GMAC®

Graduate
Management
Admission
Council®

Creating Access to Graduate Business Education®

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

Kyung (Chris) T. Han

GMAC® Research Reports • RR-09-07 • June 25, 2009

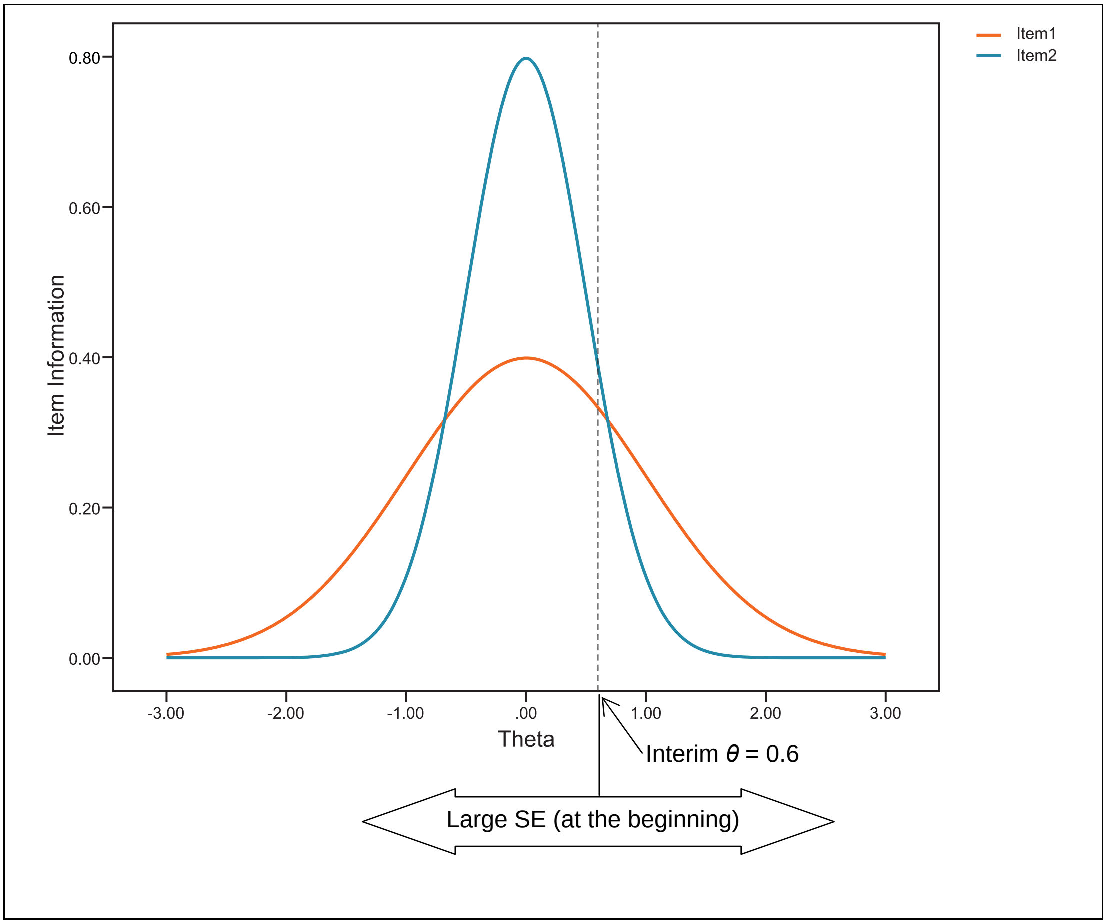
With the emergence of modern test theory, such as item response theory (IRT) and the rapid advancement of computer technology in the last few decades, computerized adaptive testing (CAT) has entered the mainstream of educational measurement. The most distinctive advantage of CAT is that a test can be altered to best fit each test taker's ability level (e.g., the test difficulty is matched to the test taker's expected proficiency). As a result, the test accuracy and reliability can be improved substantially, while test length and time remain the same or can be reduced compared to the 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 test taker, while satisfying the other requirements such as content balancing. For long-term quality control of CAT programs, optimizing the usage of the item pool (i.e., controlling item exposure rate) is also 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 the maximized Fisher information (MFI) at the interim proficiency 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 proficiency estimate $\hat{\theta}$ and m-1 as the number of item administered so far [Weiss, 1982]). For example, with a typical case of a multiple-choice item pool, 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

