Graduate Management Admission Council®

A Comparison of Three Methods of Item Selection for Computerized Adaptive Testing

Denise R. Costa and Camila A. Karino
Center for Selection and Promotion of Events (Cespe), University of Brasilia

Fernando A. S. Moura
Federal University of Rio of Janeiro

Dalton F. Andrade
Federal University of Santa Catarina

Presented at the CAT Research and Applications Around the World Poster Session, June 2, 2009

Abstract

One of the most important components of CAT is the set of procedures used for item selection. Unlike traditional paper-and-pencil tests, adaptive procedures administer items that fit the examinee's level of ability. This selection is based both on the characteristics of the items (e.g., item difficulty or discrimination parameters) and on the estimated ability of the examinee. This study is a work in progress that aims to evaluate the performance of three different CAT item selection methods. The first is based the maximum information criterion, one of the most popular item selection methods in CAT. The second method is based on the global information method as defined by Chang and Ying (1996), which uses the Kullback-Leibler measure. The third selection method is based on the predictive analysis defined by the maximum expected information criterion proposed by van der Linden (1998). To evaluate the three different methods, five different simulation studies were conducted for an item bank containing 246 items of the University of Brasilia's Instrumental English test. The resulting database was fit by a three-parameter logistic model on a scale with mean 0 and standard deviation of 1. The examinees' θs were estimated using expected a posteriori (EAP). An initial analysis of bias and mean square error suggested that all methods performed similarly to estimate examinees' θs. Databank-related characteristics, however, might have influenced those measures, since it is not yet an ideal item bank for CAT implementation. With these results, it can be concluded that there is no apparent statistical difference in relation to θ estimation for the three methods examines with the analyzed item bank.

Acknowledgment

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

Citation

Costa, D. R., Karino, C. A., Moura, F. A. S., & Andrade, D. F. (2009). A comparison of three methods of item selection for 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

Denise R. Costa: Rua 13 de Maio Quadra 53 Lote 8B, Planatina, Distrito Federal, Brazil.
Email: denise.reis@gmail.com

A Comparison of Three Methods of Item Selection for Computerized Adaptive Testing

As the main idea of computerized adaptive testing (CAT) is to efficiently select items for a good estimate of the examinee's ability, this evaluated the performance of three adaptive selection methods in CAT: maximum information (MI); Kullback-Leibler (KL) information defined by Chang and Ying (1996); and maximum expected information (MEI), proposed by van der Linden (1998). In order to do so, an item bank from the Instrumental English I test, provided by the Center for Selection and Promotion of Events (Cespe) of the University of Brasilia, was used.

The item bank used refers to the Instrumental English proficiency test taken by graduating students at the University. This test has been used since 2004 and consists of 450 questions (items). It takes place twice a year and fundamentally aims to induce the student to practice reading comprehension strategies that allow a more efficient reading of varied texts in English. Any regular student of the University of Brasilia is eligible for enrollment, and the approved students will receive credits for the Instrumental English I course. In order to evaluate the adequacy of the bank items for CAT implementation, classical test theory (CTT) and item response theory (IRT) analyses were conducted. During those analyses, some items were removed because they lacked psychometric quality. As a result, 246 items of the original 450 were used and calibrated by a 3-parameter IRT logistic model on a $(0,1)$ scale.

To compare the three best-known procedures for CAT item selection, five simulation studies were implemented. After the selection of an item, the examinee's responses were simulated from a Bernoulli $[P(\hat{\theta})]$ distribution, where $P(\hat{\theta})$ represents the probability of getting an item correct according to the 3-parameter logistic model, considering both the item bank parameters and the current θ estimate fixed.

The iterative method for θ estimation was e*xpected a posteriori* (EAP). After the observation of each response (u_i =1 for correct or u_i =0 for incorrect), the EAP estimator for θ after $k-1$ items was numerically calculated by:

$$\hat{\theta}_{\dot{u}_1,\ldots,u_{k-1}}^{EAP} \approx \frac{\sum_{t=1}^{q} X_t A_t \Delta_t^{-1} \left\{ \prod_{i=1}^{k-1} P_{ti}(X_t)^{u_i} \left[1 - P_{ti}(X_t)\right]^{1-u_i} \right\}}{\sum_{t=1}^{q} A_t \Delta_t^{-1} \left\{ \prod_{i=1}^{k-1} P_{ti}(X_t)^{u_i} \left[1 - P_{ti}(X_t)\right]^{1-u_i} \right\}}, \tag{1}$$

where X_t represents the quadrature points, $t = 1,\ldots,q$,

A_t, the weights associated to X_t;

Δ_t, the interval ranges;

$P_{ti}(X_t) = c_i + \frac{(1-c_i)}{1+\exp\left[-Da_i\left(X_t - b_i\right)\right]}$, the 3-parameter IRT logistic model in X_t.

The posterior variance associated withthe EAP estimatewas computed as

$$Var[\theta_j \mid u_1,\ldots,u_{k-1}] \approx \frac{\sum_{t=1}^{q}\left[X_t - \hat{\theta}_{j_{u_1,\ldots,u_{k-1}}}^{EAP}\right]^2 A_t\Delta_t^{-1}\left\{\prod_{i=1}^{k-1}P_{ti}(X_t)^{u_i}\left[1-P_{ti}(X_t)\right]^{1-u_i}\right\}}{\sum_{t=1}^{q}A_t\Delta_t^{-1}\left\{\prod_{i=1}^{k-1}P_{ti}(X_t)^{u_i}\left[1-P_{ti}(X_t)\right]^{1-u_i}\right\}}. \quad (2)$$

The prior assumed distribution was the uniform $(-6.0, 6.0)$, which implies that the examinee distribution was completely concentrated in the -6.0 to 6.0 interval. 40 equally spaced points in the ± 6 standard deviation interval were used, and the weights associated with those points equaled the prior density.

