ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2019; 10(3); 315-326

Investigation of Item Selection Methods According to Test Termination Rules in CAT Applications *

Sema SULAK **   Hülya KELECİOĞLU ***

Abstract

In this research, computerized adaptive testing item selection methods were investigated in regard to ability estimation methods and test termination rules. For this purpose, an item pool including 250 items and 2000 people were simulated (M = 0, SD = 1). A total of thirty computerized adaptive testing (CAT) conditions were created according to item selection methods (Maximum Fisher Information, a-stratification, Likelihood Weight Information Criterion, Gradual Information Ratio, and Kullback-Leibler), ability estimation methods (Maximum Likelihood Estimation, Expected a Posteriori Distribution), and test termination rules (40 items, SE < .20 and SE < .40). According to the fixed test-length stopping rule, the SE values that were obtained by using the Maximum Likelihood Estimation method were found to be higher than the SE values that were obtained by using the Expected a Posteriori Distribution ability estimation method. When ability estimation was Maximum Likelihood, the highest SE value was obtained from a-stratification item selection method when the test length is smaller then 30. Whereas, Kullback-Leibler item selection method yielded the highest SE value when the test length is larger then 30. According to Expected a Posteriori ability estimation method, the highest SE value was obtained from a-stratification item selection method in all test lengths. In the conditions where test termination rule was SE < .20, and Maximum Likelihood Ability Estimation method was used, the lowest and highest average number of items were obtained from the Gradual Information Ratio and Maximum Fisher Information item selection method, respectively. Furthermore, when the SE is lower than .20 and Expected a Posteriori ability estimation method was utilized, the lowest average number of items was obtained through Kullback-Leibler, and the highest was obtained through Likelihood Weight Information Criterion item selection method. In the conditions where the test termination rule was SE < .40, and ability estimation method was Maximum Likelihood Estimation, the maximum and minimum number of items were obtained by using Maximum Fisher Information and Kullback-Leibler item selection methods respectively. Additionally, when Expected a Posteriori ability estimation was used, the maximum and minimum number of items were obtained via Maximum Fisher Information and a-stratification item selection methods. For the cases where the stopping rule was SE < .20 and SE < .40 and Maximum Likelihood Estimation method was used, the average number of items were found to be highest in all item selection methods.

Key Words: Computerized adaptive testing, maximum fisher information, a-stratification, likelihood weight information criterion, gradual information ratio, kullback-leibler.

INTRODUCTION

Computerized Adaptive Test (CAT) algorithm consists of applying selected items to the examinee in computer environment, estimating examinee ability level through given responses, selecting new items according to the most recent estimated ability, and administrating test until the specified test termination rule is conducted (Orcutt, 2002; Thissen & Mislevy, 2000; Wainer, 2000; Weiss, 1983).

The key questions for CAT are (Wainer, 2000);

- How is the first item selected to start the test?

* The present study is a part of PhD Thesis entitled “Bireyselleştirilmiş Bilgisayarlı Test Uygulamalarında Kullanılan Madde Seçme Yöntemlerinin Karşılaştırılması” completed within Hacettepe University Graduate School of Educational Sciences.

** Assist. Prof. PhD., Bartin University, Faculty of Education, Bartın-Turkey, semasulak@gmail.com, ORCID ID: 0000-0002-2849-321X
*** Prof. PhD., Hacettepe University, Faculty of Education, Ankara-Turkey, hulyakelecioglu@gmail.com, ORCID ID: 0000-0002-0741-9934

To cite this article:
Sulak, S. & Kelecioğlu, H. (2019). Investigation of item selection methods according to test termination rules in CAT applications. *Journal of Measurement and Evaluation in Education and Psychology, 10*(3), 315-326. doi: 10.21031/epod.530528

Received: 21.02.2019
Accepted: 06.07.2019

- How are the subsequent items selected from the item pool based on examinee responses, and how is the examinee ability predicted based on given responses?
- How is the test terminated?

There are different methods for selecting the first item to start testing. Either relevant information about examinees (i.e., previous test scores, grades, etc.) are used or a set of items, which do not impact examinees' final scores, are applied to all examinees to determine the first item. (Slater, 2001; Sireci, 2003). The most commonly used ability estimation methods in CAT applications are Maximum Likelihood and Bayesian Based Estimation. The major item selection methods used in CAT applications are Maximum Fisher Information (MFI), a-stratification, Likelihood Weight Information Criterion (LWIC), Gradual Information Ratio (GIR) and Kullback-Leibler (KL). The methods used in this study are explained below.

Maximum Fisher Information

The MFI item selection method aims to find the maximal interim ability to estimate regarding every previously administered item. MFI item selection investigate the i^th^ item that results in the largest value of,

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

In the Equation 1, a_i, b_i, and c_i; represent the discrimination, difficulty, and pseudo-guessing parameters in 3PLM respectively, and D stands for the scaling constant, 1.702. (Han, 2010).

Kullback-Leibler

The KL information selection method was developed by Chang and Ying (1996) based on the global knowledge approach. KL information for an item is defined as Equation 2.

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

KL information is a function of two variables (θ and θ_0) and is a surface in three-dimensional space. As a function of these two θ levels, KL information characterizes the change capacity of an item between two θ levels.

