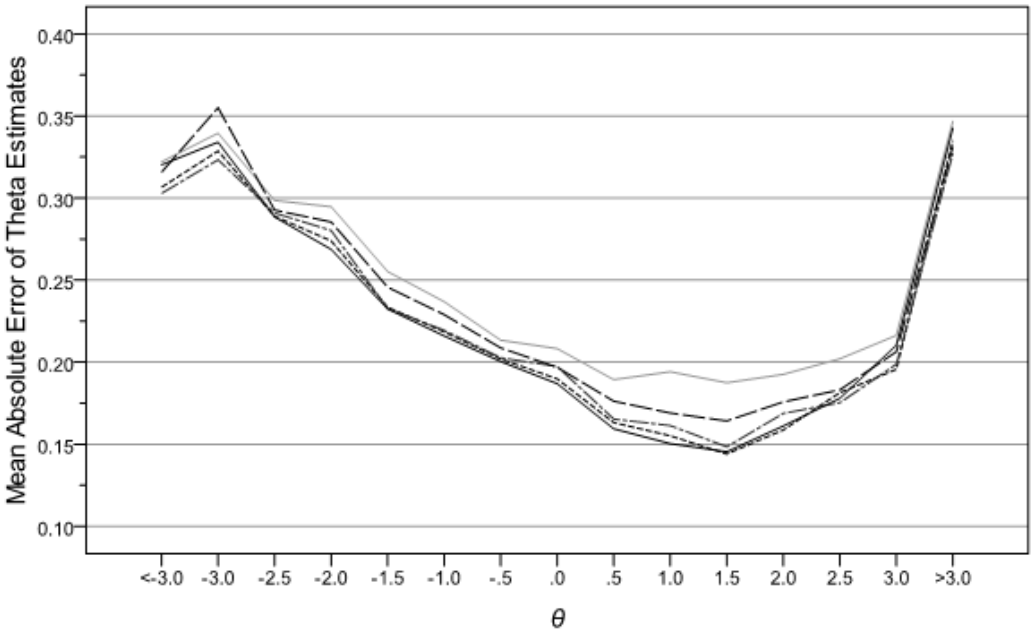
In terms of the estimation bias, there was no meaningful difference among the item selection methods (Figure 4a). For the majority of the θ area (–2.0 to 2.0), small positive bias was observed, but the magnitude of the bias was minimal (about 0.025 on average). As an empirical measure of the θ estimation errors, the mean absolute errors (MAE) are reported in Figure 4b. Overall, the patterns of the MAE were almost identical to the SEE in Figure 3.

