

Running head: Non-Fisher-Information Item Selection Criteria

Comparison of Non-Fisher-Information Item Selection Criteria

in Fixed-Length Computerized Adaptive Testing

Kyung T. Han

Graduate Management Admission Council®

Correspondence may be sent to:

Kyung T. Han

Graduate Management Admission Council

1600 Tysons Blvd. Suite 1400, McLean, VA 22102 USA

khan@gmac.com

(Phone) +1-703-245-4363

(Fax) +1-703-749-0169

The views and opinions expressed in this article are those of the author and do not necessarily reflect those of the Graduate Management Admission Council®.

Abstract

The maximized Fisher information (MFI) criterion has been the mainstream of the item selection algorithm in many computerized adaptive test (CAT) programs because of its effectiveness and simplicity. There are still several issues with the MFI criterion, however, that need to be resolved in the field, specifically regarding estimation accuracy at the beginning of CAT administration and item pool utilization. Several non-MFI criteria (or approaches) to item selection have been developed, but practitioners still lack insight into the consequential differences among those non-MFI approaches in terms of proficiency estimation and item pool utilization. This study compared five non-MFI item selection approaches (a-stratification, interval information, likelihood weighted information, global information, and gradual maximum information ratio) and attempted to find the item selection criteria that struck a good balance between performance and efficiency in pool utilization. The study found that the gradual maximum information ratio criterion resulted in the most efficient pool use while achieving proficiency estimation as good as the MFI with various test lengths.

Keywords: computerized adaptive test, item selection, item exposure control

**Comparison of Non-Fisher-Information Item Selection Approaches
in Fixed-Length Computerized Adaptive Testing**

One of the most widely used—and probably the oldest—item selection methods in computerized adaptive testing (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. Basically, this involves finding item x maximizing $I[\hat{\theta}_{m-1}]$ for an examinee with the interim proficiency estimate $\hat{\theta}$ and $m-1$ as the number of items administered so far. (Weiss, 1982). Taking a typical case of a multiple-choice item pool, where item characteristics are defined by the three-parameter logistic model, or 3PLM (Birnbaum, 1968), the item selection method based on the MFI criterion looks for item i that results in the largest value of

$$I_i[\hat{\theta}_{m-1}] = \frac{(Da_i)^2(1-c_i)}{[c_i + e^{Da_i(\hat{\theta}_{m-1}-b_i)}][1 + e^{-Da_i(\hat{\theta}_{m-1}-b_i)}]^2}, \quad (1)$$

where a_i , b_i , and c_i are the discrimination, difficulty, and pseudo-guessing parameters in 3PLM, respectively, and D is a scaling constant whose value is 1.702. The MFI criterion has been popular because it is an effective means of administering CAT that results in maximized test information for each individual. In the early stage of testing, however, when five or fewer items have been administered, for example, the interim proficiency estimates are rarely accurate. So at the start of testing, items selected according to the MFI criterion tend not to provide as much information as they were supposed to at the interim proficiency estimates. Another problem with the MFI approach is that this method tends to select items with higher a -parameter values more frequently than it selects items with lower a -parameter values. Such uneven usage of items with

the MFI method may create serious problems in item pool maintenance. To overcome these problems with the MFI criterion, especially at the early stage of CAT administration, several non-MFI item selection criteria have been developed.

Non-Fisher-Information Item Selection Approaches

Chang and Ying (1996) came up with the global information approach that uses the moving average of Kullback-Leibler information (KLI) (Cover & Thomas, 1991; Kullback, 1959) to select items. The KLI for any θ for the i th item with response X_i is defined by

$$K_i(\theta \parallel \theta_0) = P_i(\theta_0) \log \left[\frac{P_i(\theta_0)}{P_i(\theta)} \right] + [1 - P_i(\theta_0)] \log \left[\frac{1 - P_i(\theta_0)}{1 - P_i(\theta)} \right], \quad (2)$$

where $P_i(\theta_0)$ is the probability that a random examinee at the proficiency level θ_0 answers the item correctly. The moving average of KLI is then taken and used as the item selection criterion, as follows,

$$K_i(\theta_0) = \int_{\theta_0 - \delta}^{\theta_0 + \delta} K_i(\theta \parallel \theta_0) d\theta, \quad (3)$$

where δ specifies the range of the moving average. Determining δ could be a little ambiguous, but Chang and Ying (1996) proposed c / \sqrt{m} as a reasonable choice for δ with constant c selected according to a specified coverage probability and with m being the number of items administered thus far. Chang and Ying (1996) found that replacing the MFI criterion with the KLI criterion could often reduce the biases and mean squared errors of proficiency estimation when the test length was short or the CAT administration was in its early stage ($m < 30$).

Veerkamp and Berger (1997) came up with two other alternatives to MFI. In the first, called the interval information criterion (IIC), the information function is averaged across the

confidence interval of an interim proficiency estimate. The mathematic expression of IIC for item i is

$$\int_{\hat{\theta}_L}^{\hat{\theta}_R} I_i[\theta]d\theta, \quad (4)$$

where $\hat{\theta}_L$ and $\hat{\theta}_R$ are a confidence interval of θ . The actual mean value for IIC is Equation 4 divided by the length of the confidence interval, but taking an average of it is unnecessary for the purpose of item selection.

