ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
Research Article; 2024; 15(1); 5-17

Investigation of Measurement Precision and Test Length in Computerized Adaptive Tests under Different Conditions

Hüseyin YILDIZ* Ceren TUNABOYLU DEMİR** Süleyman ÜLKÜ***
Gamze GİRAY**** Hülya KELECİOĞLU*****

Abstract

In this study, it is aimed to examine item exposure rate, content balancing, and ability estimation in terms of termination rules with regard to testing lengths and testing accuracy in computerized adaptive testing. In this context, EAP and MLE ability estimation methods were compared in terms of correlation, bias, RMSE, and test length. In the study EAP and MLE were compared with a total of 72 different conditions; including 1, 2, and 4 group content balancing patterns; 0.50, 0.75, and 1.00 exposure rates; 0.35 and 0.40 standard error-based and the termination rule based on the test length of 15 and 30. This research is Monte-Carlo simulation study, which was carried out in relational screening model of the quantitative research methods. The production and analysis of the data were performed in the Rstudio. As a result, the best performance in the measurement is a fixed test length of 30 items with 0.35 standard error; in group 1 pattern where the content balancing is not a group limitation; the exposure rate was displayed in the range of 0.75 and 1.00. Depending on the test length of ability estimation methods, scope balancing patterns and exposure rates, the number of items changes in the range of 22 and 25; Based on the termination rule, it was estimated that at least 0.40 standard errors with a standard error based on 39 items.

Keywords: computerized adaptive testing, contend balancing, exposure rate, simulation study

Introduction

With the developments of technology field, the need for the use of computerized adaptive testing (CAT) instead of the classical paper-pencil tests in the measurement and evaluation applications has increased, and the studies have become widespread. CAT is the form of creating tests, testing individuals and scoring individuals in the computer environment (Reckase, 2009). The most important feature that separates CAT from the paper-pencil tests is that how the test starts, continues and terminates may differentiate according to the individual. The individualization mentioned here works as a set of algorithms and rules.

Classical Test Theory (CTT) was used in the first examples of CAT applications (Betz & Weiss, 1973; Larkin & Weiss, 1974; Vale & Weiss, 1975). In CTT, test and item parameters may vary according to the ability level of the group. Due to its parameter invariance feature, Item Response Theory (IRT) eliminates this disadvantage of CTT. In IRT, item parameters do not change according to the ability

* Researcher, Australian Council for Educational Research (ACER), Methodology and Measurement Department, Melbourne-Australia, huseyin.yildiz@acer.org ORCID ID: 0000-0003-2387-263X
** Branch Manager, Republic of Türkiye Ministry of National Education, General Directorate of European Union and Foreing Relations, Ankara-Türkiye, cerentunaboylu@gmail.com, ORCID ID: 0000-0001-8090-8913
*** National Education Expert, Republic of Türkiye Ministry of National Education, General Directorate of Lifelong Learning, Ankara-Türkiye, ulkusuleyman@gmail.com, ORCID ID: 0000-0003-1965-0671
**** Phd. Candidate, Hacettepe University, Faculty of Education, Ankara-Türkiye, giraygamze@gmail.com, ORCID ID: 0000-0002-5795-4521
***** Prof.Dr., Hacettepe University, Faculty of Education, Ankara-Turkey, hulyakelecioglu@gmail.com, ORCID ID: 0000-0002-0741-9934

To cite this article:
Yıldız, H., Tunaboylu Demir, C., Ülkü, S., Giray, G. & Kelecioğlu, H. (2024). Investigation of measurement precision and test length in computerized adaptive tests under different conditions, *15*(1), 5-17. https://doi.org/10.21031/epod.1068572

Received: 5.02.2022
Accepted: 1.06.2023

distribution of individuals who take the test. The predicted ability parameters do not change according to the properties of the items in the applied test (Hambleton, Swaminathan, & Rogers, 1991). IRT is a model that explains the relationship between an individual's ability level and his/her responses to the measured feature with a mathematical function (Embretson & Reise, 2000; Hambleton & Swaminathan, 1989). Although CAT applications are not dependent on IRT, the preference for IRT in CAT applications allows the results to be more effective (Weiss, 1983).

CAT starts with choosing items to start the test, estimating the ability of the test taker according to the responses given by them, and administering next item chosen based on the estimated ability level or stopping the test (Wainer, 2000). The process must be planned very well in a detailed manner to serve the purpose of the application.