Likelihood Weight Information Criterion

LWIC item selection method was developed by Veerkamp and Berger (1997). In this method, the information function is collected along the θ scale and weighted by the likelihood function after the administration of the item.

The item to be selected in the LWIC criterion is determined by selecting the item that will maximize the value of the Equation 3.

$$\int_{\theta=-\infty}^{\infty} L(\theta;x_{m-1})I_i[\theta]d\theta \tag{3}$$

a-Stratification

The method of a-stratification item selection is constituted with the suggestion of layering according to the a parameter values in the item pool by Chang and Ying (1999). In this method, items are stratified into K strata based on their a values. Accordingly, the item selection process is divided into K stages. In the first stage, items are selected from the first stratum, which corresponds to the items with the lowest a values. In the second stage, items are selected from the second stratum. In the K^th^ stage, items are selected within the Kth level (Chang, Qian, & Ying, 2001). This method utilizes low a-items at early stages of the test. By doing so, the test precision and efficiency are maintained (Chang & Ying, 1996).

Gradual Information Ratio

The GIR item selection method was developed by Han (2009). Han proposed an alternative method based on the expected item effectiveness to improve the use of item pool instead of MFI method.

Han (2009) proposed to take the item efficacy (expected item information) into account on the item adequacy. Thus, this method looks for the item that makes the following criteria maximum,

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

In Equation 4, M shows the length of the test, and m denotes the number of administered items +1.

There are two test stopping methods in CAT applications; fixed-length tests and standard error termination (Sireci, 2003; Wainer, 2000; Weiss & Kingsbury, 1984). Fixed-length termination rules continue until an examinee takes a predetermined number of items. According to the standard error (SE) termination rule, the exam continues until the estimate of the θ reaches a certain level.

CAT applications have numerous advantages. The most important advantage provided by CAT applications is that the test can be tailored to the examinees' ability level. In order to obtain valid results from CAT applications, it is critical to select the item that maximizes the test information about the examinee. MFI is widely used in CAT applications; however, this method tends to use items with a high a parameter and is insufficient in the ability estimation at the beginning of the test (Van der Linden & Glas, 2010; Wainer, 2000; Weiss, 1983). Veldkamp (2012) stated that it is important to investigate different item selection methods in order to eliminate the aforementioned (proposed) limitations of MFI item selection method. There are researches indicating a-stratification item selection method is preferred to MFI due to selecting high a parameter items (Chang & Ying, 1999; Deng, Ansley, & Chang, 2010; Deng & Chang, 2001). Additionaly, Eggen (1999) found that KL item selection method provides more accurate ability estimation in comparison to MFI. Weissman (2003) stated in his study that ability estimation methods affect item selection methods. Bock and Mislevy (1982) indicated that Expected a Posteriori (EAP) ability estimation method was better than Maximum Likelihood Estimation (MLE) methods; while Wang and Visposel (1998) proposed that EAP ability estimation method was more biased. There are additional researches regarding the relationship between the test termination rules and item selection methods (Han, 2009; Weissman, 2003).

Purpose of the Study

The key point of the item selection process in the CAT applications is to match the ability of the respondent with the difficulty of the item. Namely, in CAT, ability estimation is reperformed after each item is answered, and the most recent ability estimation is used in the selection of subsequent items. MLE and EAP which are among the ability estimation methods were included in the research, and it was attempted to determine how ability estimation methods affect the item selection methods. There are studies suggesting that item selection methods are inadequate (especially when the test length is smaller than five items) at the beginning of CAT applications (Han, 2009; Linda, 1996; Van der Linden & Glas, 2010). According to the literature when the CAT has more than 20 items, the difference in the performance of a newly proposed method and MFI turns out to be trivial (Passos, Berger & Tan, 2007; as cited in Şahin & Özbaşı, 2017). Chen, Ankenmann and Chang (2000) conducted a simulation study to compare item selection methods, and they found that for CATs with more than 10 items, there is no difference between item selection methods. Veerkamp and Berger (1997) conducted a simulation study according to 60 items termination rule and found that item selection performances vary over 20 items. One of the advantages of CAT applications is to shorten the test. An item pool of 60 items was not selected, and an item pool of more than 20 items was used.

Thus, different test lengths (5, 10, 20, 30, and 40 items) were also taken as a variable to determine how the item selection methods differ depending on the test length. In order to compare the item selection methods in CAT applications where the test stopping rule was determined based on a fixed standard error, conditions were established in which the standard error was .20 and .40.

This study aims to answer the following questions:

1) How do standard errors in relation to the methods used in item selection (Maximum Fisher Information, a-stratification, Likelihood Weight Information Criterion, Gradual Information Ratio, and Kullback-Leibler) differ in terms of

   a) test length (5, 10, 20, 30 and 40 items)

   b) ability estimation (Maximum Likelihood and Expected a Posteriori) methods?

2) How do the average number of items utilized in item selection methods (Maximum Fisher Information, a-stratification, Likelihood Weight Information Criterion, Gradual Information Ratio, and Kullback-Leibler) differ in terms of

   a) test termination rules (SE < .20 and SE < .40)

   b) ability estimation methods (Maximum Likelihood and Expected a Posteriori)?