In the other alternative approach proposed by Veerkamp and Berger (1997), the so-called likelihood weighted information criterion (LWI), the information function is summed up throughout the theta scale, weighted by the likelihood function after item administrations so far. Thus, with the LWI criterion, the item to be selected is item i that results in the maximized value of

$$\int_{\theta=-\infty}^{\infty} L(\theta; x_{m-1})I_i[\theta]d\theta. \quad (5)$$

, where $L(\theta; x_{m-1})$ is the likelihood function of the response vector x_{m-1} after $(m-1)$ th item administration. According to Veerkamp and Berger's simulation study (1997), both IIC and LWI resulted in item selection performance comparable to but not necessarily superior to the MFI criterion.

As clearly shown in Equations 3, 4, and 5, the KLI, IIC, and LWI approaches evaluate items based on item information functions (Kullback-Leibler information or Fisher information) across a range of θ instead of a point estimate of θ . This means all three non-MFI approaches tried to cope with the large standard errors of Fisher information by replacing a point estimate with a range of estimates. It should be noted, however, that these three non-MFI item selection criteria were not intended to resolve the other critical problem with the MFI criterion—the

excessive use of items with higher a -parameter values (and poor utilization of items with lower a -parameter values).

Chang and Ying (1999) looked into the problem with the MFI criterion in terms of poor item pool utilization. To prevent a waste of items with higher a -parameter values at the beginning of CAT, Chang and Ying (1999) proposed stratifying items in the item pool by a -parameter values. In their method, so called “ a -stratification,” an item with a b -parameter value that is closest to the interim $\hat{\theta}$ is the one to be selected from the item stratum with the lowest a -parameter values at the beginning of CAT. Using this method, items are selected from item strata with higher a -parameter values as the CAT administration proceeds. According to their simulation results, the a -stratification method was the best method for reducing the uneven use of items by far, while also managing the increase in estimation errors and biases, compared with the MFI method. The a -stratification did have some downsides. First, stratifying items could substantially limit the number of available items to the size of each item stratum and might eventually cause another type of item overexposure especially when the size of the item pool is small and/or there are additional item selection controls such as content balancing. Second, item stratification is solely based on a -parameter values; the c -parameter values are left uncontrolled. Third, it is not unusual to observe a correlational relationship between a - and b -parameter values, so if items are stratified according to a -parameter values, there is a chance that the item strata will not be equivalent to each other in terms of b -parameters. Fourth, when the test length is not fixed, item selection performance with this method may not be as effective as expected. Some of these problems with the original a -stratification method were relieved somewhat by a number of variations of the a -stratification method that were developed later (Chang, Qian, & Ying, 2001; Yi & Chang, 2001; Deng & Chang, 2001).

In his recent study, Han (2009) also looked at the poor item pool utilization issue associated with the MFI method. To improve the level of item pool utilization, he proposed selecting items at the beginning of CAT administration based on expected item efficiency instead of the MFI criterion. Expected item efficiency is defined as the level of realization of item's potential information at interim $\hat{\theta}$. Thus, if item i results in its maximum potential information at θ_i^* , the expected item efficiency at interim $\hat{\theta}$ is computed by

$$\frac{I_i[\hat{\theta}_{m-1}]}{I_i[\theta_i^*]}, \quad (6)$$

where θ_i^* is equal to b_i when either a 1PL or 2PL model is used. If a 3PL model is used and $c_i \neq 0$, θ_i^* can be computed using Birnbaum's solution (1968):

$$\theta_i^* = b_i + \frac{1}{Da_i} \log\left(\frac{1 + \sqrt{1 + 8c_i}}{2}\right). \quad (7)$$

Han (2009) suggested taking item effectiveness (i.e., expected item information) into account over the item efficiency as CAT administration proceeds and reaches to the end. Thus, this approach looks for an item that maximizes the criterion,

$$\frac{I_i[\hat{\theta}_{m-1}]}{I_i[\theta_i^*]} \left(1 - \frac{m}{M}\right) + I_i[\hat{\theta}_{m-1}] \frac{m}{M}, \quad (8)$$

where M is the test length, and m is 1 plus the number of items administered thus far. This method is referred to as the gradual maximum information ratio (GMIR) approach. The first part of Equation 8 is the item efficiency term; the second part is the item effectiveness term (the Fisher information; Equation 1). Each part of Equation 8 is inversely weighted by the progress of the CAT administration. Based on his simulation results, Han (2009) found that the GMIR approach could generally improve the item pool utilization compared to the MFI criterion.

Purpose of Study

A few studies have provided evidence showing the effectiveness of each of the non-MFI criteria in item selection in comparison to the MFI criterion (Chang & Ying, 1996, 1999; Veerkamp & Berger, 1997; Han, 2009). When practitioners choose an item selection criterion for their CAT programs, however, they still do not have a clear picture of the differences among each of the non-MFI item selection criteria because few studies have actually compared one non-MFI criterion to another as well as to the MFI criterion. This is especially true regarding the level of item pool utilization by these non-MFI item selection criteria, where knowledge in the field is even shallower. Thus, in this article, the five non-MFI item selection criteria will be compared to each other as well as to the MFI criterion. Based on the simulation study results, each of the item selection criteria will also be discussed in terms of how well they balance between proficiency estimation and item pool utilization.