Procedures

The item selection procedures used in this study can be synthesized the following manner:

***Beginning*:** Specification of an initial value, $\hat{\theta}_0$, for each examinee.

***Iteration*:** Estimation of the examinee's $\hat{\theta}$ immediately after responding to each item. The choice of the following item in CAT considers the information of the items at $\hat{\theta}$. Let i be the *i*th item of the $i = 1,...,I$ item bank and k the order in which that *i*th item is shown on the adaptive test. Suppose that $k-1$ items were administered by CAT. The administered items index from the following set: $S_{k-1} = \{i_1,\ldots,i_{k-1}\}$. The remaining items form another set, $R_k = \{1,\ldots,I\} \setminus S_{k-1}$. The selection of the *k*th item obeyed one of the following rules:

***(1) Maximum Information (MI)*:** Given $\hat{\theta}$, the *k*th item was selected such that $I_{F,k}(\hat{\theta}_{k-1})$ was the item with the greatest value, i.e.:

$$i_k \equiv argmax_s \left\{ I_{F,U_s}\left(\hat{\theta}_{u_1,\ldots,u_{k-1}}\right) : s \in R_k \right\}, \quad (3)$$

where:

$$I_{F,k}(\theta) = \frac{\left[\frac{\partial P_k(\theta)}{\partial \theta}\right]^2}{P_k(\theta)[1-P_k(\theta)]}. \quad (4)$$

***(2) Kullback-Leibler (KL)*:** Given $\hat{\theta}$, the *k*th item was selected such that $K_k(\hat{\theta}_{k-1})$ was the item with the greatest value, i.e.:

$$i_k \equiv argmax_s \left\{ \int_{\hat{\theta}_{k-1}-\delta_k}^{\hat{\theta}_{k-1}+\delta_k} K_s\left(\theta \,||\, \hat{\theta}_{k-1}\right) : s \in R_k \right\}, \tag{5}$$

where:

$$K_i(\theta \,||\, \theta_0) = P_i(\theta_0)\log\left[\frac{P_i(\theta_0)}{P_i(\theta)}\right] + \left[1-P_i(\theta_0)\right]\log\left[\frac{1-P_i(\theta_0)}{1-P_i(\theta)}\right] \tag{6}$$

and $\delta_k = \frac{3}{\sqrt{k}}$.

***(3) Maximum Expected Information (MEI)*:** Given $\hat{\theta}$, the *k*th item was selected such that $J_k(\hat{\theta}_{k-1})$ weighted by the predictive probability $P_k(u_k \,|\, u_1,\ldots,u_{k-1})$ was the item with the largest value of the bank, i.e.,

$$\begin{aligned} i_k \equiv argmax_s \{ & P_s(0 \,|\, u_1,\ldots,u_{k-1}) J_{u_1,\ldots,u_{k-1},U_s=0}(\hat{\theta}_{u_1,\ldots,u_{k-1},U_s=0}) \\ & + P_s(1 \,|\, u_1,\ldots,u_{k-1}) J_{u_1,\ldots,u_{k-1},U_s=1}(\hat{\theta}_{u_1,\ldots,u_{k-1},U_s=1}) : s \in R_k \}, \end{aligned} \tag{7}$$

where $P_s(u_s \,|\, u_1,\ldots,u_{k-1}) = \int P_s(u_s \,|\, \theta) g(\theta \,|\, u_1,\ldots,u_{k-1}) d\theta$ is the predictive probability of the response u_s,

$J_{u_1,\ldots,u_{k-1}}(\theta) = -\frac{\partial^2 \log L(\theta; u_1,\ldots,u_{k-1})}{\partial \theta^2}$ is the curvature of the likelihood,

and

$L(\theta; u_1,\ldots,u_{k-1}) = \prod_{i=1}^{k-1} P_i(\theta)^{u_i} \left[1-P_i(\theta)\right]^{1-u_i}$.

***Stop*:** The algorithm halted when a predefined number of items had been administered or when a precision level for θ had been achieved.

The simulations of the adaptive tests were made in the R software environment. In each study, different start and stop criteria were used.

It is important to remember that some items were administered sequentially in CAT as they refer to a common element (a text or a figure), which happens often in language tests. According to the literature, whenever an item bank is developed in such a manner, this set of connected questions is known as a testlet (Wainer, Bradlow, & Du, 2003). Thus, restrictions were applied to the item selection algorithm so that, for each testlet, the examinees had the opportunity to respond to the whole group and then that group was never again selected during the CAT.

Results

Simulation Study 1

This study aimed to assess the necessary number of CAT items in order to obtain estimates of θ with a standard error less than or equal to 0.4 and 0.2. In order to do this, 500 adaptive tests were simulated for each of the adaptive item selection methods (MI, KL and MEI). The individuals were simulated as having θs ranging from -3 to $+3$. For this simulation, it was determined that the examinees would answer questions from the Instrumental English I bank until the predetermined precision level of the estimate was reached.

Table 1 presents the distribution of the number of items necessary to reach a standard error of 0.4 and 0.2 for each item selection method. It should be noted that the standard error measure associated with the θ estimate is the square root of the posterior variance.

Table 1. Distribution of 500 Simulated Cases to Evaluate the Number of Items Required by the Adaptive Test With Standard Errors (SE) of .4 and .2

Number of Items Required by CAT	SE = .4			SE = .2		
	MI	KL	MIE	MI	KL	MIE
Less than or equal to 10	41	38	81	-	-	-
Between 11 and 20	300	310	270	-	-	-
Between 21 and 30	95	99	95	-	-	-
Between 31 and 40	32	23	20	-	-	-
Between 41 and 50	10	10	11	77	79	68
Between 51 and 60	4	8	8	60	57	68
Between 61 and 70	2	2	4	42	50	36
Greater than or equal to 71	16	10	11	321	314	328

When the stopping criterion was set at .4 standard errors, it was observed that 341 of the 500 adaptive simulated tests under the MI item selection criterion had 20 items or less. For the KL criterion, this number rose to 348. The MEI method showed that 70.2% of the tests had a maximum of 20 items. When the stoppage criterion was set to .2, however, all simulated cases required more tha 40 items.

Table 2 presents the average number of items, as well as their standard deviations (SD), required by CAT in order to obtain estimated θs with a standard error of .4 and .2 for each of the three methods.