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

where a_x, b_x, and c_x are the discrimination, difficulty, and pseudo-guessing parameters in 3PLM, respectively, and D is a scaling constant whose value is 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. First, the MFI approach itself is not capable of controlling item exposure rate, and as a result, a portion of the items in the item pool may be used excessively while the rest of items may 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). The second major problem of the MFI approach, and the more challenging to solve, is that the interim proficiency 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 may not be very efficient, and may 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 a test taker 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 one and four 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 Item Selection at Earlier Stage 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 pool 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. This approach also can be problematic when the *a*- and *b*-parameters are correlated. Subsequently, Chang and van der Linden (2003) proposed to use the 0-1 linear programming optimization method to stratify item pools to overcome the situation where there was a correlation between *a*- and *b*- parameters. This method clearly improved the item exposure control even when the *a*- and *b*- parameters are correlated. Several problems inherited from the stratification of item pool 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.

Gradual Maximum Information Ratio Approach

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

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

where θ^* is a certain theta point where the information is maximized. When the *c*-parameter is equal to zero (i.e., when 1PLM or 2PLM is used), θ^* is equal to b_x. If $c_x \neq 0$, θ^* can be easily computed using Birnbaum's solution (1968):

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

This study 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 stage of CAT administration. As the CAT administration progresses toward the end and the SE of interim $\hat{\theta}$ gets smaller, however, the new approach considers the 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

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

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:

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

Simulation Study

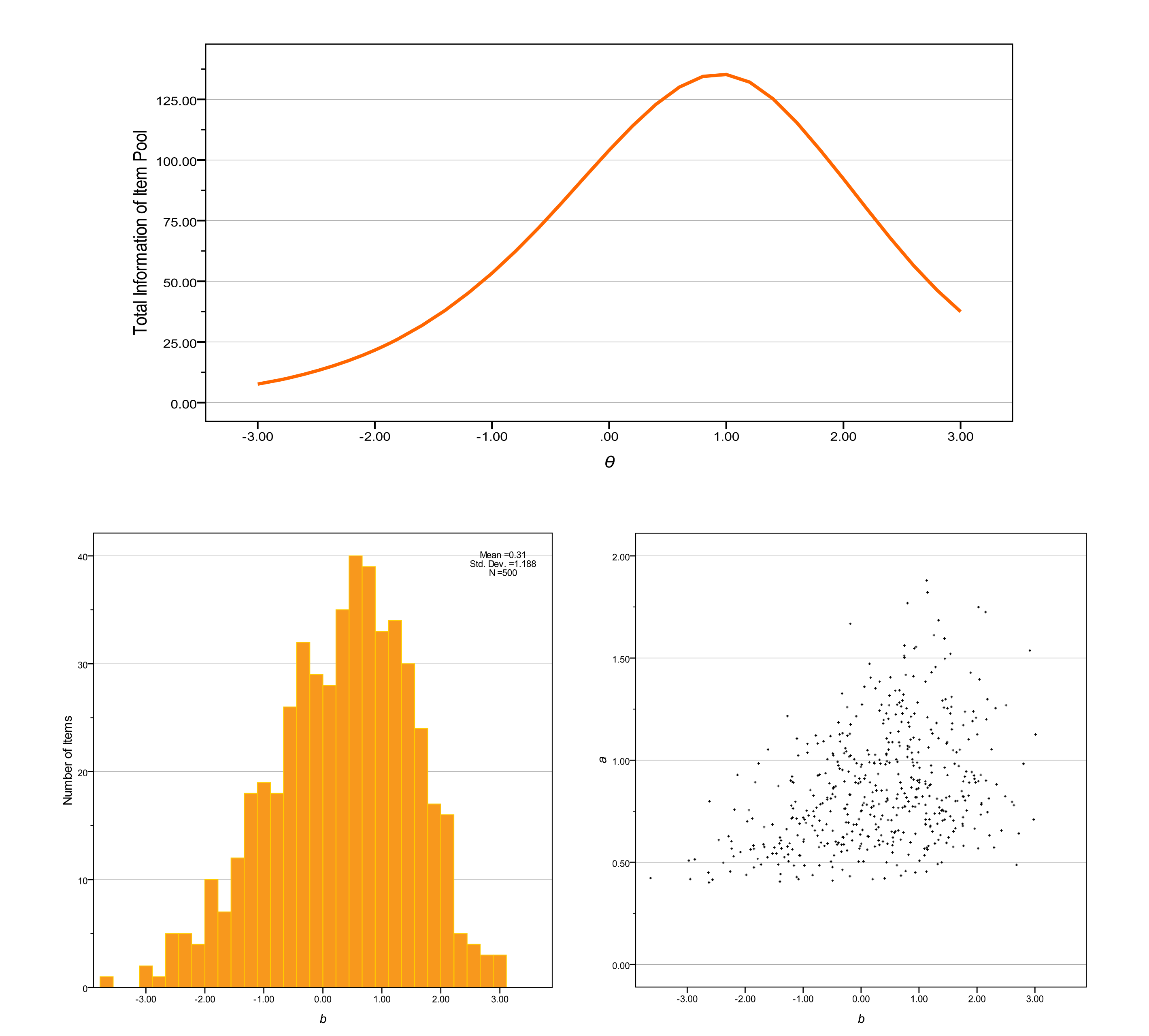
A series of simulation studies was conducted to evaluate the effectiveness of the GMIR approach. The simulation studies mimicked 1 month (20 administration days) of an existing CAT program for higher education (with simplified content balancing) and used the evaluation criteria of item exposure rate, test information, item pool usage, and proficiency estimation bias and errors.