Figure 4. Bias and Mean Absolute Error of θ Estimates

a. Bias



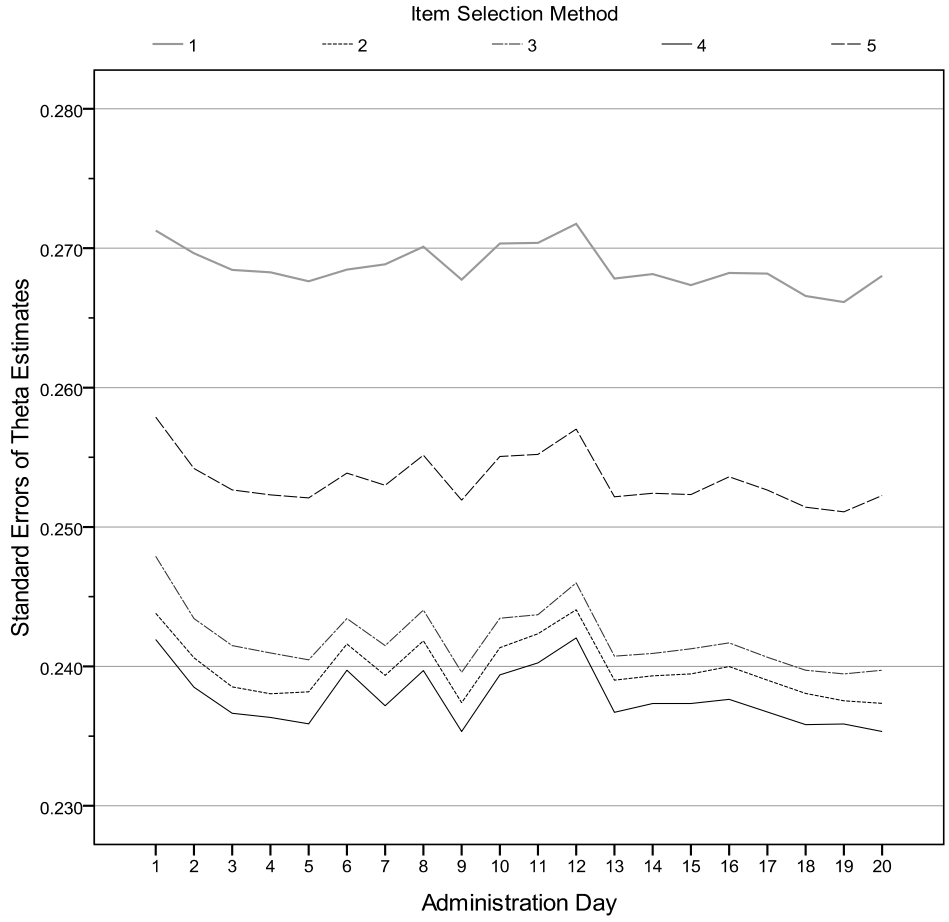
c. Mean Absolute Error



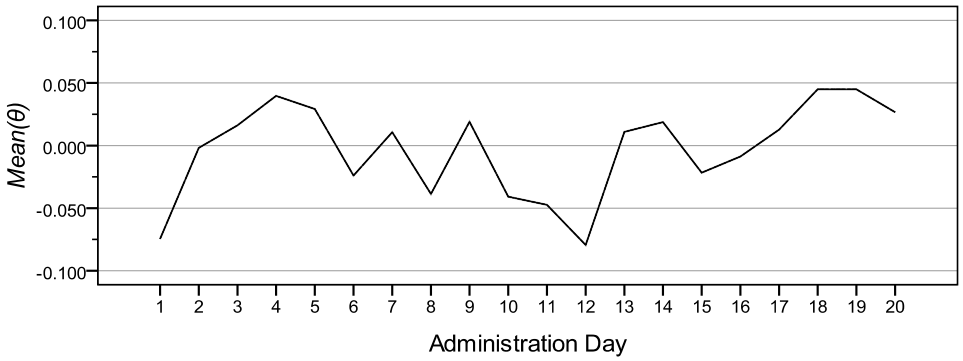
To examine whether the quality of the CAT administration was stably maintained over time with the various combinations of the item selection and exposure control methods, the mean SEE for each administration day was analyzed, as shown in Figure 5. There were visible fluctuations in the mean SEE over time (Figure 5a), but these were mainly due to the change in the examinee distribution day by day (Figure 5b). Within the 20-day period, each item selection method seemed to succeed in maintaining the quality of the CAT implementation. In fact, Figure 5 also clearly indicated the difference in the mean SEE among the item selection methods. Method 4 (GMIR plus item exposure constraint) resulted in the smallest SEE over time, and Methods 2 and 3 (MFI plus item exposure constraint and MFI plus fade-away method) closely followed. With Method 5 (GMIR plus fade-away method), the SEE was slightly increased, and Method 1 (modified randomesque plus item exposure constraint) resulted in substantially increased SEEs. As shown in Figure 5, the impact of a choice of item selection persisted over time.