When the literature regarding the current study is reviewed, the following results are found:

In their study, Veerkamp and Berger (1997) compared the Interval Information Criteria and LWIC methods with MFI method, and the authors concluded that these methods did not have a substantial superiority to MFI. Eggen (1999) have compared KL and MFI item selection methods. According to the results of this study, KL item selection method performed better than the MFI. In a simulation study, Wen, Chang and Hau (2001) compared a-stratification item selection method and MFI item selection method. They concluded that MFI item selection method yielded more effective results than a-stratification item selection method. Weissman (2003) investigated the effectiveness of item selection methods in CAT applications. According to the findings, the ability estimation method impacted the effectiveness of item selection more than item selection method. Han (2009) explored random selective MFI, fade-away selective MFI, GIR, and fade-away selective GIR item selection methods in CAT application. It was concluded that MFI and GIR item selection methods exhibited lowest SE through theta criteria. Costa, Karino, Moura and Andrade (2009) evaluated the performance of MFI, KL, and Maximum Expected Information item selection methods. They concluded that all methods performed similarly to estimate examinees' θs by means of bias and mean square error.

Deng et al. (2010) compared MFI, a-stratification, and refined a-stratification item selection methods. The study findings yielded that MFI was more effective in predicting ability in comparison to other methods. Han (2010) compared five different item selection methods, which are a-stratification, Interval Information Criteria, Likelihood Weighted Information Criterion (LWIC), Kullback-Leibler Information, and Gradual Information Ratio (GIR). The study results showed that SE values decreased in all item selection methods due to test length. Low SE values were calculated under MFI, KL and GIR item selection methods, whereas high SE values were calculated under a-stratification item selection methods.

Research findings related to different item selection methods in the literature indicated that item selection methods have strengths as well as weaknesses in different conditions (Deng et al., 2010; Eggen, 2009; Wen, et al., 2001; Yi & Chang, 2003) and two-item selection method were compared. In the studies investigating more than two item selection methods (Han, 2010; Weissman, 2003), stopping rules and ability estimation methods were not elaborated together.

METHOD

The data of the study were simulated by SimulCAT computer program, which was developed by Han (2012). In data collection stages, first, the group where the research was to be carried out, then the item pool and CAT conditions were formed.

Participants

2000 hypothetical examinee were simulated. Examinee ability parameters (N = 2000) were randomly drawn from a normal distribution ~N(0, 1). Dichotomous item responses for the entire item bank were generated using the SimulCAT program (Han, 2012).

Data Collection Instruments

Item pool

An item pool with 250 dichotomously-scored items was created using the three-parameter logistic (3PL) item response model. In his research, Urry (1977) found that an item pool of at least 100 items is adequate to estimate ability. Kingsbury and Zara (1989) indicated that item pool size for adaptive tests should always be -more is better-. Stocking (1992) determined that an item pool size should be 6 to 12 times more than the item number.

Item discrimination parameters were randomly drawn from a uniform distribution ~U(0.8, 1.5); item difficulty parameters were randomly drawn a uniform distribution ~U(-3, 3); guessing parameters were randomly drawn from a uniform distribution ~U(.05, .15). Following the suggestions from previous studies regarding data simulation for the 3PL model, the simulation was conducted. Feinberg and Rubright (2016) indicated 3PL IRT model parameters are often simulated as uniform. Ree and Jensen (1983) said that “a values below 0.5 are insufficiently discriminating for most testing purposes, and a values above 2.0 are infrequently found … most test items have c parameters less than or equal to .30” (pp. 135-146).

Process

The data collection process was simulated using the SimulCAT computer program. As the first step, examinee and item pool files were created and uploaded to the computer program. In the second step, item selection and stopping rules were specified, and in the final step, ability estimation methods, test initiation rule, number of replications and output files were selected. The test initiation rule was determined as θ = 0.5, and 100 replications were performed for all simulation conditions. A crossed-factorial design resulted in a total of 30 simulation conditions; 5 item selection methods * 2 ability estimation methods * 3 stopping rules. For each crossed condition, 100 replications were conducted. The number of replications depends on the research question. However, with too many replications simulation may be more complex and might take a long time to complete (Bulut & Önder, 2017; Feinberg & Rubright, 2016). Because of the 30 conditions, the researcher decided to make 100 replications. Harwell, Stone, Hsu and Kirisci (1996) suggested a minimum of 25 replications and indicated that “aggregating results over replications produces more stable and reliable results” (p. 110). Thus, the simulation study was ended after 100 replications and interim, and final θ values were aggregated over the 100 replications.

Data Analysis

In order to determine how item selection methods differ according to the test length in the CAT conditions, where the stopping rule was specified as 40 items, interim θ and standard error (SE) of the estimation were calculated for 5, 10, 20, 30 and 40 items. The standard error of the estimation is calculated via the Equation 5.

$$\mathrm{SE}(\hat{\theta}) = \frac{1}{\sqrt{\mathrm{I}(\theta)}} \tag{5}$$

In the conditions with test stopping rule of SE < .20 and SE < .40, item selection methods were evaluated according to the average number of items. Since CAT administration would terminate at a specific standard error value, the average number of items used until reaching this standard error value was investigated.

RESULTS

To determine how standard error associated with different item selection methods (MFI, GIR, LWIC, a-stratification, KL), the test length (5, 10, 20, 30 and 40 items) and ability estimation methods (MLE and EAP), mean of the interim ability estimations ($\hat{\theta}$) were used in the analysis of the results. Item selection methods were compared according to SE values, and the results are presented Table 1.