Data

To construct the item pool, 500 multiple-choice math items[1] were drawn from the GMAT® item bank. The aggregated total information of the item pool showed the peak around $\theta = 1$ (top of Figure 2), not only because there was a large number of hard items (bottom-left of Figure 2) but also because the hard items tended to be more discriminating (bottom-right of Figure 2). To simplify the study and to increase the generalizability of the results, the constraints on the content balancing were ignored.

[1] The size of the item pool (n=500) in this study was not the actual size of the operational GMAT® item bank.

Figure 2. Aggregated Information of Item Pool (Top), Item Difficulty (Bottom-Left), and Correlation Between a-b Parameters (Bottom-Right)



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 the test items 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

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

where C_x was the constraint of the item exposure rate, which was 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 pool. 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

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

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 theta 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 conducted based on the item usage information that was updated up to the previous time slot.

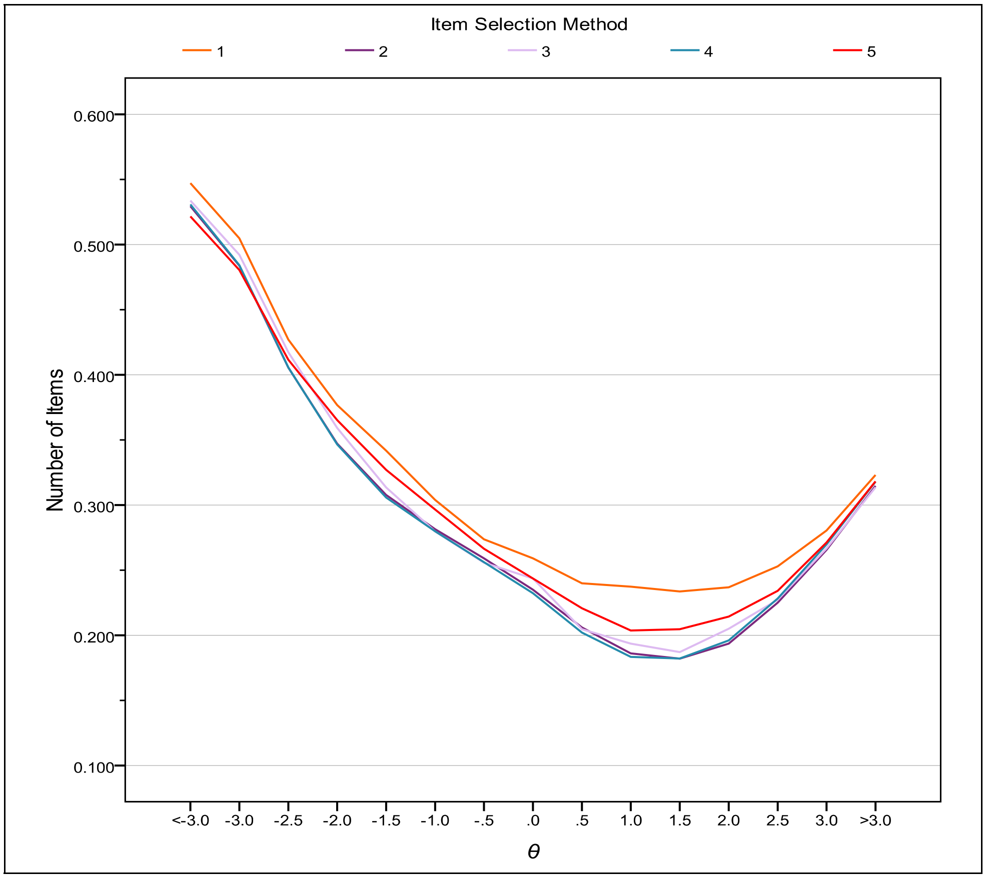
The evaluation of the five item selection methods focused on two major points: (1) performance of theta estimation, and (2) item pool usage. First, the performance of the theta estimation was evaluated by the standard errors of theta estimates (SEE) across the theta scale. The bias and mean absolute error of the theta 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 see which item selection method resulted in the most optimal item pool usage.