Simulation Study

A series of simulation studies was conducted to evaluate the effectiveness of the KLI, IIC, LWI, a -stratification, and GMIR criteria for item selection. The simulation studies mimicked one month (20 administration days) of an existing CAT program for higher education and used the evaluation criteria of biases and errors in proficiency estimation and level of item pool utilization.

Method

Data. To construct the item pool, 500 multiple-choice math items¹ were drawn from the Graduate Management Admission Test[®] (GMAT[®]) item bank. The aggregated total information of the item pool showed the peak around $\theta = 1$, not only because there was a large number of hard items but also because the hard items tended to be slightly more discriminating (Figure 1). To simplify the study and to increase the generalizability of the results, the constraints of content balancing were ignored.

Three different conditions of test length were simulated: 10, 20, and 40 items per examinee. To make the item usage equivalent and comparable across the three test-length conditions, the numbers of examinees also varied to 80,000, 40,000, and 20,000, respectively. Therefore, the total item usage was fixed to 800,000 across the conditions. Examinee's true proficiency was drawn from the standard normal distribution ($\sim N(0,1)$).

Item exposure control. The main computer server collected item usage information for each item after each test time slot via the online network that connected the server and client computers (i.e., test terminals) in test centers. Each client computer received the updated item usage information just before the start of the next test administration. Once a client computer started a test administration, there was no communication between the server and client computer until the test administration was finished. Such network technology enabled the CAT system to use near real-time item exposure information, precluding the need to predict the item exposure by other means, for example, using the Sympon-Hetter method (Sympon & Hetter, 1985), which involves iterative simulations.

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

For this study, the KLI, IIC, LWI, α -stratification, and GMIR criteria were implemented in the simulation as well as the MFI criterion, which served as a baseline. Each of the criteria was teamed up with the item exposure control (IEC), which consisted of three components, all of which were based on the up-to-date item exposure information.

In the first IEC component, the absolute limit on the exposure count of each item was applied. In this simulation, the CAT system let the item retire from the item pool once an item was used 3,000 times.

The second component of the IEC involved applying relative exposure limit per item. The relative exposure rate was constrained to be less than 0.20; those items exceeding the constraint were temporarily prevented from being selected. In other words, the CAT system was designed to limit the percentage of examinees seeing the same item to 20% during the 20-day test period.

For the third IEC component, the item selection criteria value for each eligible item in the pool was inversely weighted by the ratio between the updated item usage and absolute item exposure limit. For example, with the MFI criterion, it looked for an item that maximized

$$I_i[\hat{\theta}_{m-1}] \frac{U_i}{C}, \quad (6)$$

where C was the absolute item usage limit (of the first exposure control component), which was 3,000 in this study. U_i was the item usage for the life of item i . With this method, rarely used items were expected to be promoted more frequently, whereas excessively used items were likely to “fade away” from the item selection (this method will be referred to hereafter as the fade-away method (FAM)).

CAT implementation. Twenty-five test centers were simulated, with the assumption that each test center accommodated 20 test takers per test time slot. Thus, in each test time slot, 500 examinees were administered simultaneously. The numbers of test time slots per day were eight, four, and two for the test-length conditions of 10, 20, and 40, respectively. The test was administered for the 20-day administration period.

The computer software *SimulCAT* (Han, 2010) was used to simulate the CAT administration. The first item for each examinee was chosen randomly among items whose b -parameter value was between -0.5 and 0.5. The interim $\hat{\theta}$ was estimated using the Bayesian maximum a posteriori (MAP) method after each item administration. The theta estimation algorithm limited the change of the interim $\hat{\theta}$ from the previous estimate to ± 1 .

In the simulation, it was assumed that each client computer (i.e., test terminal) only communicated with the item bank server 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 non-MFI item selection methods focused on two major points: (a) performance of theta estimation, and (b) item pool utilization. First, the study evaluated the performance of the theta estimation using the standard errors of theta estimates (SEE) across the theta scale. The study also computed the bias and mean absolute error of the theta estimates. Second, the study analyzed the item exposure rate at the item level to see which item selection method resulted in optimal item pool usage. Standard deviation (SD) of item usage was evaluated as well as the bipolarized usage index (BUI), which was computed by

$$\text{BUI} = \frac{R + U}{N}, \quad (7)$$

where R , U , and N were the numbers of retired items, unused items, and all items in the item pool, respectively. Item usage by a -parameter value was also investigated.

Results

Proficiency Estimation

Mean SEE statistics from the 20 days of test administration with each of the item selection methods are shown in Figure 2. The a -stratification method resulted in the largest mean SEE for most of the theta area. With the LWI criterion, the mean SEE tended to be larger in the lower theta area, but it started to show smaller SEE when $\theta > -0.5$. With the IIC criterion, a strange bumpy pattern for the mean SEE was observed around $\theta = -0.5$. One reason for this pattern with the IIC criterion is that there were several cases where the interim $\hat{\theta}$ changed dramatically from the previous interim $\hat{\theta}$ even in the later part of the CAT administration; as a result, the SEE jumped for those cases. The MFI, KLI, and GMIR criteria consistently resulted in the lowest SEE compared to the other three criteria. Increase in test length resulted in a decrease in the mean SEE regardless of the choice of item selection criterion. The difference in the mean SEE among the six-item selection criteria became much smaller when the test length increased to 40.