Table 2. Average Number of Items in 500 Simulations

Measure	SE = .4			SE = .2		
	MI	KL	MEI	MI	KL	MEI
Mean	23	22	21	136	137	138
SD	30.4	22.1	27.1	84.2	85.4	84.5

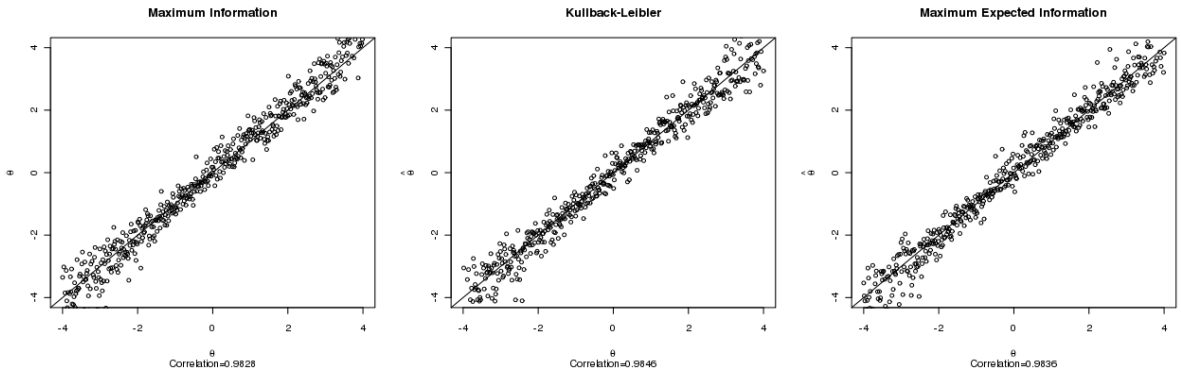
It was observed that the average number of items on the 500 simulated tests administered in order to obtain an SE less than .4 using the MI method was 23, against 22 for KL and 21for MEI. When theSE was .2, the average number of items on the 500 tests was six times greater (136 by MI, 137 by KL, and 138 by MEI).

Simulation Study 2

In this study, 500 answers from examinees with θs between -4 and $+4$ were simulated. The initial value of θ for each examinee was equal to $\hat{\theta}_0 = 0.00$. The first CAT item was selected at random among all of those with difficulty parameter below the initial $\hat{\theta}_0$ (i.e., $b < 0.00$) and the test terminated when 25 items had been administered.

Figure 1 shows the distribution of θ estimates by true θ, according to each item selection criterion.

Figure 1. True θ Versus Estimated θ for Each Item Selection Method



All three methods estimated θ quite well for a test with 25 items. The smallest correlation between the estimates and their true value was achieved by the MI method (0.9828).

As expected, θs at the distribution's tails ($|\theta| \geq -3.50$) were those that had performed the worst under the three item selection methods. This happened because there were few items in the bank with such extreme difficulty parameters.

It is worth mentioning that, in order to generate the 500 CAT simulations through the MI method, the algorithm implemented in R took around 50 minutes. The KL adaptive method took approximately an hour and 25 minutes to run, and MEI took about one hour and 34 minutes.

Table 3 summarize the logic behind the adaptive process, for a single examinee. The items are presented in the same order they were administered by CAT, as well as the IRT item parameters (*a*, *b*, and *c*), the examinee's answers, the iterative estimates, and the associated standard errors . Three examples of θ levels ($\theta = -1.50$, $\theta = 0.00$ and $\theta = 1.50$) were simulated using MI for selection of items. For each simulation, the initial θ value was $\hat{\theta}_0 = 0.00$. The item that started all algorithms was ING10836, which had a *b* parameter value of -2.51.

Table 3. Simulated Adaptive Test for $\theta = -1.50$ by the MI Method

Order	Item	*a*	*.b*	*c*	Answer	$\hat{\theta}$	SE
1	ING10836	0.80	-2.51	0.10	0	-4.00	1.36
2	ING10833	0.88	-3.56	0.11	1	-3.18	1.36
3	ING10834	0.97	-2.40	0.11	1	-2.30	1.29
4	ING20626	1.05	-1.93	0.04	0	-2.94	1.04
5	ING10701	0.92	-2.32	0.12	1	-2.44	0.89
6	ING10622	0.93	-1.69	0.04	1	-1.92	0.73
7	ING10614	1.12	-1.58	0.09	1	-1.57	0.67
8	ING20627	1.68	-1.03	0.03	0	-1.79	0.56
9	ING20625	1.20	-1.17	0.04	0	-1.92	0.54
10	ING20611	0.96	-1.58	0.04	0	-2.07	0.54
11	ING10831	0.91	-1.73	0.10	1	-1.89	0.47
12	ING10624	1.00	-1.50	0.10	1	-1.74	0.43
13	ING20705	1.15	-1.38	0.17	1	-1.61	0.41
14	ING10613	1.17	-1.14	0.08	0	-1.70	0.40
15	ING10547	0.89	-1.61	0.05	1	-1.60	0.37
16	ING20622	1.29	-0.91	0.04	0	-1.65	0.36
17	ING20807	1.08	-1.24	0.11	0	-1.72	0.36
18	ING20605	0.87	-1.42	0.04	1	-1.63	0.33
19	ING20701	1.07	-1.12	0.11	0	-1.68	0.33
20	ING10839	0.90	-1.68	0.11	1	-1.62	0.31
21	ING10724	0.91	-1.51	0.11	1	-1.56	0.30
22	ING10937	0.97	-1.31	0.12	1	-1.50	0.29
23	ING10634	1.28	-0.91	0.11	0	-1.54	0.29
24	ING10628	0.91	-1.08	0.03	0	-1.58	0.29
25	ING10629	0.90	-1.10	0.04	1	-1.51	0.27