The whole procedure was replicated 100 times and the median values were taken to be reported.

Results

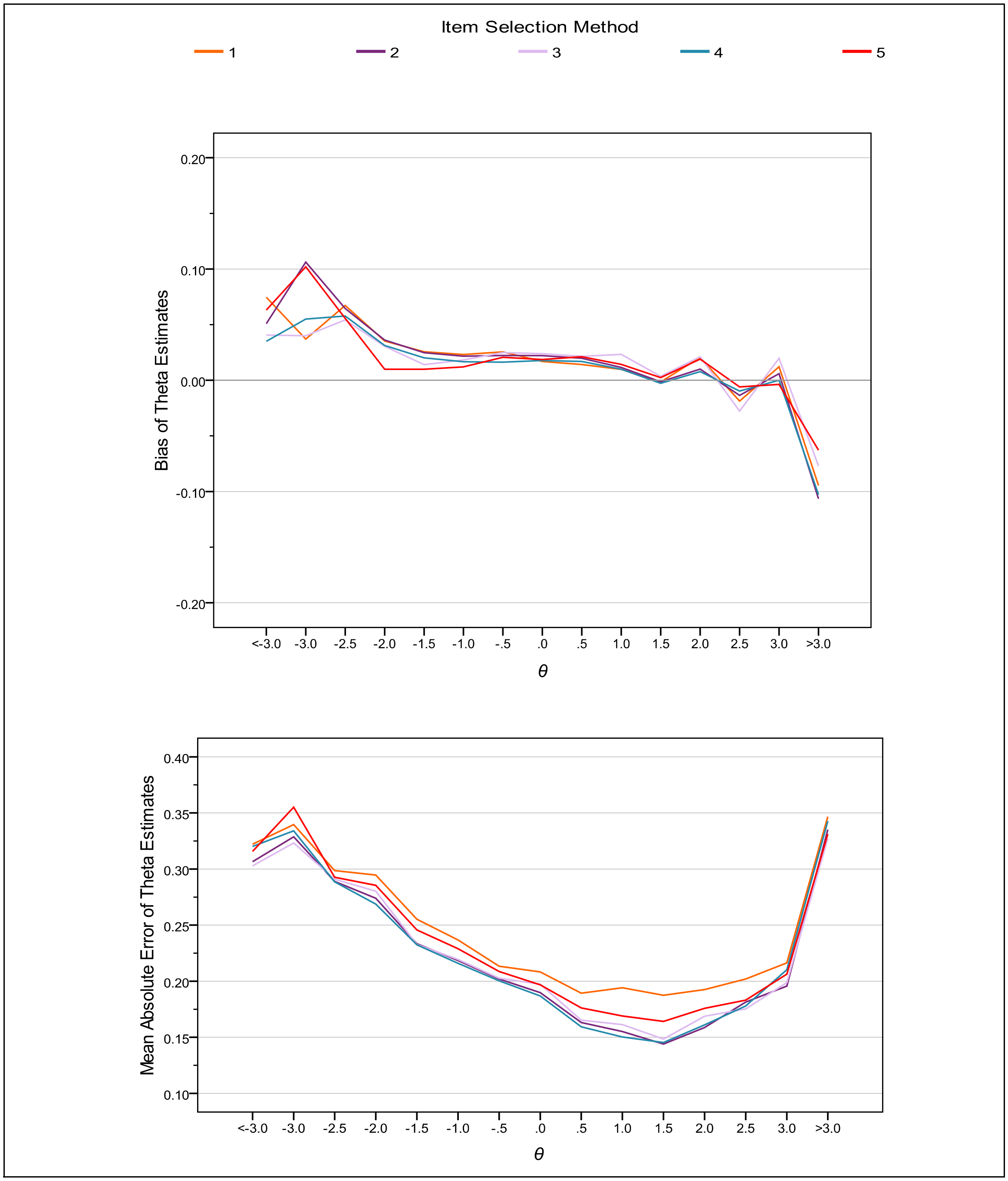
The standard errors of theta estimation are plotted in Figure 3. The item selection in Methods 2 and 4 showed the smallest SEE across the theta scale, whereas Method 1 resulted in the largest SEE. This was the expected result; the MFI and GMIR approaches select items to maximize the test information 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 the item exposure more rigorously (Methods 3 and 5), the SEE was slightly increased across the theta scale.