Figure 5. Mean Standard Errors of θ Estimation and the True Mean θ for Each Administration Day

a. Mean Standard Error



b. Mean True θ



The study also evaluated the effectiveness of the item selection methods in terms of item bank utilization. In Figure 6, the item exposure rates of the all 500 items in the item bank are plotted for each item selection method. With Method 1, no items were excessively used up to the item exposure constraint (0.20), and the item exposure rates were relatively evenly distributed. On the other hand, Method 2 resulted in extremely unbalanced item usage. A large group of items were used up to the maximum exposure rate, and another large group of items were not used at all. When the fade-away exposure control method was used in Method 3, no items were used up to the maximum constraint. Several items that were not used at all were still found frequently with Method 3, however. Method 4 resulted in item usage similar to Method 2: a large number of items were not used at all while many other items were used up to the maximum exposure rate. Method 5 had no items that were either used up to the exposure limit or not used at all. More importantly, the item bank usage was very well balanced in Method 5. Figure 6 shows that the exposure rate of a majority of the items in the bank clustered around 0.10, with few items exposed over 0.15.

In Figure 7, the items in the bank were categorized by the usage (item exposure rate divided by the maximum item exposure constraint, which was 0.20). By summarizing the number of items in each category, Figure 7 shows how well each item selection method utilized the item bank. With Methods 2 and 4, there were approximately 150 items that were not used at all during the 20 administration days, representing about 30 percent of the item bank. On the other hand, those two methods caused excess usage on more than 125 items, or about 25 percent of the item bank. Such extremely unbalanced item bank usage can be a serious problem in maintaining the item bank over the long term. With Methods 1 and 3, there were fewer extreme cases of unbalanced item bank usage; however item usage still varied considerably. Method 5, as illustrated in Figures 6 and 7, resulted in the most balanced item bank usage. Nearly 240 items (or about 48 percent of the item bank) were exposed between 40 percent and 60 percent of the exposure limit.

Figure 6. Item Exposure Rate of the Individual Items in the Bank With Each Item Selection Method