The pre-condition of the CAT application is the creation of a large pool of test items. In order to achieve the advantages of CAT over paper-pencil tests, the item pool must contain high quality items in accordance with different ability levels (Flaugher, 2000). In the item pool used, sufficient number of items in accordance with each ability level must be present (Reckase, 1989). The CAT consists of four basic processing steps, including the starting of the test, item selection method, ability estimation and test termination rules (De Ayala & Koch, 1995; Reckase, 1989; Thompson & Weiss, 2011). The test begins with choosing the first item. The test can be initiated with the best distinctive or average difficulty item in accordance with the preliminary information about ability level of the individual. After the estimation of the individual's ability level according to the given response, the second item suitable for individual from the item pool is selected using different methods. A new ability estimation is performed according to the responses to the items. According to different termination criterion, the test is terminated, and the ability level of the individual is determined. As a result, the CAT application has a cycle in which the creation of the item pool, how to select the items from the item pool, how the scoring will be done, and in which situations the application will be stopped, are determined and implemented.

In CAT applications, the item to be answered by the individual is determined according to his/ her measured ability level. In this way, in test applications where maximum performance is measured, individuals with high ability levels take the more difficult items, while individuals with low ability levels take the easier items and a customized test form is formed for each individual. The basic rationale behind individualized tests is to avoid items that may be too difficult or too easy for the person taking the test and to choose the items that best suit the individual's ability. Individuals take items that provide more information for their own ability level, so that differences between individuals can be determined more clearly (Wang, 2013). The process results in shorter tests, as individuals receive items suitable for their ability level and do not waste their time dealing with more difficult or easier items for themselves (Wainer, 2000).

One of the most important advantages of CAT is its estimation individuals' abilities with a small number of items in comparison to the classical paper-pencil tests. Embretson and Reise (2000) stated that a well-patterned CAT application could reduce the test length up to 50% without significantly losing measurement accuracy. In classical paper-pencil tests, the measurement precision may vary according to the ability levels of the individuals in the group, while accurate measurements can be made according to the ability levels of the individuals in the CAT applications. However, there are problems such as the individual does not have a chance to return to the previous item in CAT applications, security violation problems caused by the disclosure of the item pool and the frequency of item use (Aybek et.al., 2014). However, in CAT applications, problems such as the individual's lack of a chance to return to the previous item, a security violation based on the disclosure of the item pool, and the frequency of item use may occur.

The item pools used in CAT applications include a large number of items. However, in some cases, the frequency of use is seen to be rather high for some of the items and for some items pretty low. When such cases are encountered, the possibility of recalling of items for the individuals can be considered high. In order to increase the utility level of the item pool, Item Bank Constraining methods have been developed. These methods are counted among the basic components of CAT applications since they offer solutions for the application problems (Davis, 2002; Boyd, 2003). These methods include Content Balancing and Item Exposure Rate.

Content Balancing

In the tests containing two or more contents, the items may vary with low and high ability levels in accordance with the content. Student group with any level of ability may be exposed to a restricted test. For example, Mathematics course, let's think about a test where four operations skills are included in a single scope. While the student group at the high ability level may only encounter with items for division skill, the student group at the low ability level may only encounter items for addition skills. In such a case, content balancing is needed.

The tests in which the content balancing is used are longer than the tests which is not. The ability, personality and preference measurements are relatively more homogeneous and one-dimensional in the content, so they do not need content balancing; however, content balancing is required for achievement tests (Weiss, 2004).

Content balancing methods can be evaluated into two categories those based on a methodological approach and approaches that select items by trying to meet a complex set of constraints (Demir, 2019). In the first approach, an item pool is divided into several sub-pools based on item attributes, and items are selected from sub-pools to meet predetermined content areas. On the other, it relies on a different approach, which makes item selection try to meet a complex set of constraints, and an item can contribute to satisfying more than one constraint at the same time.

There are Constrained CAT, Modified CCAT, and Modified Multinomial Model among the methods to ensure a fixed content balance (Lin, 2011). In addition, Weighted Deviations Model, Shadow Test Approach, Two-Phase Item Selection Procedure, Weighted Penalty Model, and Maximum Priority Index methods can be counted for large-scale applications (as cited in He, Diao, & Hauser, 2014).