Figure 3. Standard Error of Theta Estimation



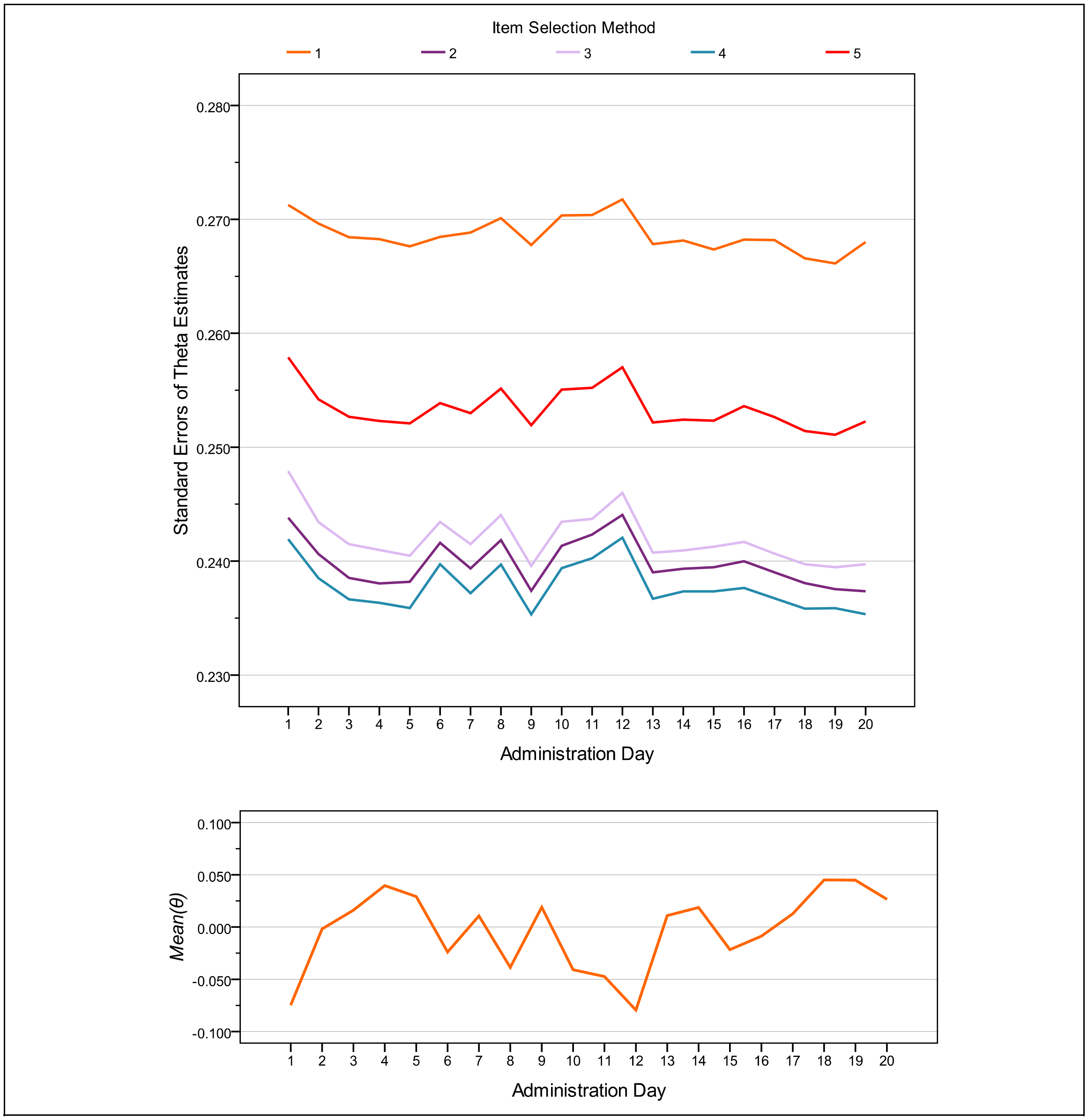
In terms of the estimation bias, there was no meaningful difference among the item selection methods (top of Figure 4). For the majority of the theta area (-2.0 ~ 2.0), a small positive bias was observed, but the magnitude of the bias was minimal (about 0.025 in average). As an empirical measure of the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