Empirical estimation error and bias were also assessed across the item selection criteria. As shown in Figure 3, when the test length was short ($M = 10$), the IIC criterion showed relatively larger mean absolute error (MAE) in the middle of the θ scale, and the LWI tended to have larger MAE at both extremes on the θ scale compared to the other item selection criteria. When the test length was 20 or larger, the difference in MAE among the item selection criteria

became negligible except for the a-stratification method. When the a-stratification method was used, MAE tended to be flatter throughout the θ scale; the MAE observed in the higher θ was larger than that observed with the other item selection criteria.

Estimation bias is summarized in Figure 4. Overall, examinees' proficiency tended to be overestimated for those who were at the lower proficiency level and underestimated for those who were at the higher proficiency level. This kind of bias pattern is not unusual when the MAP method is used to estimate proficiency because the scale of estimates tended to be shrunk toward the mean. When the test length was long ($M = 40$), the estimation bias became minimal regardless of the item selection criterion chosen. When the test length was 10, the LWI criterion showed the smallest positive bias for those who were at the lower proficiency level, but unlike the other item selection criteria, its bias positively increased even more for those who were at the higher proficiency level.

Based on the SEE, MAE, and bias results, it was clear that the MFI, KLI, and GMIR criteria nearly equally outperformed the other three item selection criteria (IIC, LWI, and a-stratification) in proficiency estimation, especially when the test length was shorter.

Item Pool Utilization

The test length was fixed to 10, 20, or 40 for studied conditions and the number of examinees was changed to 80,000, 40,000, or 20,000 for the corresponding test-length conditions, allowing total item usage to remain the same (800,000) across conditions. This enabled evaluation of the item pool utilization with each six-item selection criterion to be conducted by directly comparing the SD of item usage, bipolarized usage index, and the difference in item usage by a-parameter value.

Table 1 shows the number of items retired after 3,000 exposures, which was the absolute exposure limit of the IEC in this study, as well as the number of items that were never used during the 20 days of the CAT administration. The first thing to note in Table 1 is that the number of items either retired or never used was very large when the test length was short. This is because of the initial value for $\hat{\theta}$, which, for each examinee was a random value between -0.5 and 0.5. The first few items administered for each examinee were very likely to be items within the difficulty of that $\hat{\theta}$ range, whereas other items with a difficulty level farther away from that initial $\hat{\theta}$ range would have much less chance of being used. As a result, when the test length was short (e.g., $M = 10$), observations of extremely uneven item use was not unexpected. Because comparison of different test-length conditions was not meaningful, the study focused on comparisons among the item selection criteria within each test-length condition.

Based on the bipolarized usage index, the MFI and KLI criteria resulted in the most uneven item use. The a-stratification and GMIR approaches showed much better item pool utilization when the test length was 10 and 20. The IIC and LWI criteria seemed to better utilize the item pool than the MFI and KLI criteria, but not as well as the a-stratification and GMIR methods. The results based on SD of the item usage also concurred with the findings based on the bipolarized usage index. When the test length was 40, however, according to the bipolarized usage index, the a-stratification method resulted in the next worst item pool utilization after the MFI criterion. A primary reason for this was that there were fewer items available within each item stratum in the a-stratification method, which resulted in excessive use of some items over others within each item stratum. On the other hand, the GMIR criterion showed the most impressive level of item pool utilization, with only 1% of items retired (five out of 500) and another 1% of items not used.

Although the bipolarized usage index and SD of item usage provided information indicating how even or uneven the item usage was with each item selection criteria, it was still not known which items were used more than the others were. Generally, when the MFI is used, items with higher a -parameter values are used more often than other items. A correlation coefficient between the item usage and a -parameter value is often used to evaluate such a tendency, but a correlation analysis could be misleading in this case because the absolute usage limit (= 3,000) for IEC created a ceiling for the item usage. Instead, for this study, a visual investigation of the item usage was conducted. On the horizontal axis of Figure 5, the 500 items in the pool are listed by a -parameter value in ascending order from left to right. Each bar in each cell represents the usage of each item. When the test length was 10 or 20, the items with higher a -parameters (right side in each cell) were used more intensively with the MFI, IIC, LWI, and KLI criteria; items with low a -parameters (left side in each cell) were used rarely. With the a -stratification method, many items were used up to the exposure limit, but there was no correlational relationship between the item usage and a -parameter value. With the GMIR criterion, items with higher a -parameters tended to be used more often, but the items with very low a -parameters were still used frequently. When the test length was 40, the differences in item usage among the item selection criteria were more obvious.

Evaluating the overall results of the item pool utilization, the GMIR criterion clearly outperformed and stood out from the MFI, IIC, LWI, and KLI in utilizing the entire item pool. The result for the a -stratification method was somewhat mixed. It was very effective in utilizing items with very low a -parameters, but it often used many items up to the usage limit regardless of a -parameter value.

Discussion and Conclusion

If item pool utilization occurred under ideal circumstances, every item in the pool would be used with the same usage level. In CAT, however, it is nearly impossible to achieve equal item usage across all items without seriously compromising test quality, unless the item pool is constructed so that it follows examinees' proficiency distribution exactly. Such a scenario again would be practically impossible. Therefore, it is critically important to find an item selection method that best utilizes an entire item pool while producing accurate and stable proficiency estimates.