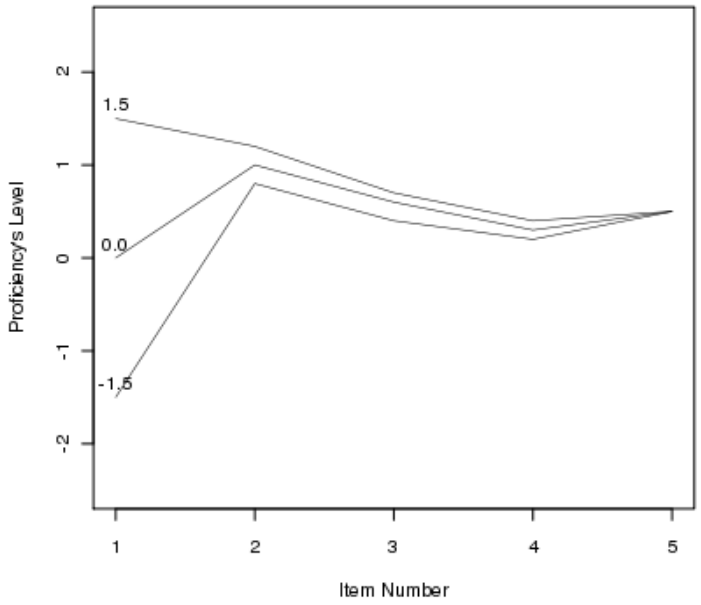
Table 3 shows that, by incorrectly answering question ING10836, the θ estimate changed from $\hat{\theta}_0 = 0.00$ to $\hat{\theta}_1 = -4.00$, with standard deviation equal to 1,36. After the computation of Fisher's Information for $\hat{\theta}_1 = -4.00$, the MI-simulated examinee correctly answered the question, and the θ estimate increased to $\hat{\theta}_1 = -3.18$. The iterative procedure was conducted until the 25th item was answered. The value of the estimate at the end of the process was $\hat{\theta}_{25} = -1.51$, close to its true value, $\theta = -1.50$. It is interesting that 56% of the 25 administered items were answered correctly, resulting in an estimated θ with precision equal to or greater to that of a traditional 50-item test.

Simulation Study 3

This study was designed to assess the difference between the methods for adaptive item selection when setting different initial values for the examinees' θs . Thus, three values were fixed for the algorithm's initialization. The first was $\hat{\theta}_0 = -1.50$; the second, $\hat{\theta}_0 = 0.00$ and the third, $\hat{\theta}_0 = 1.50$. As ilustrated in Figure 2, this simulation study had as its objective to assess

whether, when starting the adaptive test with different θ estimates, the methods of item selection would converge the same way or if one would show better performance than the others.

Figure 2. Example of Simulation 3



For each of he three starting v alues, θ was replicated 100 times to evaluate the variability of the estimates, since individuals with the same θ can have different patterns of response and, consequently, result in different estimated θs. The maximum test length in this simulation was fixed at 25. The results can be seen in Figures 3 to 5.

Figure 3. Distribution of Estimates When $\theta = -1.50$ for Different $\hat{\theta}_0$

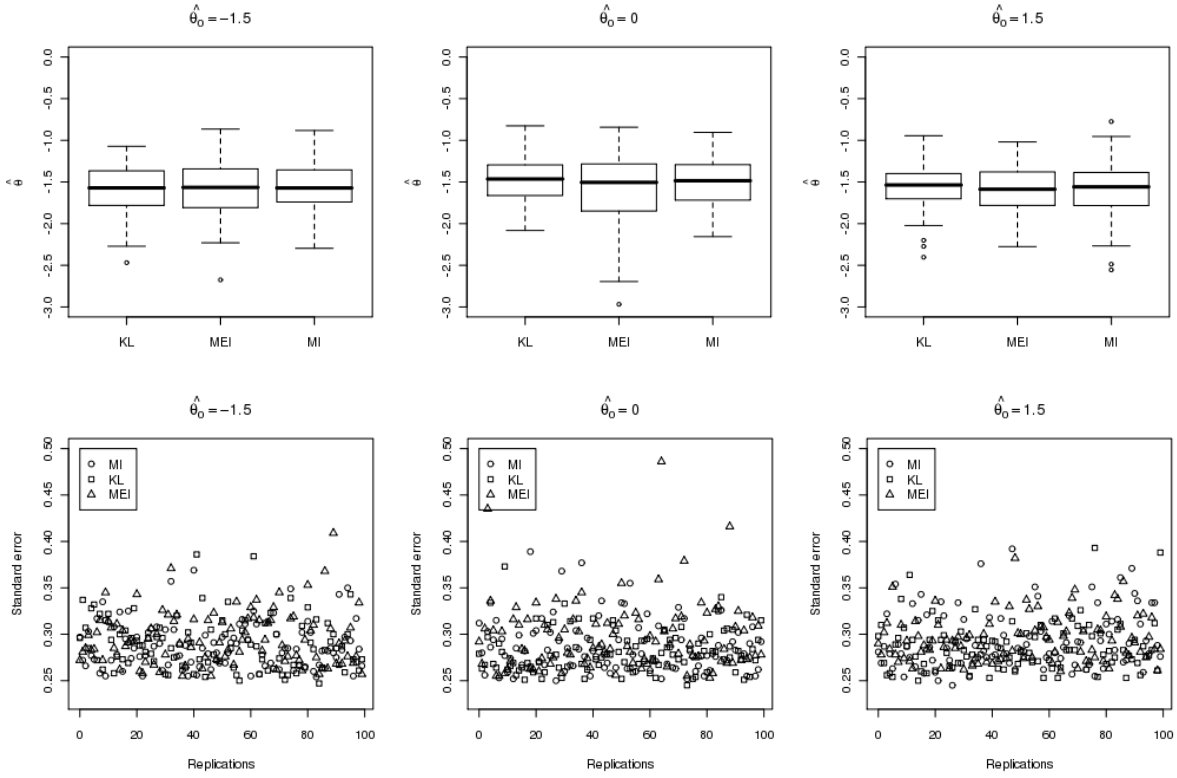


Figure 3 shows how the θ estimates for the simulated individuals with $\theta = -1.50$ were distributed among different initial values for $\hat{\theta}_0$. For $\hat{\theta}_0 = -1.50$, it was observed that, although the true value of θ was fixed as the initial value for the examinee's theta , it was later found that the θ estimates ranged from -1.77 to -1.35. In general, it could be that the three methods for item selection had similar performance when $\hat{\theta}_0 = -1.50$ with 50% of the SEs less than or equal to .28.

When the initial value was $\hat{\theta}_0 = 0.00$ (Figure 4), it was observed that MEI had the largest range for the values of the estimates (the lowest value was $\hat{\theta} = -2.97$ and the highest was -1.28). In general, the value of the standard errors in the estimation of $\theta = -1.50$ varied between 0.24 and 0.30 for the three item selection methods.

For $\hat{\theta}_0 = 1.50$ (Figure 5), it was observed that the three item selection criteria presented estimates close to their true values even when the algorithm started with a θ value above the true value. The SEs in this simulation varied, on average, between 0.25 and 0.32.