Figure 4. Bias (Top) and Mean Absolute Error (Bottom) of Theta Estimation



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 in Figure 5. There were visible fluctuations in the mean SEE over time (see top of Figure 5), but these were mainly due to the change in the examinee distribution day by day (bottom of Figure 5). 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 + item exposure constraint) resulted in the smallest SEE over time, and Methods 2 and 3 (MFI + item exposure constraint/MFI + fade-away method) closely followed. With Method 5 (GMIR + fade-away method), the SEE was slightly increased, and Method 1 (modified randomesque + item exposure constraint) resulted in the substantially increased SEE. As shown in Figure 5, the impact of a choice of item selection persisted over time.

Figure 5. Mean Standard Errors of Theta Estimation (Top) and True Mean Theta (Bottom) for Each Administration Day



The study also evaluated the effectiveness of the item selection methods in terms of the item pool utilization. In Figure 6, the item exposure rates of all 500 items in the item pool were 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. Finally, Method 5 had no items that were either used up to the exposure limit or not used at all. More important, the item pool usage was very well balanced in Method 5. Figure 6 shows the exposure rate of a majority of the items in the pool clustered around 0.10, with few items exposed over 0.15.

In Figure 7, the items in the pool 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 pool. 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 pool. On the other hand, those two methods caused excess usage on more than 125 items, or about 25 percent of the item pool. Such extremely unbalanced item pool usage can be a serious problem in maintaining the item pool over the long term. With Methods 1 and 3, there were fewer extreme cases of unbalanced item pool usage; however item usage still varied considerably. Method 5, as illustrated in Figures 6 and 7, resulted in the most balanced item pool usage. Nearly 240 items (or about 48 percent of the item pool) were exposed between 40 percent and 60 percent of the exposure limit.