Table 1. Statistics Regarding the Item Selection Methods According to the Test Length (For a 40-Item Fixed-Length CAT Administration Where MLE Ability Estimation is Used)

	Test Length									
Item Selection	5		10		20		30		40	
Methods	$\hat{\theta}$	SE	$\hat{\theta}$	SE	$\hat{\theta}$	SE	$\hat{\theta}$	SE	$\hat{\theta}$	SE
MFI	0.12	.55	0.05	.36	0.03	.25	0.02	.20	0.02	.18
a-stratification	-1.55	.78	-1.57	.52	-1.39	.31	-1.29	.23	-1.19	.19
LWIC	-0.60	.74	-0.28	.38	-0.11	.25	-0.62	.21	-0.04	.18
GIR	-1.52	.50	-1.28	.35	-1.26	.25	-1.06	.21	-0.68	.19
KL	-1.6	.67	-1.20	.37	-1.10	.25	-0.57	.22	-0.21	.22

When Table 1 is examined, it is observed that the method of a-stratification item selection shows high SE value in cases where the test length is less than 30 items (n < 30), while the method of KL item selection shows high SE value in cases where the test length is greater than thirty items (n > 30). While the highest SE value that was obtained from the a-stratification item selection method is similar to the results of Han’s (2009) research, it differs from Linda’s (1996) study which shows that KL item selection method is better than the MFI item selection method.

Considering all item selection methods according to test lengths, it was determined that there was a great difference between the SE values of the item selection methods after administering five items. However, the difference between SE values was decreased after administering ten items. When the inadequacy of MFI item selection method in the predictive estimation at the beginning of the CAT applications (n < 5) was examined, it was found that only the GIR item selection method showed a lower SE value than MFI. These two findings indicated that all of the item selection methods included in this study were limited in their ability estimation at the beginning of CAT applications and that they did not have a significant advantage over MFI item selection method.

Table 2. Statistics on the Methods of Item Selection According to the Test Length in the CAT Conditions Where the Test Stopping Rule is Determined as 40 Items and the EAP Ability Estimation is Used

	Test Length									
Item Selection	5		10		20		30		40	
Method	$\hat{\theta}$	SE	$\hat{\theta}$	SE	$\hat{\theta}$	SE	$\hat{\theta}$	SE	$\hat{\theta}$	SE
MFI	0.01	.47	0.02	.33	0.02	.23	0.02	.20	0.02	.18
a-stratification	0.01	.70	0.01	.49	0.02	.31	0.02	.23	0.02	.18
LWIC	0.01	.55	0.02	.35	0.02	.24	0.02	.20	0.02	.18
GIR	0.01	.49	0.01	.33	0.02	.24	0.02	.20	0.02	.18
KL	0.01	0.47	0.02	0.33	0.02	0.24	0.02	0.20	0.02	0.18

As shown in Table 2, a-stratification item selection method had the highest SE value among all test lengths. When the findings were examined, it was found that there was a substantial difference between the SE values of the item selection methods while the test length was 5 items, but the difference between the SE values was decreased in the CAT conditions where the test length was specified as 10 items and higher. The differences decreased as test length increased, and the results were close to each other. In addition, when the test length reached 40 items, the SE values of the item selection methods were found equal to each other. This significant decrease in all item selection methods at the beginning of the CAT applications (n < 5) was interpreted as the absence of a significant superiority of other item selection methods except for the KL item selection method in the problem of MFI item selection method in terms of ability estimation.

When MLE and EAP ability estimation methods were examined, the highest SE value was obtained from the a-stratification item selection method in both MLE and EAP ability estimation methods. In general, the SE values obtained when the MLE ability estimation was used were found higher than the SE values obtained when EAP ability estimation was used.

The most important difference was detected when the test length was 5 items. For example, the SE value of the KL item selection method was .67 for MLE ability estimation, whereas the SE value was calculated as .47 for EAP ability estimation. Wang and Visposel (1998) found that EAP ability estimation showed a lower SE value compared to MLE ability estimation method.

The findings obtained in the present study align with these results. This finding may indicate that EAP ability estimation method should be primarily preferred especially at the beginning of the test in the application of CAT. In both cases where MLE and EAP ability estimation were used, a sharp decrease in SE values was observed when the test length reached to 10 items from 5 items.

To be able to determine how the average number of items related to item selection methods (MFB, GIR, LWIC, a-stratification, KL) changes according to test stopping rule (SE < .20 and SE < .40) and ability estimation methods (MLE and EAP), the mean number of items was calculated. The findings were presented in Table 3.