Figure 4. Distribution of Estimates When $\theta = 0.00$ for Different $\hat{\theta}_0$

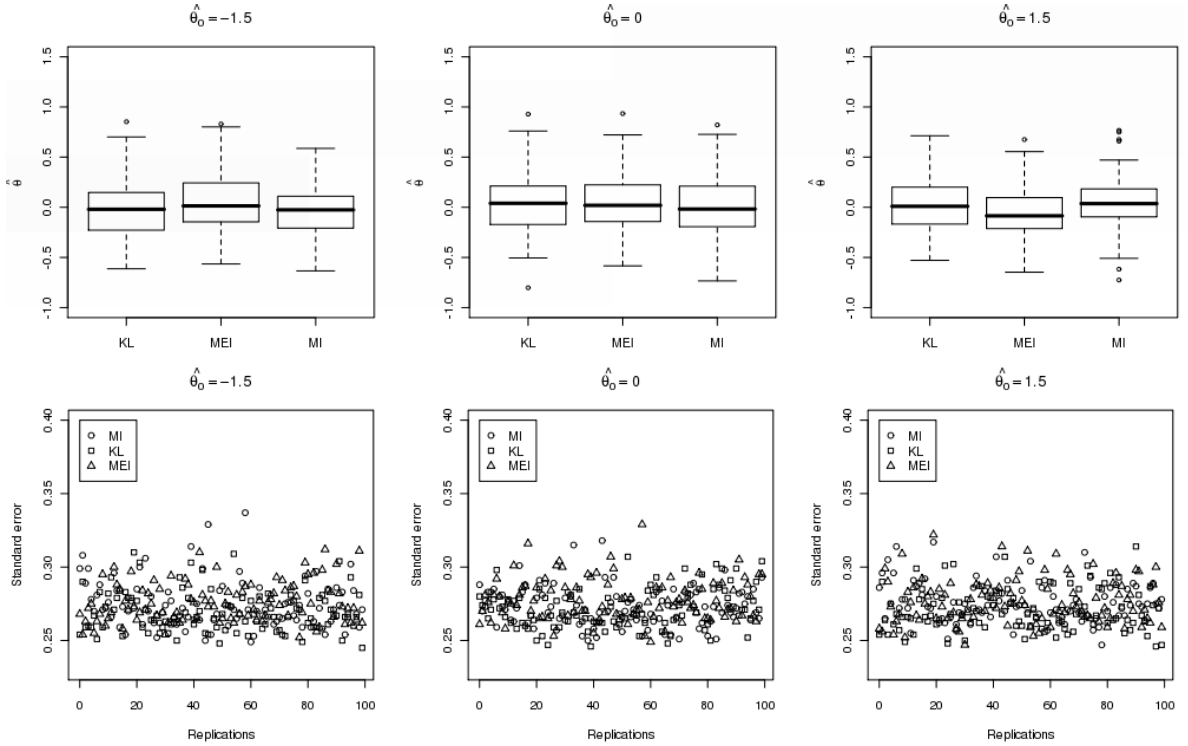
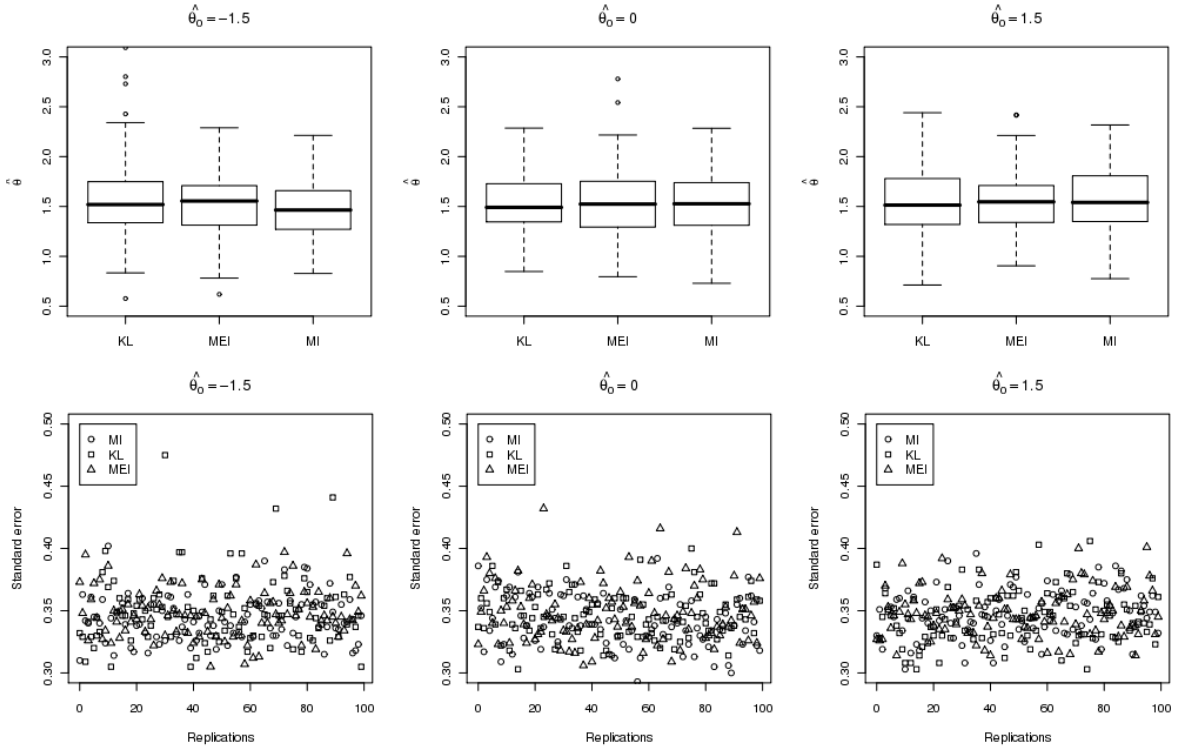


Figure 5. Distribution of Estimates When $\theta = 1.50$ for Different $\hat{\theta}_0$



In Figure 4 and 5 it can be observed that the procedures for adaptive selection efficiently estimated different levels of Instrumental English I proficiency, independent of the initial value attributed to the algorithm for estimation of the examinees' θs. All the simulations of this study verified that the performances of the three analyzed methods for item selection were very close when the maximum length of the test was fixed at 25.