In the literature, there are studies in which 2 to 6 content areas are determined and different content balancing methods are compared (Lin, 2011, Kingsbury & Zara, 2009; Kingsbury & Zara, 1980; Eggen & Netherland, 1999; Demir, 2019). These studies compared content balancing methods by keeping the number of content areas constant. In this research, using CAT, the cases where there is a different number (2, 4) of content areas and no content area were compared.

Item Exposure Rate

The use of item exposure rate is focused on protecting the integrity of the item pool and the confidentiality of the items in the item pool by blocking over-exposure to the same items (Davis & Dodd, 2005). With adaptive tests created from the same item pool, the same questions can be asked for multiple times and the individual can learn the correct answers. The most commonly used items lose their original psychometric properties by becoming popular. This situation causes the test validity to fall. The purpose of a good item exposure rate control method is to ensure the more balanced use of the item pool without reducing the measurement accuracy by defying this relationship (Pastor, Dodd & Chang, 2002). Item exposure rate control methods are used to balance the test security and measurement accuracy (Boyd, 2003; Boyd, Dodd & Fitzpatrick, 2013).

Item use frequency control methods prevent the disclosure of items by preventing excessive use of the same items, preserving the integrity of the substance pool (Davis & Dodd, 2005). It is possible to classify the item use frequency control methods into five categories. These are (1) randomization strategies, (2) conditional selection, (3) stratified strategies, (4) combined strategies, and (5) multiple-stage adaptive test designs (Lin, 2011). In this study, the frequency of use of the item was controlled by using the restricted maximum information strategy, which is one of the conditional selection methods. This method determines whether the item will be used when that item is selected by comparing it with the maximum value of the frequency of use parameter determined before the test.

Exposure rates were predetermined 0,10 and 0,20 (Chang & Ansley, 2003), 0,19 and 0,29 (Boyd et al., 2013), 0,30 (Pastor, Dodd & Chang), 0,40 (Revuelta & Ponsoda, 1998). In this research, two exposure

rates of 0,50 and 0,75 were studied due to the lack of a large item pool and the use of content balancing which is another restrictive method. To compare the effect of exposure rates, no exposure control condition was also added to the research.

When the literature is examined, there are studies in which many aspects of CAT (content balancing, item pool properties, test length, etc.) are compared under different conditions (Boyd, 2003; Eroğlu & Kelecioğlu, 2012; Demir, 2018; Aybek & Çıkrıkçı, 2018; Sulak & Kelecioğlu, 2019; Kara, 2019). It is considered that the research will contribute to the field in terms of examining the measurement accuracy and test length when ability estimation methods, content balancing approaches, item exposure rates and termination rules are changed in CAT applications. Based on the results of the research, it is evaluated that the research will contribute to the field of study by determining the conditions which provide calculations with minimum error and bias, and maximum correlation between true and estimated thetas.

Purpose of the Study

The main purpose of this study is to explain how bias, RMSE, correlation values between true and estimated thetas, and test length change according to different conditions of item exposure rate, content balancing, ability estimations methods and termination rules. Accordingly, the sub-problems of the study are given below.

a) How do bias, RMSE, correlation values between true and estimated thetas, and test length change according to different conditions of termination rules based on standard error (0,35 and 0,40), item exposure rate (0.50, 0.75, 1.00), content balancing (1 group, 2 groups and 4 groups) and ability estimation methods (Expected a Posteriori (EAP) and Maximum Likelihood (ML))?

b) How do bias, RMSE and correlation values between true and estimated thetas change according to different conditions of termination rules based on fixed length (15 and 30), item exposure rate (0.50, 0.75, 1.00), content balancing (1 group, 2 groups and 4 groups) and ability estimation methods (Expected a Posteriori (EAP) and Maximum Likelihood (ML))?

c) How do the average of bias, RMSE and correlation values between true and estimated thetas change in all conditions separately?

Simulation Methods

This study is a Monte Carlo simulation study that aims to reveal the relationship between various ability estimation methods, exposure rates, content balancing rules and termination rules in CAT applications. Collecting real data for research can be time-consuming and costly to collect. In addition, sometimes the use of real data may not be sufficient for the analyzes desired to be carried out in the research. In such cases, it may be more useful to generate the data. In the simulation study, the data is created by the researcher based on a model. Simulations have two major components. The first is a system that is of interest to the investigator, and the second is a model that represents the system. One advantage of simulation studies is that they allow researchers to compare estimated parameters against their respective true parameters, which are unknown for real data applications (Feinberg & Rubright, 2016; Wilcox, 1997). Also, simulation study is a quantitative relational research since it aims examining the relations between methods (Fraenkel & Wallen, 2006).