Table 3. Statistics of Ability Estimation and Item Selection Methods in CAT Conditions Where the Test Stopping Rule is Based on Fixed Standard Error

	Stopping rule						
Ability Estimation Method	**Item Selection Method**	**SE < .20**			**SE < .40**		
		Minimum Item	**Maximum Item**	**Average Item**	**Minimum Item**	**Maximum Item**	**Average Item**
MLE	MFI	26	95	40.71	7	9	8.72
	a-stratification	-	-	-	13	16	14.65
	LWIC	27	88	32.85	8	13	9.54
	GIR	12	41	31.75	7	10	8.96
	KL	13	38	32.63	8	12	9.72
EAP	MFI	18	124	30.07	6	11	7.07
	a-stratification	-	-	-	12	17	12.54
	LWIC	26	78	31.18	8	9	8.41
	GIR	18	43	30.23	7	12	7.46
	KL	27	48	30.13	6	11	7.16

According to the results on Table 3, the lowest and highest number of items were obtained from GIR and MFI item selection methods respectively in the CAT conditions where the standard error was less than .20 and the MLE ability estimation was used. In the CAT applications where EAP ability estimation was used, the average of the lowest and highest number of items was obtained from KL and LWIC item selection methods. The a-stratification item selection method did not function as expected in both MLE and EAP ability estimates. The computer program could not complete the simulation because no suitable item was found in the item pool. This situation was interpreted as the insufficiency of the item pool or the small size of the a-parameter range. In the method of a-stratification item selection, the item pool is stratified according to the a parameters, and in the present research, the item pool is divided into three layers. In the literature, studies have been carried out for the various size of item pools.

Wen et al. (2001) determined four layers for an item pool of 360 items and a-stratification item selection method in their research where the parameter value ranges from 0.40 to 2. On the other hand, Costa et al. (2009) were able to use the a-stratification item selection method for a standard error value of .20 using a pool of 246 items. When the existing research was examined, it was considered that keeping a parameter value between 0.80 and 1.5 could be the reason why a-stratification method has not been realized under the condition that the standard error is less than .20 as well as the effect of item pool size.

The average number of items was examined for each ability estimation methods. The mean number of items obtained from CAT conditions using MLE ability estimation was found to be higher than the mean number of items from CAT conditions using EAP ability estimation.

This was interpreted as the ability to estimate EAP ability to obtain shorter tests in CAT applications. In CAT conditions test stopping rule, where standard error is defined as less than .40 and MLE ability estimation is used, the lowest and highest number of items were obtained from MFI and a-stratification item selection methods, respectively regarding the mean number of items. In CAT conditions using EAP ability estimation, MFI and KL item selection method had the lowest value while the method of a-stratification item selection was found to be the highest in terms of the average number of items.

The lowest test length was obtained from MFI, and the highest test length was obtained from a-stratification item selection method in cases where both of the ability estimation methods were used. The a-stratification item selection method requires the highest number of items to achieve the standard error value of .40 may be related that this method selects items by stratification of the item pool.

DISCUSSION and CONCLUSION

In the beginning of CAT conditions, where MLE ability estimation method used, the lowest SE value was obtained from the GIR item selection method after five items administered (n < 5). a-stratification item selection method showed the highest SE value while the test length is shorter than 30 items (n < 30), and KL showed the highest SE value while the test length is longer than 30 items (n > 30). In the beginning of CAT conditions, where MLE ability estimation method used and the number of items was less than 10 (n < 10), it was seen that there was a great difference between the SE values of the item selection methods investigated, but this difference decreased as the test length increased.

When using EAP ability estimation, the highest SE values were obtained from a-stratification item selection method for all different test lengths included in the study. At the beginning of CAT conditions where EAP ability estimation method used and the number of items was less than 10 (n < 10), it was seen that there was a great difference between the SE values of the item selection methods investigated, but this difference decreased as the test length increased. When the test length was set to 40 items, the SE values of all the item selection methods yielded equal results. The SE values observed when MLE ability estimation was used were found to be higher than the SE values obtained when EAP ability estimation was used.

The lowest item number was obtained from GIR item selection method, and the highest item number was obtained from MFI item selection method when MLE ability estimation was used in the CAT conditions where SE was accepted as SE < .20. When EAP ability estimation is used, the lowest mean of the item number is obtained from KL item selection method, and the highest mean of the item number is obtained from the item selection method. In both cases where MLE and EAP ability estimations were used, a-stratification item selection method did not yield meaningful results. It was concluded that this finding was due to insufficient pool size and low level of the parameter value. When the average of the number of items was examined in terms of ability estimation method, it was concluded that the conditions in which MLE ability estimation was used were higher than those in which EAP ability estimation is used.

When MLE ability estimation was used in the CAT conditions where SE < .40 was used, the lowest average of item number was obtained from MFI item selection method, and the highest average of item number was obtained from KL item selection method. When EAP ability estimation was used, the lowest average of item number was obtained from MFI and KL item selection methods, and the highest average of item number was obtained from a-stratification item selection method. For all of the item selection methods included in the study, the average test length obtained from MLE ability estimation was higher than the average test length obtained from EAP ability estimation. It was concluded that EAP ability estimation shorten the test length. SE values for item selection methods were lower when EAP ability estimation was used. EAP ability estimation is recommended for operational CAT applications. One of the most important advantages of CAT applications is that it produces a shorter test length than paper-based tests. When the results are investigated, it is recommended that EAP ability estimation method can be preferred in CAT applications.

The method of a-stratification item selection did not yield meaningful result in the condition that the test stop rule was SE < .20. This finding shows that further research is needed. It is recommended that future studies may be conducted by determining different item pool sizes and a-parameter values. In addition, the relationship between the number of layers used in the method of a-stratification item selection method may be studied.