Simulation Study 4

This study was designed to verify the quality of the θ estimates for different difficulty parameters for the first CAT item. Three different values for θ were used in the simulation: $\theta = -1.50$, $\theta = 0.00$ and $\theta = 1.50$. Each simulated θ was replicated 100 times. The maximum test length was 25 for all examinees.

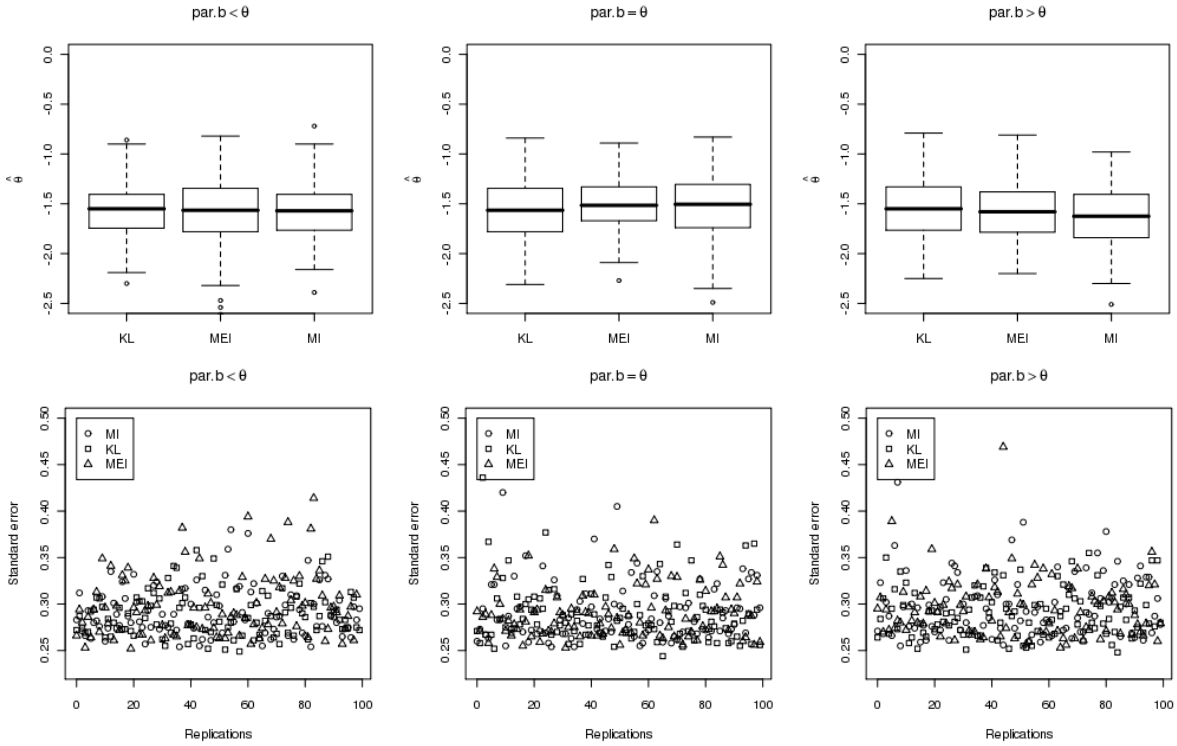
The initial θ value for the examinees was $\hat{\theta}_0 = 0.00$, and then three analyses were made: the first fixing the first item with a difficulty parameter much lower than the true value for θ, the second fixing the first item with a b parameter close to the true values, and the third fixing the first item with a b parameter above thetrue θ. The selected items in each analysis are presented in Table 4.

Table 4. Initial Item Parameters in Simulation 4

Analysis	Item	*a*	*b*	*c*
$b < \theta$	ING10836	0.80	-2.51	0.10
$b \approx \theta$	ING10706	0.60	-1.50	0.12
	ING20602	0.61	0.00	0.04
	ING10516	0.59	1.43	0.19
$> \theta$	ING20532	1.03	2.57	0.04

For $\theta = -1.50$, it was observed that both the estimates as well as the errors associated with the θs for each item selection method (MI, KL and MEI) performed similarly when a varying difficulty item was presented as the first question in the test. For any value of the first item's difficulty parameter, it was observed that the standard errors of the proficiencies were, on average, 0.29 after 25 items. Concerning the θs, it was also observed that for any of the initial analysis items, the selection methods presented values that were close to the true θs. The distributions of the θ estimates and of the SEs for $\theta = -1.50$ are presented in Figure 6.

Figure 6. Distribution of Estimates When θ = –1.50 for Different Initial Items



In relation to $\theta = 0.00$, it was also noted that both the estimates and the errors associated with each item selection method's θs had similar performance when a varying difficulty was presented as the first question in the test. The distribution of the θ estimates as well as of the SEs from this simulation are shown in Figure 7.

Figure 7. Distribution of θ Estimates When $\theta = 0.0$ for Different Initial Items