Data Generation

The data sets used in this study are produced by the help of the codes written by researchers in the R programming language. Fixed 200-item pools (Veerkamp & Berger, 1997) and 150 hypothetic participants (Guzmán & Conejero, 2004) are derived for each analysis. While producing ability parameters of individuals, standard normal distribution was used with mean of 0 and standard deviation of 1.

a Parameters of the items in the pools were obtained from the normal distribution of 0.8 mean and 0.1 standard deviation. b parameters used were obtained from uniform distribution in the range of (-3, + 3). Since the data generation is manufactured based on 2 parameter logistics model (2 PLM), c parameter (guessing parameters) is fixed 0.

Data Analysis

The CAT simulations were carried out with the "simulaterespondents" function in the "catR v.3.17" package of the programing language (Magis & Raiche, 2012). In this function, the ability parameters of individuals, item parameters, initial and termination rules are defined as compulsory arguments. In this study, the starting rule is fixed as an item that would generate the maximum information for a skill level to be chosen randomly in the range (-1.00,+1.00) for all analyses. Maximum Fischer Information, which is widely preferred in the literature (Choe, Kern, & Chang, 2017; Chen, Chao & Chen, 2019), was used as the item selection rule in all analyzes. This method is based on the principle of selecting the item that produces the highest information among the items in the estimated ability level after each response of the individual. The termination rules used are explained in the title of the simulation conditions because it is among the changed conditions. In this study, the correlation values between true theta scores and estimated theta scores were calculated with the Pearson correlation coefficient method, which is one of the parametric correlation methods.

Simulation Conditions

In this study, the ability estimation method, exposure rate, test termination rule and content balancing conditions were changed in CAT simulations. 3 different situations were used for content balancing. The first of these situations is not to use content balancing limitation, the second one is dividing item-pool into 2 content group, and the last situation is dividing the item-pool into 4 content group. In the conditions in which the item-pool had limitation of content balancing, analyzes were performed to be applied evenly for each group in terms of items applied.

Another condition that is changed in the study is the ability estimation method. To estimate ability Expected a Posteriori (EAP) and Maximum Likelihood (ML) ability estimation methods can be used. Maximum likelihood estimators are consistent, functions of sufficient statistics when sufficient statistics exist, efficient and asymptotically normally distributed (Hambleton & Swaminathan, 1985). The first of the methods used is the Maximum Likelihood Estimation (MLE) and the other is the Expected a Posteriori (EAP) method. MLE method is the most widely used method among the estimation methods based on the likelihood function, but it cannot give stable results when all answers are correct or incorrect. On the other hand, Bayesian approaches can make ability estimations for all response patterns (Embreston & Reise, 2000).

For CAT simulations, the "simulateRespondents" function allows us to determine both temporary ability estimates and final ability estimates. Within the scope of this study, the same method was used for temporary and final ability estimation.

Larger item pools are needed when content balancing and item exposure control methods are used to ensure content validity and test safety (Çoban, 2020). The item exposure rate restriction is used to allow an item in the item pool to be directed to a specified percentage of the group. In this study, 3 different exposure rate conditions were used for the restriction in question. In the first condition, the rate was accepted as 1.00. This rate means that the items are not brought to a restriction for the frequency of use. For 0.75 and 0.50 values used in other conditions, each item is allowed to be directed to the maximum of 75% and 50% of the groups respectively.