The simulation study in this article found that the KLI and GMIR criteria as well as the MFI criterion outperformed the other non-MFI criteria (a-stratification, IIC, and LWI) in the proficiency estimation regardless of test-length conditions. The a-stratification method showed estimation errors slightly higher than the other criteria most of the time, and the IIC and LWI criteria resulted in unstable estimates in certain θ areas, especially when the test length was short. On the other hand, in terms of item pool utilization, the simulation results in this study clearly revealed the effectiveness of the GMIR approach over the other item selection criteria. Practitioners often face a dilemma when they choose an item selection method because there are usually tradeoffs between proficiency estimation performance and item pool utilization. This was not the case with the GMIR approach in this study. The GMIR approach was as effective as the MFI in proficiency estimation, and its item use was very efficient at the same time, not wasting items with low a -parameter. The advantages of the GMIR approach remained the same whether the test length was short or long.

Although the findings in this study clearly favored the GMIR approach over other MFI and non-MFI item selection criteria, it is worth noting that use of different item exposure control methods with the item selection criteria could yield a different set of results. Thus, a comparison of the non-MFI item selection criteria teamed up with various item exposure control methods would be a worthy subject for a future study.

This study focused on a situation where the test length was fixed, but CAT with flexible test length would be also another important area for future examination. In some non-MFI item selection approaches like a-stratification and GMIR, the test length must be predetermined. When the difference between the predetermined (or expected) test length and the actual test length is substantially large in a flexible length CAT, the performance and efficiency of those item selection methods are likely to decline. What is still unknown, however, is the extent to which deterioration would occur if actual test length missed the expectation. Thus, the question to be addressed in the near future is how robust are the non-MFI item selection criteria against missed test length in flexible test length CAT.

References

- Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (Chap. 17–20). Reading, MA: Addison-Wesley.
- Chang, H.-H., & Ying, Z. (1996). A global information approach to computerized adaptive testing. *Applied Psychological Measurement, 20*, 213–229.
- Chang, H.-H., & Ying, Z. (1999). Alpha-stratified multistage computerized adaptive testing. *Applied Psychological Measurement, 23*, 211–222.
- Chang, H.-H., Qian, J., & Ying, Z. (2001). Alpha-stratified multistage computerized adaptive testing with beta blocking. *Applied Psychological Measurement, 25*, 333–341.
- Cover, T. M., & Thomas, J. A. (1991). *Elements of information theory*. New York: Wiley.
- Deng, H., & Chang, H.-H. (2001, April). *A-Stratified computerized adaptive testing with unequal item exposure across strata*. Paper presented at the annual meeting of the American Educational Research Association (AERA), Seattle, WA.
- Georgiadou, E., Triantafillou, E., & Economides, A. (2007). A review of item exposure control strategies for computerized adaptive testing developed from 1983 to 2005. *Journal of Technology, Learning, and Assessment, 5*(8). Retrieved from <http://www.jtla.org>.
- Han, K. T. (2009). *A gradual maximum information ratio approach to item selection in computerized adaptive testing*. Research Reports 09–07, McLean, VA: Graduate Management Admission Council.
- Han, K. T. (2010). SimulCAT: Simulation software for computerized adaptive testing [computer program]. Retrieved March 20, 2010, from <http://www.hantest.net/>

- Kingsbury, G. G., & Zara, A. R. (1989). Procedures for selecting items for computerized adaptive tests. *Applied Measurement in Education*, 2(4), 359–375.
- Kullback, S. (1959). *Information theory and statistics*. New York: Wiley.
- Luecht, R. M. (2003, April). *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*. *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 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.

Veerkamp, W. J. J., & Berger, M. P. F. (1997). Some new item selection criteria for adaptive testing, *Journal of Educational and Behavioral Statistics*, 22(2), 203–226.