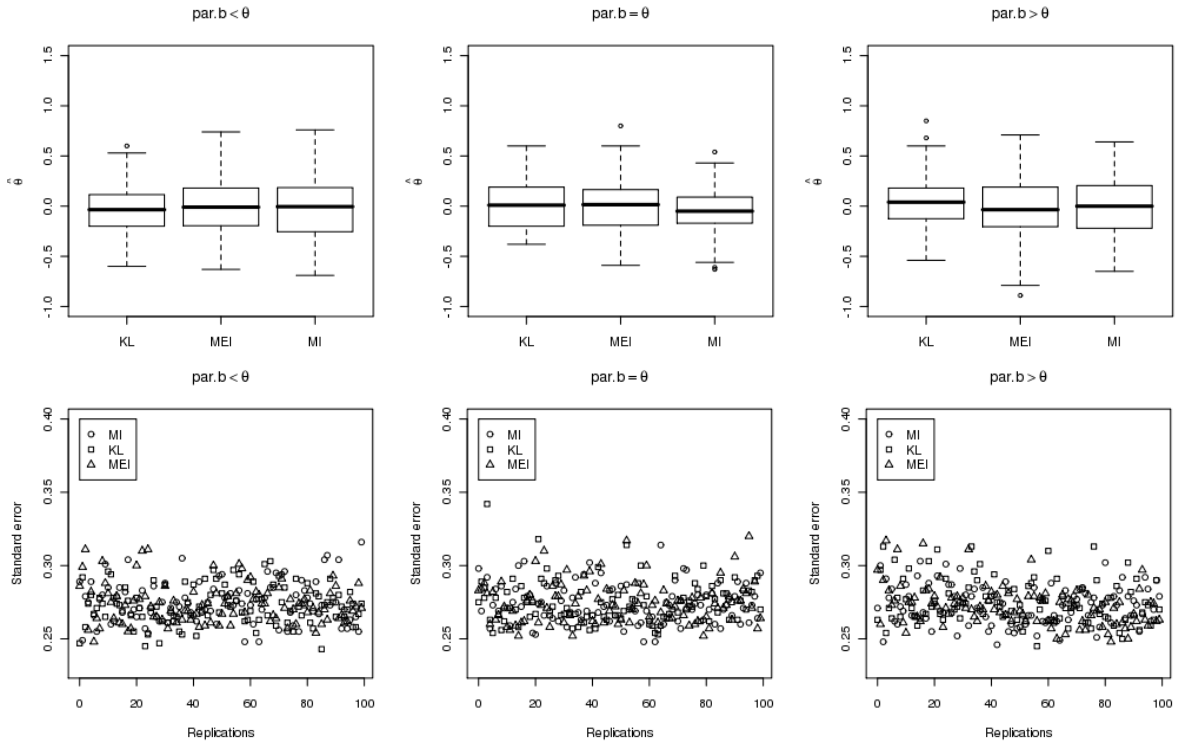
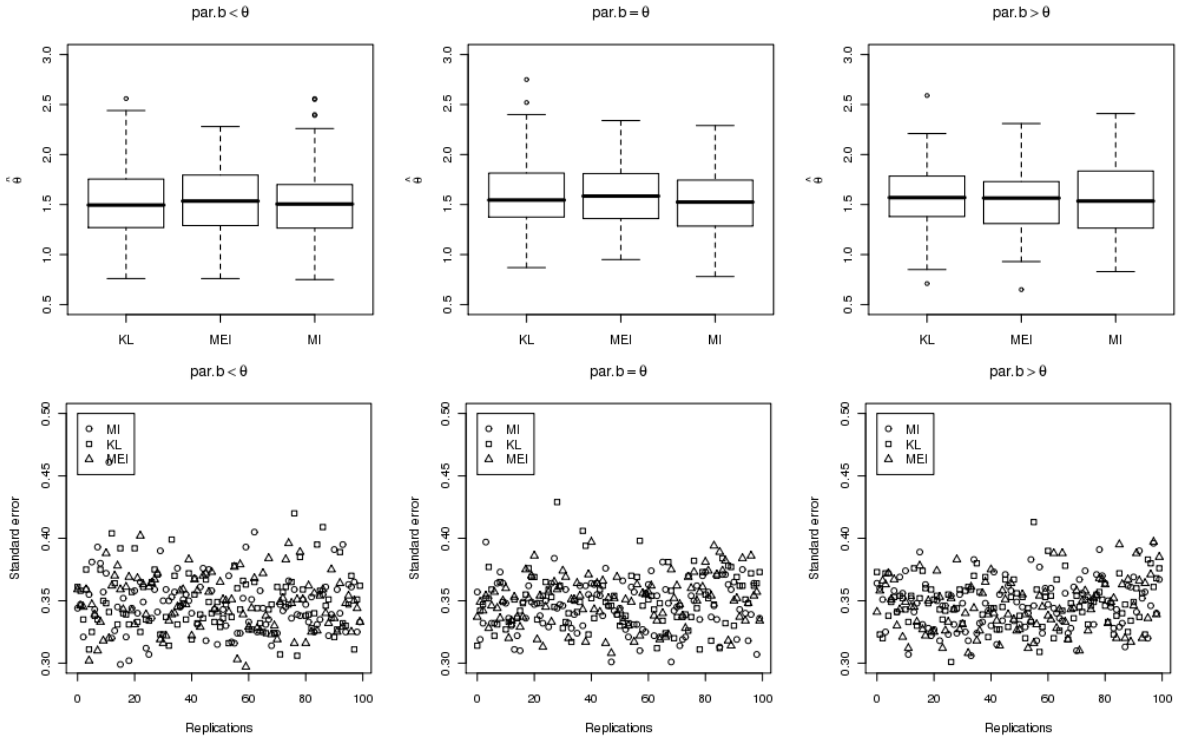


Figure 8 presents the distributions of the estimates and standard errors for the different item selection methods when $\theta = 1.0$. As in the previous cases, it can be concluded that, independent of the initial item's difficulty parameter, the three adaptive selection methods estimated θ on Instrumental English I with practically the same precision for a 25-item test.

Figure 8. Distribution of Estimates when θ = 1.50 for Different Initial Items



Simulation Study 5

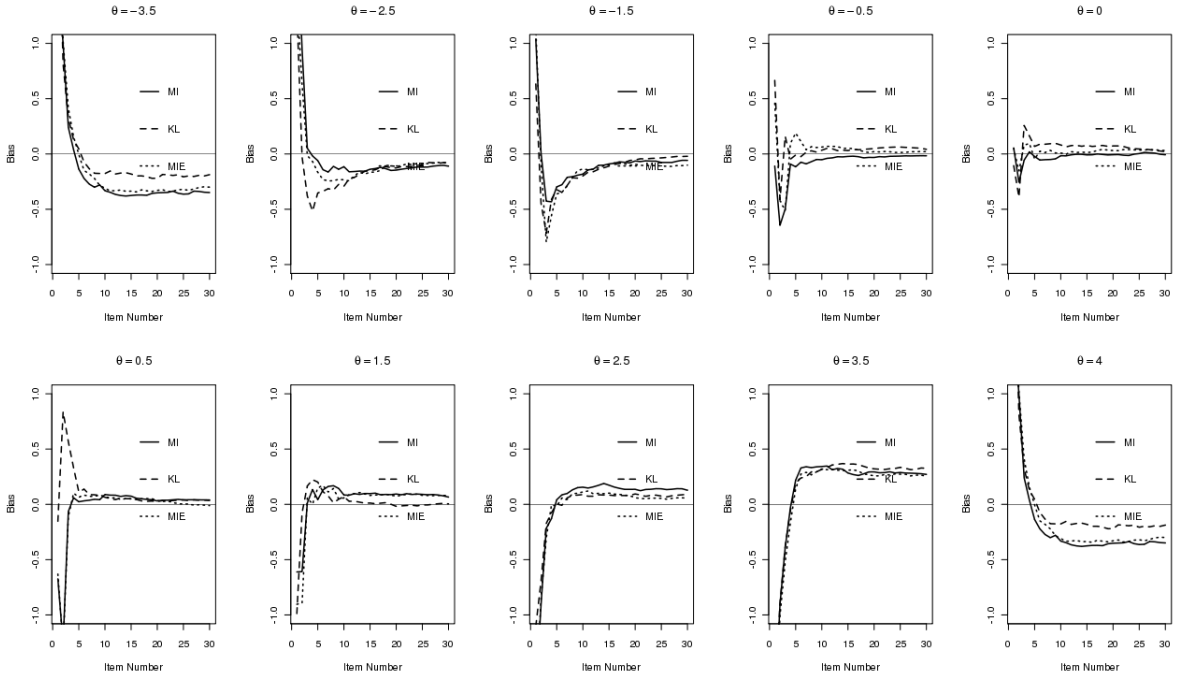
The simulation procedure used in this study was designed to assess measures for bias and mean square error (MSE) associated with 10 different θ values: $\theta = -3.50$, $\theta = -2.50$, $\theta = -1.50$, $\theta = -0.50$, $\theta = 0.00$, $\theta = 0.50$, $\theta = 1.50$, $\theta = 2.50$, $\theta = 3.50$ and $\theta = 4.00$. 100 replications of each θ were made. θs were estimated until 30 items were administered.