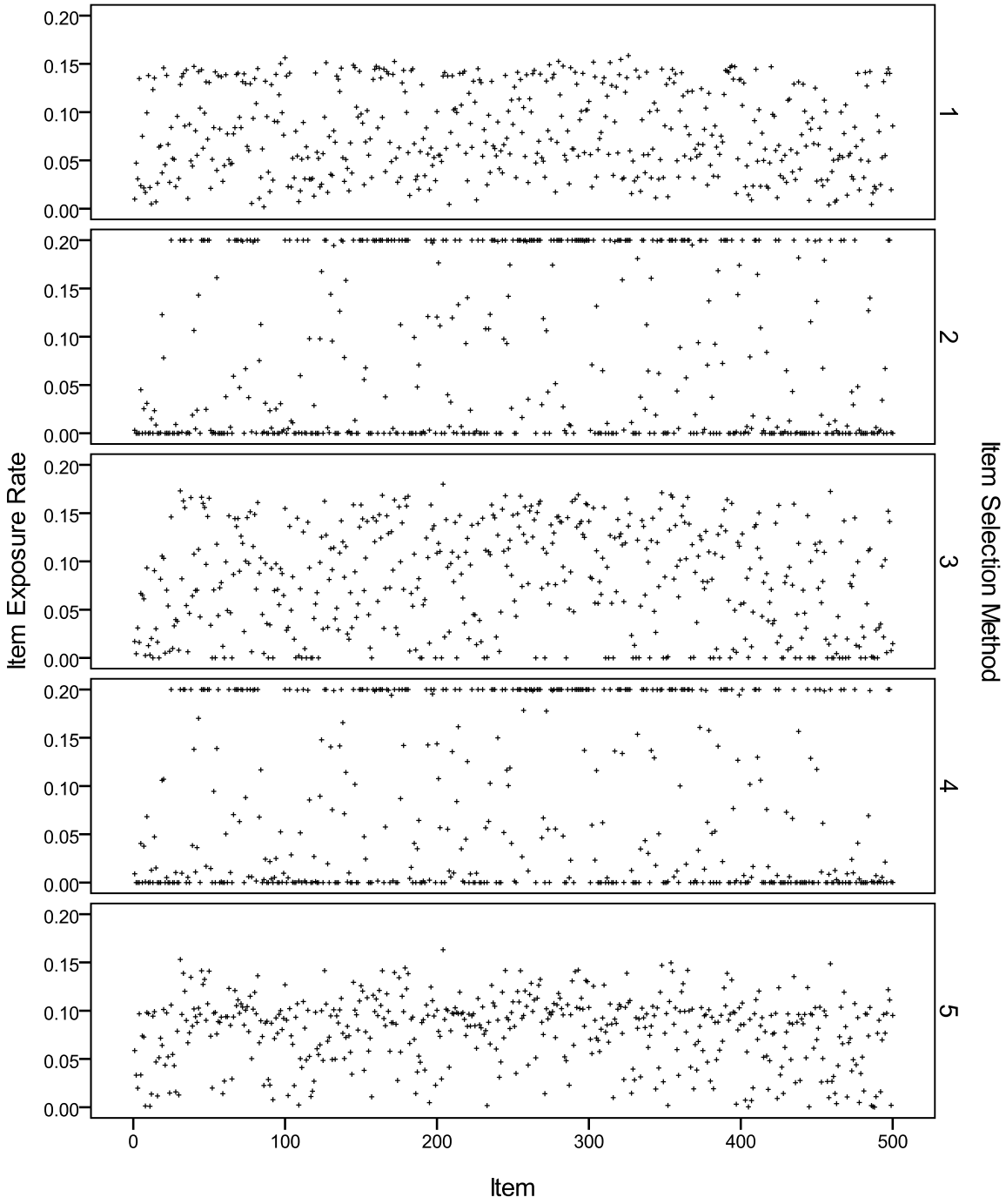
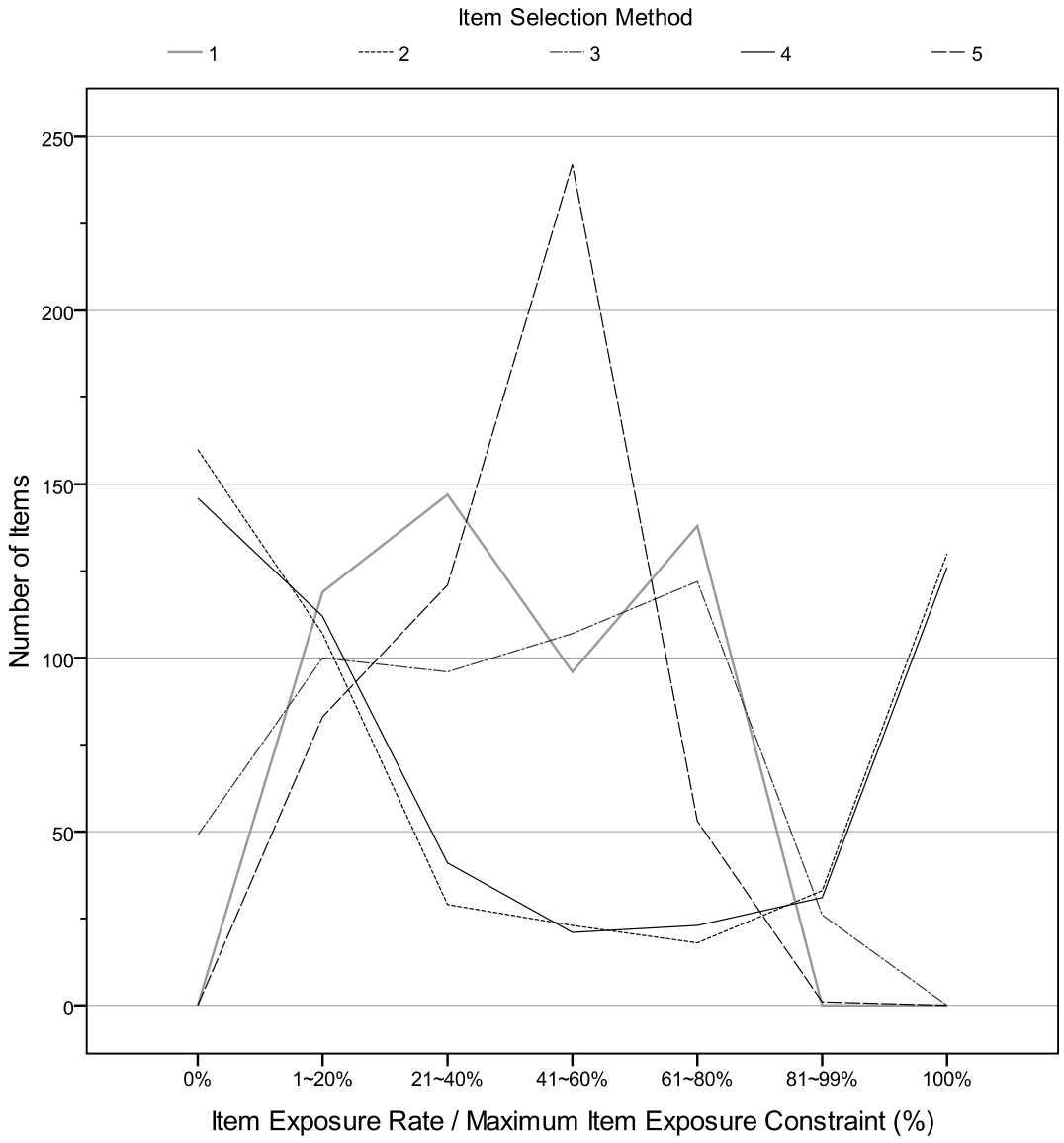