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

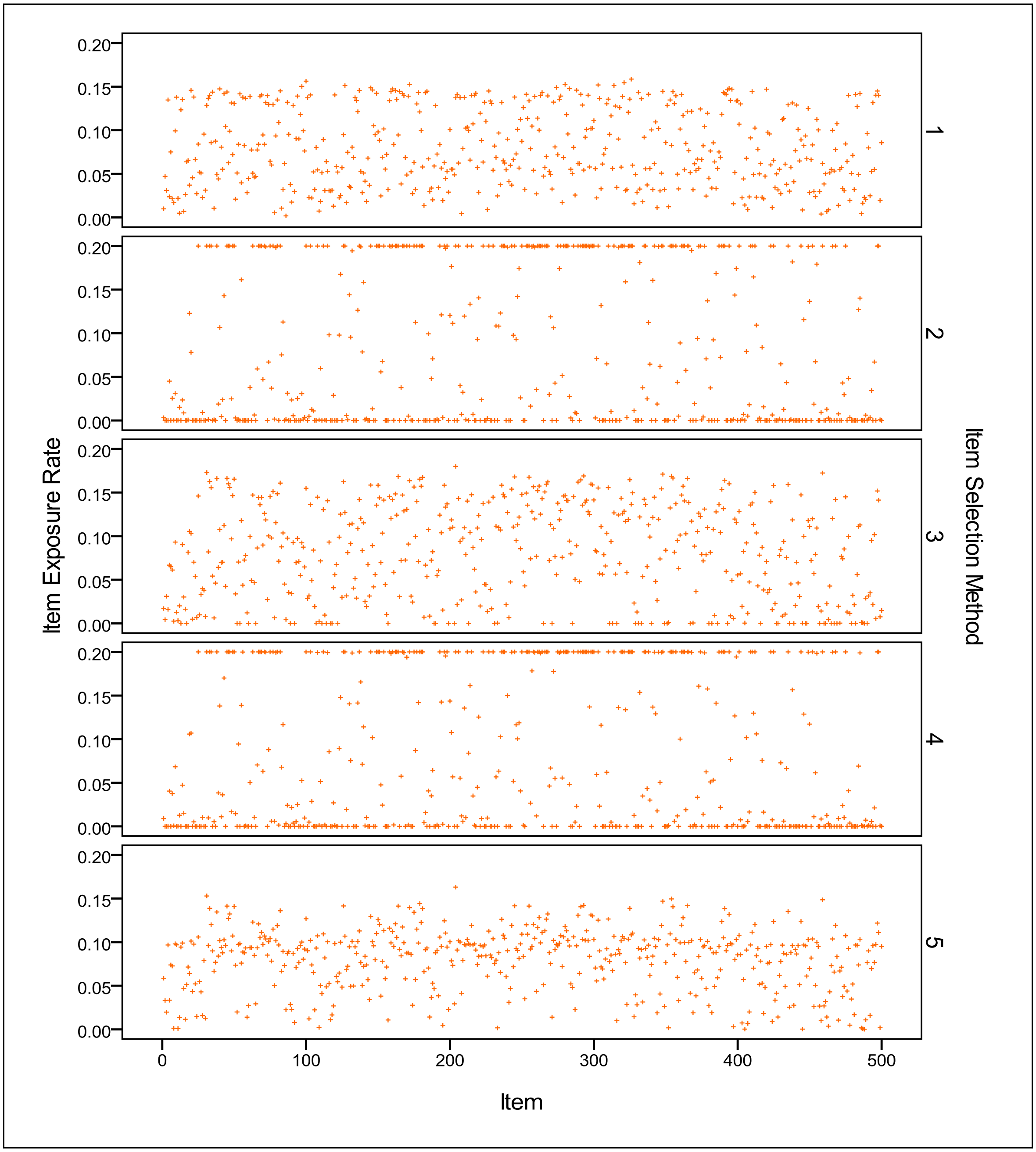
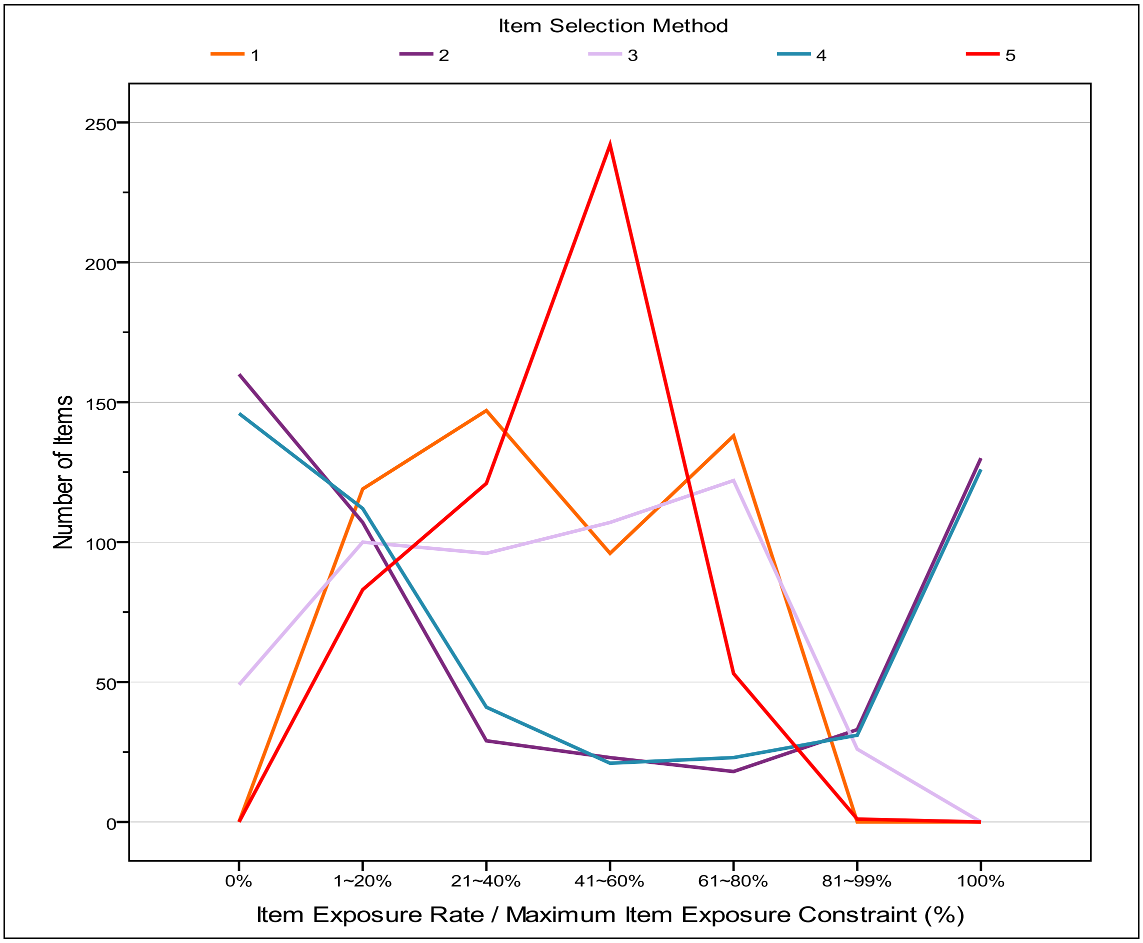