Future studies should be carried out to investigate what would happen if there were more constraints placed on the items in the pool, such as, content constraints which may differ how the item pool is conducted. Also, the effect of b parameter value (b-blocking, etc.) on item selection methods can be investigated. In this research, a parameter value range is narrow, and this research can be repeated according to different a parameter range. Different item pool sizes and ability estimation methods can be examined for the same simulative conditions of research. How different item selection methods work in an item pool weighted according to content can be examined. In this study, one-dimensional item response theory is used, in the future studies multi-dimensional item response theory can be used. The present study has been done on the simulation data, and the operational CAT applications can be investigated in future studies.

REFERENCES

Bock, R. D., & Mislevy, R. J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement*, *6*(4), 431-444. doi: 10.1177/014662168200600405

Bulut, O., & Sünbül, Ö. (2017). R programlama dili ile madde tepki kuramında monte carlo simülasyon çalışmaları. *Journal of Measurement and Evaluation in Education and Psychology*, *8*(3), 266-287. doi: 10.21031/epod.305821

Chang, H.-H, Qian, J., & Ying, Z. (2001). a-stratified multistage adaptive testing with b blocking. *Applied Psychological Measurement, 25*(4), 333-341. doi: 10.1177/01466210122032181

Chang, H.-H, & Ying, Z. (1996). A global information approach to computerized adaptive testing. *Applied Psychological Measurement, 20*(3), 213-229. doi: 10.1177/014662169602000303

Chang, H. H, & Ying, Z. (1999). a-stratified multistage computerized adaptive testing. *Applied Psychological Measurement, 23*(3), 211-222. doi: 10.1177/01466219922031338

Chen, S.-Y., Ankenmann, R. D., & Chang, H. H. (2000). A comparison of item selection rules at the early stages of computerized adaptive testing. *Applied Psychological Measurement*, *24*(3), 241-255. doi: 10.1177/01466210022031705

Costa, D., Karino, C., Moura, F., & Andrade, D. (2009, June). *A comparision of three methods of item selection for computerized adaptive testing*. Paper session presented at the meeting of 2009 GMAC Conference on Computerized Adaptive Testing. Retrieved from www.psych.umn.edu/psylabs/CATCentral/

Deng, H., & Chang, H. H. (2001, April). *A-stratified computerized adaptive testing with unequal item exposure across strata*. Paper session presented at the American Educational Research Association Annual Meeting 2001. Retrieved from https://www.learntechlib.org/p/93050/

Deng, H., Ansley, T., & Chang, H. H. (2010). Stratified and maximum information item selection procedures in computer adaptive testing. *Journal of Educational Measurement*, *47*(2), 202-226. doi: 10.1111/j.1745-3984.2010.00109.x

Eggen, T. H. J. M. (1999). Item selection in adaptive testing with the sequential probability ratio test. *Applied Psychological Measurement*, *23*(3), 249-261. doi: 10.1177/01466219922031365

Feinberg, R. A., & Rubright, J. D. (2016). Conducting simulation studies in psychometrics. *Educational Measurement: Issues and Practice, 35*(2), 36-49. doi: 10.1111/emip.12111

Han, K. (2009). *Gradual maximum information ratio approach to item selection in computerized adaptive testing* (Graduate Management Admission Council Research Reports No. 09-07). USA.

Han, K. (2010). *Comparision of non-fisher information item selection criteria in fixed length computerized adaptive testing*. Paper session presented at the annual meeting of the National Council on Measurement in Education. Denver, CO.

Han, K. (2012). SimulCAT: Windows software for simulating computerized adaptive test administration. *Applied Psychological Measurement*, *36*(1), 64-66. doi: 10.1177/0146621611414407

Harwell, M. R., Stone, C. A., Hsu, T., & Kirisci, L. (1996). Monte Carlo studies in item response theory. *Applied Psychological Measurement*, *20*(2), 101-125. doi: 10.1177/014662169602000201

Kingsbury, G. G., Zara, A. R. (1989). Procedures for selecting ıtems for computerized adaptive tests. *Applied Measurement in Education*, *2*(4), 359-375.

Linda, T. (1996, April). *A comparision of the traditional maximum information method and the global information method in cat item selection*. Paper session at annual meeting of the National Council on Measurement in Education, New York, NY.

Orcutt, V. L. (2002, February). *Computerized adaptive testing: Some issues in development.* Paper session at the annual meeting of the Educational Research Exchange. Denton, TX.

Ree, M. J., & Jensen, H. E. (1983). Effects of sample size on linear equating of item characteristic curve parameters, In Weiss, D. (Ed). *New horizons in testing latent trait test theory and computerized adaptive testing* (pp.135-146). London: Academic Press. doi: 10.1016/B978-0-12-742780-5.50017-2

Şahin, A., & Özbaşı, D. (2017). Effects of content balancing and item selection method on ability estimation in computerized adaptive testing. *Eurasian Journal of Educational Research*, *17*(69), 21-36. Retrieved from http://dergipark.org.tr/ejer/issue/42462/511414

Sireci, S. (2003). Computerized adaptive testing: An introduction. In Wall, & Walz (Eds), *Measuring up: Assessment issues for teachers, counselors and administrators* (pp. 684-694). USA: CAPS Press.

Slater, S. C. (2001). *Pretest ıtem calibration within the computerized adaptive testing environment* (Unpublished doctoral dissertation, Graduate School of the University Massachusetts). Retrieved from https://elibrary.ru/item.asp?id=5337539. Amherst.