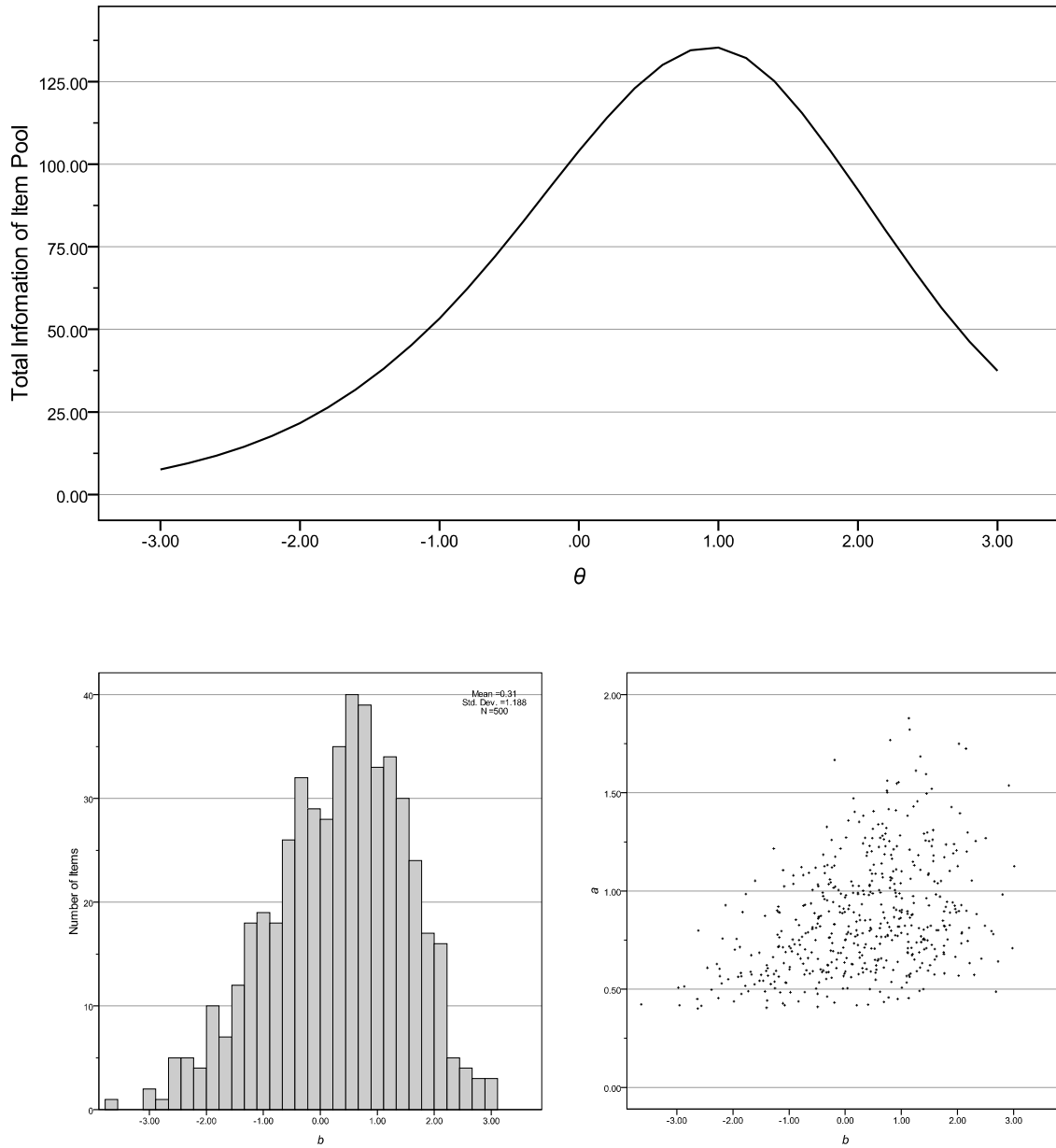
Weiss, D. J. (1982). Improving measurement quality and efficiency with adaptive testing. *Applied Psychological Measurement*, 6, 473–492.

Yi, Q., & Chang, H.-H. (2001, June). *a-Stratified computerized adaptive testing with content blocking*. Paper presented at the Annual Meeting of the Psychometric Society, King of Prussia, PA.

Table 1. Level of Item Pool Utilization

Test Length	Item Selection Method	Number of Retired Items	Number of Items Never Used	Bipolarized Usage Index	SD of Item Usage
10	MFI	232	189	0.84	1423.20
	a-Strat.	208	141	0.70	1368.30
	IIC	213	181	0.79	1389.62
	LWI	228	173	0.80	1400.70
	KLI	234	185	0.84	1432.96
	GMIR	184	152	0.67	1312.47
20	MFI	199	142	0.68	1345.70
	a-Strat.	160	39	0.40	1197.87
	IIC	171	131	0.60	1287.70
	LWI	182	130	0.62	1323.17
	KLI	192	135	0.65	1343.62
	GMIR	110	70	0.36	1094.98
40	MFI	59	57	0.23	1071.18
	a-Strat.	100	14	0.23	1027.80
	IIC	26	37	0.13	940.38
	LWI	41	39	0.16	995.96
	KLI	53	48	0.20	1048.36
	GMIR	5	5	0.02	743.82

Figure 1. Aggregated Information of the Item Pool (Top), Item Difficulty (Bottom-Left), and Correlational Relationship between a - b Parameters (Bottom-Right)²



² This figure was adapted from Han (2009).