Figure 7. Item Pool Usage with Each Item Selection Method



Discussion

Developing an item selection algorithm that includes item exposure control might not necessarily be a process of establishing an ideal theory. Rather, it could be viewed as a process of searching for the most empirically effective mechanism. In theory, the MFI-based methods should result in maximized information (in other words, minimized SEE). The simulation study showed, however, that the new GMIR approach (Method 4), in which item efficiency was considered in the early stage of CAT, resulted in slightly less SEE than the MFI-based methods (Methods 2 and 3). It is possible that the GMIR strategy of selecting the most efficient item was more robust against the instability of the interim $\hat{\theta}$ in the early stages of CAT compared with the MFI strategy of selecting the most effective item. Because the difference in SEE between the

MFI and GMIR methods was not meaningfully substantial, the MFI method could be still seen as one of the most effective methods resulting in maximized test information.

When it comes to the effectiveness of utilizing the item bank, however; the GMIR approach with the fade-away item exposure control (Method 5) substantially outperformed the other studied methods. Because the process of constructing and managing parallel item banks is usually very complicated, testing programs try to maintain item banks without significant changes over time. For example item bank rotation is one strategy that many testing programs are employing to stretch out the lifespan of item banks. If the item bank usage is unbalanced, as seen in this study with Methods 2 and 4, the usage of each item bank is likely to vary significantly from one item bank to another. If the item bank usage is not parallel among the item banks, the properties of the item banks might fluctuate across the item banks as well, in which case the quality control of the testing program could face serious problems over time, especially when item banks are rotated.

Another potential problem with unbalanced item bank usage involves test security. With Methods 2 and 4, approximately 150 out of the 500 items were never used, which means that the actual size of the item bank used in CAT administration was only about 350 items, and not 500. Such a decrease in the item bank size increases the chance that more examinees will receive the same items. Although the maximum item exposure rate is usually limited by the constraints, simply keeping the items under the item exposure constraints does not necessarily guarantee freedom from test security problems. Smaller item exposure rates would lead to reduced chances of test security issues due to item exposure. In the simulation study, nearly 90 percent of the item bank in Method 5 had an exposure rate less than 0.12 (or 60 percent of the maximum item usage). In addition, no items were exposed more than 0.16 (or 80 percent of the maximum item usage) with Method 5 (Figure 7). Therefore, compared with the other studied methods, Method 5 was clearly much more effective and efficient in utilizing the item bank.