Figure 7. Item Pool Usage With Each Item Selection Method



Discussion

Developing an item selection algorithm including the item exposure control may not necessarily be an ideal process for establishing a theory. Rather, it can 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, the 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 a slightly lower SEE than the MFI-based methods (Methods 2 and 3). It is possible the GMIR strategy of selecting the most efficient item was more robust against the instability of the interim $\hat{\theta}$ in the early stage 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 still could be seen as one of the most effective methods resulting in the maximized test information.

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

Another potential problem with the unbalanced item pool 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 pool used in CAT administration was only about 350 items, not 500. Such a decrease in the item pool size increases the chance that more examinees see 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 the item exposure. In the simulation study, nearly 90 percent of the item pool 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 the Method 5 (Figure 7). Therefore, compared with the other studied methods, Method 5 was clearly more effective and efficient in utilizing the item pool.

The trade-off between maximizing the 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 the item pool utilization with the 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 pool 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 pool usage. When the GMIR approach was used with the fade-away item exposure control method (Method 5), item pool 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.

Conclusion

The goal of any kind of testing is to produce an accurate measure of what is to be assessed and the accuracy of test measurements is mainly determined by the test information at each examinee's proficiency level. Computerized adaptive testing (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 the 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 stage 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 pool utilization compared with the MFI method while minimizing the compromise of the test precision.

Contact Information

For questions or comments regarding study findings, methodology or data, please contact the GMAC Research and Development Department at research@gmac.com.

Author

Kyung (Chris) T. Han is Associate Psychometrician in the Research and Development Department at Graduate Management Admission Council®.

Acknowledgements

An earlier version of this paper was presented by Kyung (Chris) T. Han at the GMAC Computer Adaptive Testing Conference, Minneapolis, MN, June 2–3, 2009.

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* (chap.17-20). Reading, MA: Addison-Wesley.

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.

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

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*(4), 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*(4) 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 Centre.

van der Linden, W. J. (2000). Constrained adaptive testing with shadow test. 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.