Figure 2. Standard Error of Theta Estimation

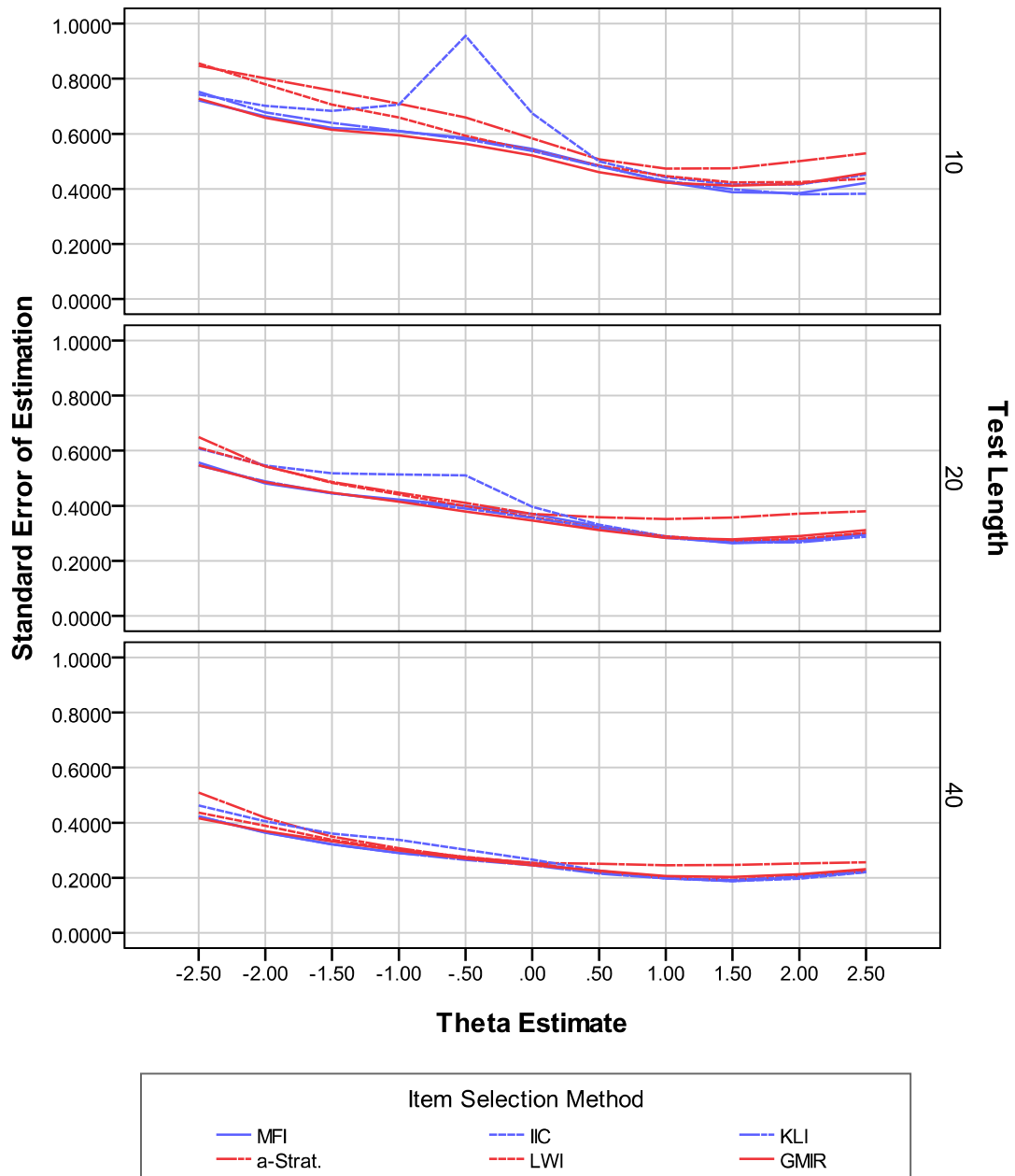


Figure 3. Mean Absolute Error of Theta Estimation