Stocking, M. L. (1992). *Controlling item exposure rates in a realistic adaptive testing paradigm* (Research Report No. 93-2). Princeton, NJ: Educational Testing Service. doi: 10.1002/j.2333-8504.1993.tb01513.x

Thissen, D., & Mislevy, R. J. (2000). Testing algorithms. In H. Wainer, (Ed.), *Computerized adaptive testing: A primer* (pp. 101-133). Mahwah, NH: Lawrence Erlbaum Associates, Inc.

Urry, V. W. (1977). Tailored testing: A successful application of latent trait theory. *Journal of Educational Measurement, 14*(2), 181-196.

Van Der Linden, W. J., & Glas, C. A. W. (2010). *Elements of adaptive testing, statistics for social and behaviorel sciences*, New York, NY: Springer.

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. doi: 10.3102/10769986022002203

Veldkamp, B. P. (2012). Ensuring the future of computerized adaptive testing. In T. J. H. M. Eggen & B. P. Veldkamp (Eds.). *Psychometrics in practice at RCEC* (pp. 35-46). Netherlands: RCEC, Cito.

Wainer, H. (Ed.) (2000). *Computerized adaptive testing: A primer*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Wang, T., & Visposel, W. (1998). Properties of ability estimation methods in computerized adaptive testing. *Journal of Educational Measurement*, *35*(2), 109-135.

Weiss, D. J. (1983). Latent trait theory and adaptive testing. In D. J. Weiss (Ed.). *New horizons in testing: latent trait test theory and computerized adaptive testing* (pp. 5-7). New York, NY: Academic Press.

Weiss, D. J., & Kingsbury, G. G. (1984). Application of computerized adaptive testing to educational problems. *Journal of Educational Measurement*, *21*(4), 361-375. doi: 10.1111/j.1745-3984.1984.tb01040.x

Weissman, A. (2003, April). *Assessing the efficiency of item selection in computerized adaptive testing*. Paper session presented at the annual meeting of the American Educational Research Association. Chicago, IL.

Wen, H., Chang, H., Hau, K. (2001, April). *Adaption of a-stratified method in variable length computerized adaptive testing*. Paper session at the American Educational Research Association Annual Meeting, Seattle, WA.

Yi, Q., Chang, H. (2003). a-stratified CAT design with content blocking. *British Journal of Mathematical and Statistical Psychology*, *56*(2), 359-378. doi: 10.1348/000711003770480084

Bireyselleştirilmiş Bilgisayarlı Test Uygulamalarında Madde Seçme Yöntemlerinin Test Durdurma Kurallarına Göre İncelenmesi

Giriş

Bireyselleştirilmiş Bilgisayarlı Test (BBT) algoritması, seçilen maddelerin bilgisayar ortamında cevaplayıcıya sunulması, verilen cevaplar yoluyla yetenek düzeyinin kestirilmesi, hesaplanan yetenek düzeyine göre yeni maddelerin seçilmesi ve testin durdurma kuralı yerine gelinceye kadar test etme sürecine devam edilmesine göre gerçekleşir (Orcutt, 2002; Thissen & Mislevy, 2000; Wainer, 2000; Weiss, 1983).

Teste başlamak için ilk maddenin seçilmesinde farklı yöntemler vardır. Cevaplayıcı hakkında önceden sahip olunan bilgi (önceki testlerden aldığı puanlar, karne notu vb.) veya BBT uygulamalarına başlamadan önce cevaplayıcıların nihai test puanlarına etki etmeyecek madde setleri, tüm cevaplayıcılara uygulanır ve elde edilen yetenek düzeyi ilk maddenin seçilmesinde kullanılabilir (Sireci, 2003; Slater, 2001). BBT uygulamalarında yaygın olarak kullanılan yetenek kestirim yöntemleri, En Çok Olabilirlik ve Bayes kestirimine dayalı olan yöntemlerdir. BBT uygulamalarında kullanılan belli başlı madde seçme yöntemleri ise, Maksimum Fisher Bilgisi (MFB), Kullback-Leibler Bilgisi (KL), Aralık Bilgisi Ölçütü (ABÖ), Olabilirlik Ağırlıklı Bilgi Ölçütü (OAB), a-tabakalama, Aşamalı Maksimum Bilgi Oranıdır (AMBO). BBT uygulamalarında testi durdurmak için; sabit test uzunluğu ve değişken test uzunluğu olmak üzere iki yöntem vardır (Sireci, 2003; Wainer, 2000; Weiss & Kingsbury, 1984). BBT uygulamalarında MFB yaygın olarak kullanılır; ancak, bu yöntem yüksek a parametresine sahip maddeleri kullanmaya meyillidir ve özellikle testin başlangıcında yetenek kestiriminde yetersiz kalmaktadır (Van der Linden & Glas,2010; Wainer,2000; Weiss, 1984). Bu araştırmada, MFB madde seçme yönteminin yüksek a parametresine sahip maddeleri seçme özelliğinin farklı madde seçme yöntemleri ile karşılaştırılması yapılmıştır.