The bias and the MSE of the estimates were calculated for test lengths (n) were $1, 2, ..., 30$ items where

$$\text{Bias}_n = \frac{\sum_{j=1}^{100} \hat{\theta}_{j,n}}{100} - \theta \quad \text{and} \quad \text{MSE}_n = \frac{\sum_{j=1}^{100} (\hat{\theta}_{j,n} - \theta)^2}{100}, \quad \text{n} = \{1, \ldots, 30\}.$$

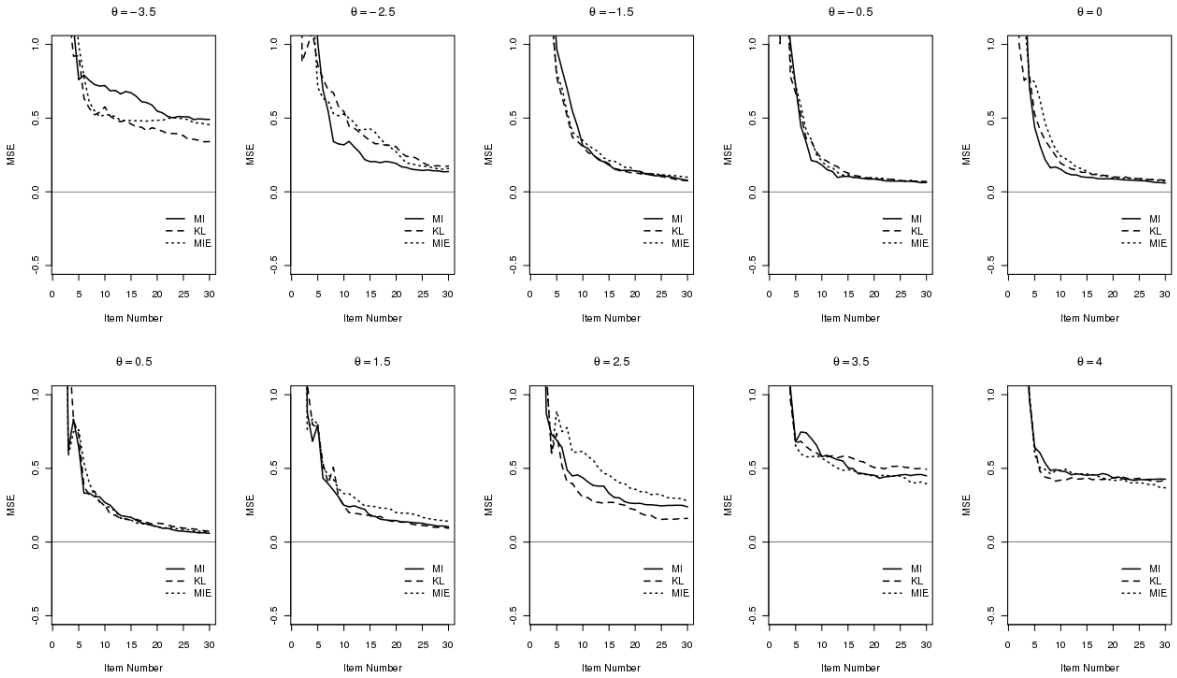
Figure 9 shows that bias decreased as test length increased for all three item selection methods.

Figure 9. Biases of the Estimates



As expected, the MSE of the estimates diminished as the number of items on the adaptive test increased, as shown in Figure 10.

Figure 10. Mean Square Errors of the Estimates



Conclusions and Future Research

The simulation studies showed that the three methods for adaptive item selection have similar performance when the item bank of Instrumental English I was used. As expected, these procedures reduced substantially the number of items in a test without compromising the precision of the theta estimates. It is also important to note that, as CAT is extremely sensitive to the item bank it uses, the item bank of Instrumental English I needs to be improved because θs outside of the interval (–2.5, 2.5) could not be estimated well enough.

Future research will refine the algorithms for item selection, incorporate constraints in the adaptive selection, such as item exposure control and testlets, evaluate the θ estimates using MCMC (Markov chain monte carlo) methods for estimating item parameters, and define other criteria to terminate CAT item selection .

References

Chang, H.H., Ying, Z. (1996). A global information approach to computerized adaptive testing. *Applied Psychological Measurement, 20,* 213.

R Development Core Team. R: A language and environment for statistical computing. *R Foundation for Statistical Computing*, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

van der Linden, W. (1998). Bayesian item selection criteria for adaptive testing. *Psychometrika*, 63, 201.

Wainer, H., Bradlow, E. T. and Du, Z. (2003). Testlet response theory: An analog for the 3PL model useful in testlet-based adaptive testing. In *Computerized Adaptive Testing: Theory and Practice*, (Eds. W. J. van der Linden & C. A. W. Glas). Netherlands: Kluwer Academic.