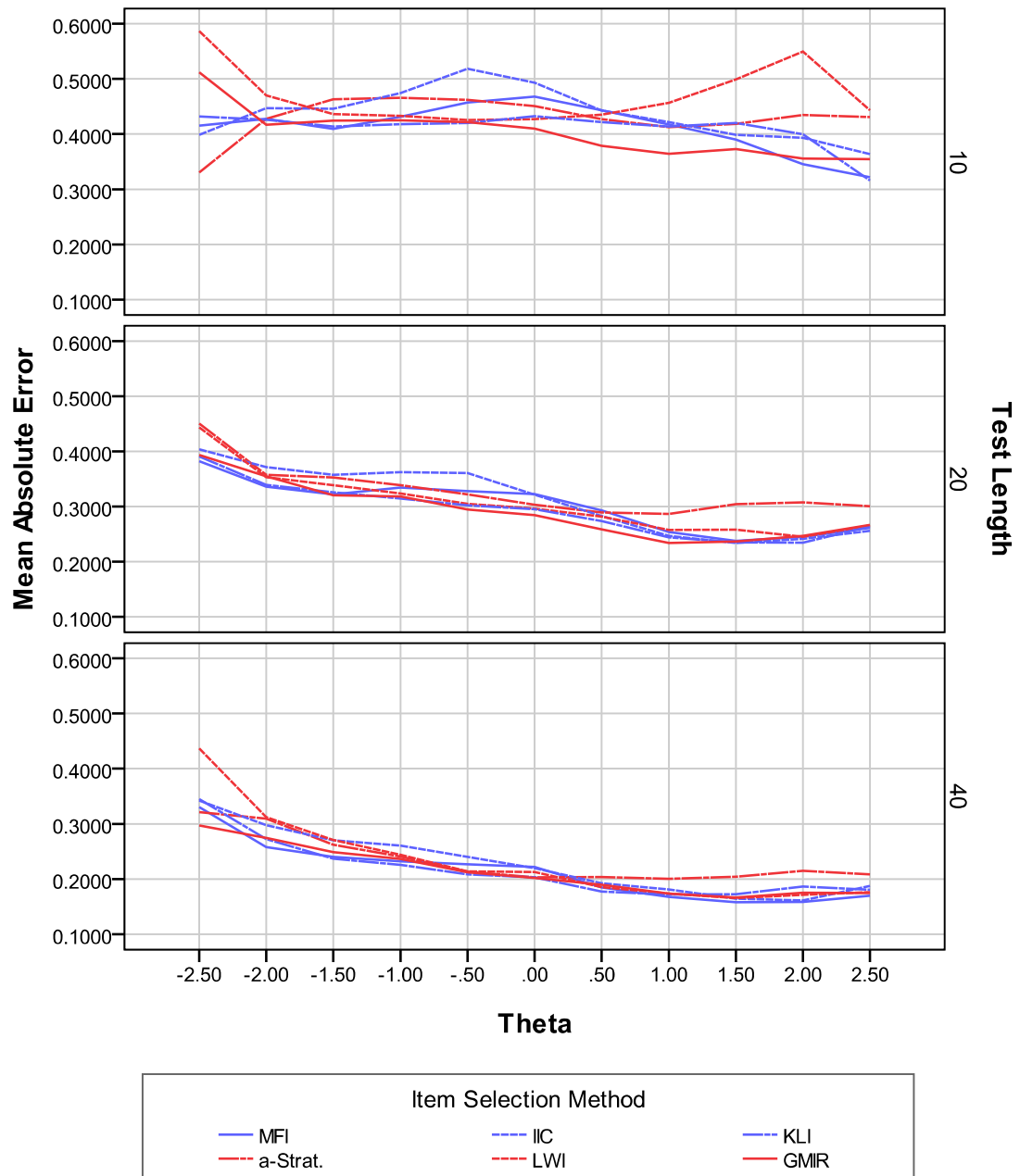


Figure 4. Bias of Theta Estimation

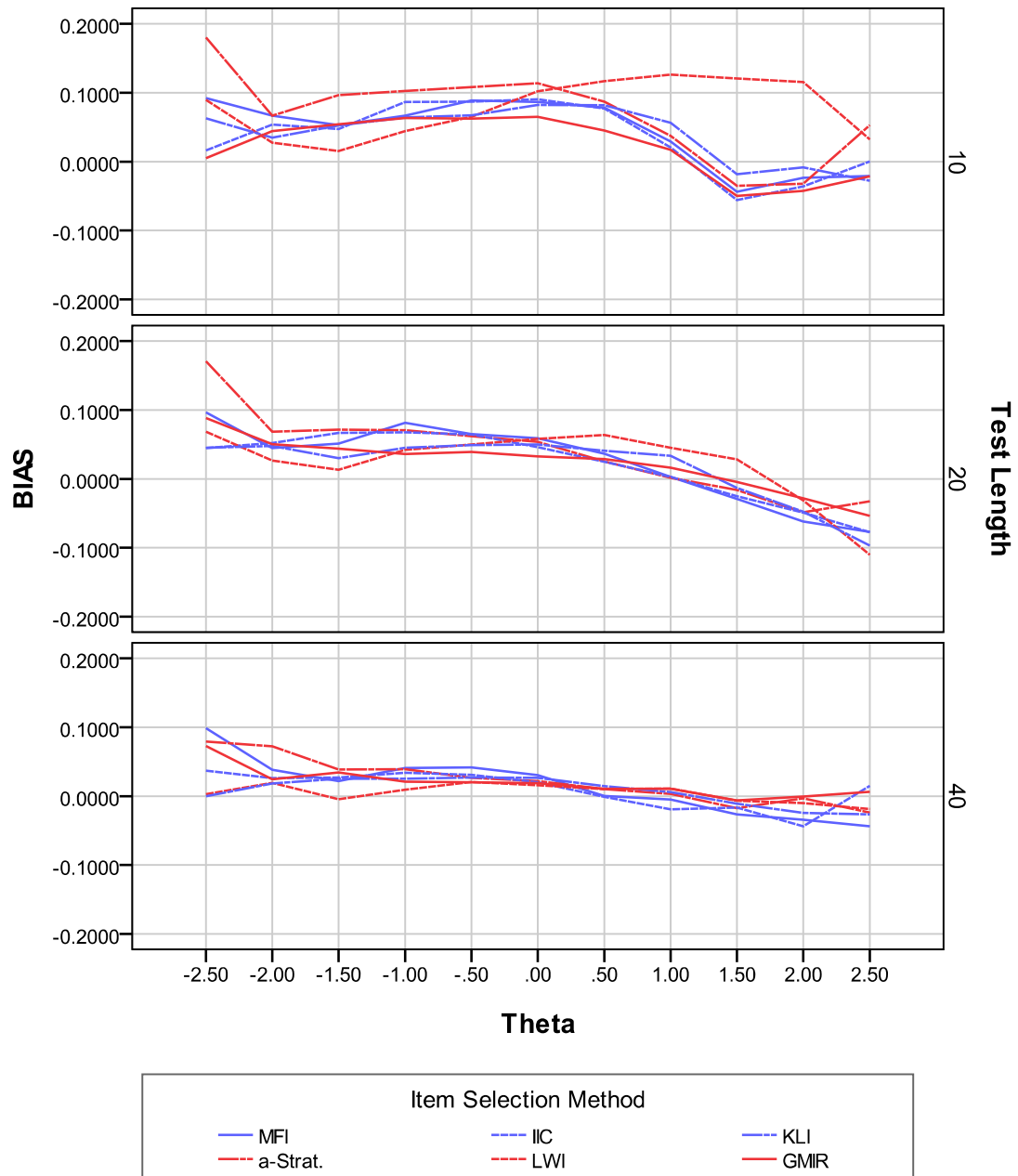


Figure 5. Item Usage by a-Parameter Value