BBT uygulamalarında madde seçme sürecinin anahtar noktası, cevaplayıcının yeteneği ile madde güçlüğünü eşleştirmektir. Şöyle ki; BBT uygulamalarında her madde cevaplandıktan sonra yetenek kestirimi yapılmaktadır ve bu yetenek kestiriminin sonucu madde seçiminde kullanılmaktadır. Yetenek kestirim yöntemlerinden En Çok Olabilirlik Tahmini (EOT) ve Beklenen Sonsal Dağılım (BSD) araştırmaya dahil edilerek madde seçme yöntemlerini nasıl etkilediği belirlenmeye çalışılmıştır. BBT uygulamalarının başında (özellikle test uzunluğu beş maddeden küçük olduğunda) madde seçme yöntemlerinin yetersiz kaldığı yönünde araştırmalar mevcuttur. Test uzunluğuna bağlı olarak madde seçme yöntemlerinin nasıl farklılaştığını belirlemek için farklı test uzunlukları (5, 10, 20, 30 ve 40 madde) da bir değişken olarak alınmıştır. Testi durdurma kuralının sabit standart hataya bağlı olarak belirlendiği BBT uygulamalarında madde seçme yöntemlerini karşılaştırmak için ise, standart hatanın .20 ve .40 olduğu koşullar oluşturulmuştur. Eldeki araştırmanın amacı yetenek kestirim yöntemi, sabit madde sayısı ve standart hataya dayalı durdurma kuralının madde seçme yöntemlerini nasıl etkilediğini belirlemektir.

Yöntem

Bu araştırma simülatif olarak gerçekleştirilmiştir. 250 maddelik bir madde havuzu, ortalaması 0 ve standart sapması 1 olacak şekilde normal dağılım gösteren 2000 kişi simülatif olarak oluşturulmuştur. BBT koşulları, madde seçme yöntemleri (MFB, KL, OAB, a-tabakalama, AMBO), yetenek kestirim yöntemleri (EOT, BSD) test durdurma kuralları (40 madde, SH < .20 ve SH < .40) olmak üzere toplam otuz koşuldan oluşturulmuştur. Test durdurma kuralı 40 madde olarak belirlenen BBT koşullarında, test uzunluğuna göre madde seçme yöntemlerinin nasıl farklılaştığını bulmak amacıyla interim θ ve tahminin standart hatası (SH) hesaplanmıştır. Test durdurma kuralı SH < .20 ve SH < .40 olan BBT koşullarında, madde seçme yöntemleri, madde sayısına göre değerlendirilmiştir.

Sonuç ve Tartışma

Test uzunluğu 5, 10, 20, 30 ve 40 madde olarak belirlendiği ve EOT yetenek kestiriminin kullanıldığı BBT koşullarında; ilk beş madde kullanıldıktan sonra (n < 5) en düşük SH değeri AMBO madde seçme yönteminden elde edilmiştir. Test uzunluğu n < 30 iken, a-tabakalama; n > 30 iken KL madde seçme yöntemi en yüksek SH değerini göstermiştir. BBT koşullarının başında (n < 10), araştırmaya alınan madde seçme yöntemlerinin SH değerleri arasında büyük farklar olduğu, ancak test uzunluğu arttıkça bu farkın azaldığı görülmüştür. BSD yetenek kestirimi kullanıldığında ise; araştırmaya alınan bütün farklı test uzunluklarında en yüksek SH değeri a-tabakalama madde seçme yönteminden elde edilmiştir.

Test uzunluğu 40 madde olduğunda bütün madde seçme yöntemlerinin SH değerleri birbirine eşit sonuçlar vermiştir. EOT yetenek kestirimi kullanıldığında elde edilen SH değerleri, BSD yetenek kestirimi kullanıldığında elde edilen SH değerlerinden daha yüksek bulunmuştur.

SH < .20 olduğu BBT koşullarında EOT yetenek kestirimi kullanıldığında en düşük madde sayısı ortalaması AMBO madde seçme yönteminden, en yüksek madde sayısı ortalaması MFB madde seçme yönteminden elde edilmiştir. EOT ve BSD yetenek kestirimlerinin kullanıldığı her iki durumda da a-tabakalama madde seçme yöntemi sonuç vermemiştir. Bu durumun madde havuzu büyüklüğünün yetersiz kalmasından ve araştırmaya alınan a parametre değeri ranjının düşük olmasından kaynaklandığı sonucuna varılmıştır. Madde sayısı ortalamaları, yetenek kestirim yöntemleri bakımından incelendiğinde; EOT yetenek kestiriminin kullanıldığı koşulların, BSD yetenek kestiriminin kullanıldığı koşullardan daha yüksek olduğu sonucuna varılmıştır.

SH < .40 olduğu BBT koşullarında EOT yetenek kestirimi kullanıldığında en düşük madde sayısı ortalaması MFB madde seçme yönteminden, en yüksek madde sayısı ortalaması KL madde seçme yönteminden elde edilmiştir. BSD yetenek kestirimi kullanıldığında en düşük madde sayısı ortalaması MFB ve KL madde seçme yöntemlerinden, en yüksek madde sayısı ortalaması a-tabakalama madde seçme yönteminden elde edilmiştir. Araştırmaya alınan bütün madde seçme yöntemleri için, EOT yetenek kestiriminden elde edilen ortalama test uzunluğu, BSD yetenek kestiriminden elde edilen ortalama test uzunluğundan yüksek bulunmuştur. BSD yetenek kestiriminin kullanıldığı BBT uygulamalarında daha kısa testler elde edileceği sonucuna varılmıştır. Madde seçme yöntemlerine ait SH değerleri, BSD yetenek kestirimi kullanıldığında daha düşük sonuç vermiştir.