The trade-off between maximizing test information and reducing the item exposure rate is often considered unavoidable. Indeed, there was a slight increase in SEE (in other words, a slight decrease in the test information) with Method 5 compared with Methods 2, 3, and 4 (Figure 3). Considering the improvement in item bank utilization with Method 5, however, such a small decrease in the test information with Method 5 would not be a meaningful drawback. In fact, Method 5 showed improvements in both test information and item bank utilization over Method 1 (the partial randomization method).

This study tested the GMIR approach using two different item exposure control methods. With the simple exposure constraint (Method 4), the GMIR approach showed similar results to the MFI method (Method 2) in item bank usage. When the GMIR approach was used with the fade-away item exposure control method (Method 5), item bank utilization was improved by far over the other combinations of item selection and item exposure control methods. Thus, it is very important to continue investigating how well the GMIR approach performs with other item exposure control techniques. It is also suggested that future studies examine what happens with the GMIR approach when there are a number of content constraints.

Conclusions

Producing an accurate measure of what is to be measured would be the goal of any kind of testing, and the accuracy of test measurements is mainly determined by the test information at each examinee's level. CAT has been considered as the ultimate solution for realizing the most accurate assessment and maximizing test information for each individual, and the MFI approach has been the most popular item selection criterion since CAT joined the mainstream of the measurement field.

In this study, the newly proposed GMIR approach, in which the efficiency of items is considered in the early stages of CAT administration, was compared with the partial randomization method and the MFI method. The simulation study found that the GMIR approach greatly improved item bank utilization compared with the MFI method while minimizing the compromise of test precision.

References

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord and M. R. Novick (Eds.), *Statistical theories of mental test scores* (chapters 17-20). Reading, MA: Addison-Wesley.

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

Chang, H.-H., & van der Linden, W. J. (2003). Optimal stratification of item pools in alpha-stratified computerized adaptive testing. *Applied Psychological Measurement, 27*, 262-274.

Georgiadou, E., Triantafillow, 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)*. Retrieved from http://www.jtla.org.

Han, K. T. (2007). WinGen: Windows software that generates IRT parameters and item responses. *Applied Psychological Measurement, 31,* 457-459.

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

Luecht, R. M. (April 2003). *Exposure control using adaptive multi-stage item bundles.* Paper presented at the annual meeting of the National Council on Measurement in Education (NCME), Chicago, IL.

McBride, J. R., & Martin, J. T. (1983). Reliability and validity of adaptive ability tests in a military setting. In D. J. Weiss (Ed.), *New horizons in testing: Latent trait test theory and computerized adaptive testing* (pp. 224-236). New York: Academic Press.

Revuelta, J., & Ponsoda, V. (1998). A comparison of item exposure control methods in computerized adaptive testing. *Journal of Educational Measurement, 35,* 311-327.

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. In Proceedings of the 27th annual meeting of the Military Association, (pp. 973-977). San Diego, CA: Navy Personnel Research and Development Center.

van der Linden, W. J. (2000). Constrained adaptive testing with shadow tests. In W. J. van der Linden & C. A. W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 27-52). Norwell, MA: Kluwer.

van der Linden, W. J., & Veldkamp, B. P. (December 2005). *Constraining item exposure in computerized adaptive testing with shadow tests.* Law School Admission Council Computerized Testing Report 02-03.

Weiss, D. J. (1982). Improving measurement quality and efficiency with adaptive testing. *Applied Psychological Measurement, 6*, 473-492.