When to terminate the test is one of the important factors in estimating the ability level (Kezer, 2013). In this study, 4 different test termination rules are included. Two of these rules are based on the standard error limit of the ability estimation, while the other two conditions are the termination rules based on the fixed test length. While the fixed-length method is about the number of questions applied to the

individuals, the variable-length methods are related to the precision of the measurement. When the predetermined criteria are met in variable-length test termination methods, the individual's test is terminated. The minimum standard error method is the most widely used test termination method. According to this method, an individual's ability level depends on a certain standard error and if a certain measurement precision is reached, the test is terminated (Demir, 2018).As a standard error-based termination rule, 0.30 and 0.40 cutting scores were used. These values were frequently studied in the literature and were critical values in terms of test termination rules to obtain a measurement precision (Aybek & Çıkrıkçı, 2018; Sulak & Kelecioğlu, 2019; Yao, 2012). For termination rules based on fixed test length, 15 and 30 items are preferred. According to Stocking (1994), the item pool size should be at least 12 times the test length in CAT applications applied according to the fixed test length. Therefore, an item pool of 15 items was chosen to correspond to the item pool of 200 items, and an item pool of 30 items was chosen to disrupt this situation. 15 and 30 item values were preferred because they were found to be related to the test lengths obtained from the standard error-based termination rules of 0.35 and 0.40 in the preliminary analysis.

For 4 different changing conditions, 3x2x3x4= 72 simulation conditions were studied. We can not use the high number of replications (e.g. 100) used in simulation studies on different subjects. It is seen that 10-15 replications are made in similar simulation studies in which a large number of conditions are used. Basically, considering that everyone's CAT simulation is completed independently of each other and ability estimations are made separately, there is no difference between making 100 replications for 100 participants and 10 replications for 1000 participants. In addition, it took approximately 90 hours to complete 720 (72x10) simulations using a computer with high processing power, even under 10 replication conditions. So, for each condition, 720 different CAT analysis were performed in total with 10 replications (Gorin et.al., 2005; Kara, 2019).

Results

In this study, it is aimed to examine the measurement accuracy and test length of Computerized Adaptive Testing (CAT) when the ability estimation methods, content balancing patterns, exposure rates and termination rules are changed. In this context, EAP and MLE as an ability estimation, 1, 2, and 4 as content balancing pattern group patterns; 0.5, 0.75 and 1.00 as exposure rate, 0.35 and 0.40 standard error as termination rule and fixed length testing in the form of 15 and 30 items were used, and 72 different conditions were created and compared in terms of correlation, bias, RMSE and test length.

In the first stage, the analysis findings carried out according to 0.35 and 0.40 standard error-based termination rule are given in Table 1.

Table 1.

Correlation, Bias, RMSE and Test Length Results by Standard Error Termination Rule

Content Balancing	Estimation Method	Exposure Rate	Correlation		Bias		RMSE		Test Length	
			SE < 0,35	SE < 0,40	SE < 0,35	SE < 0,40	SE < 0,35	SE < 0,40	SE < 0,35	SE < 0,40
1 Group	EAP	0,50	0,867	0,612	0,006	0,037	0,496	0,785	35,823	9,565
		0,75	0,901	0,554	-0,026	-0,026	0,422	0,850	38,729	6,872
		1,00	0,907	0,695	0,003	0,016	0,409	0,701	38,414	13,692
	MLE	0,50	0,863	0,432	-0,028	-0,016	0,494	0,917	35,370	3,169
		0,75	0,894	0,603	-0,006	0,006	0,448	0,777	38,817	10,953
		1,00	0,921	0,514	0,028	-0,015	0,393	0,833	38,846	6,272

Table 1.

Correlation, Bias, RMSE and Test Length Results by Standard Error Termination Rule (Continued)

Content Balancing	Estimation Method	Exposure Rate	Correlation		Bias		RMSE		Test Length	
			SE 0,35	SE 0,40	SE 0,35	SE 0,40	SE 0,35	SE 0,40	SE 0,35	SE 0,40
2 Group	EAP	0,50	0,733	0,487	-0,058	-0,021	0,782	0,881	38,637	8,919
		0,75	0,749	0,590	-0,002	0,024	0,801	0,802	42,095	9,186
		1,00	0,871	0,570	0,012	0,038	0,508	0,780	39,365	11,136
	MLE	0,50	0,852	0,548	-0,023	0,016	0,528	0,850	37,140	8,522
		0,75	0,861	0,560	0,015	-0,035	0,492	0,828	38,951	7,355
		1,00	0,886	0,585	0,007	-0,056	0,481	0,786	39,580	10,300
4 Group	EAP	0,50	0,818	0,573	-0,007	-0,013	0,581	0,830	38,051	8,122
		0,75	0,865	0,610	-0,026	0,026	0,519	0,784	39,932	11,837
		1,00	0,850	0,665	0,028	0,042	0,538	0,734	40,042	14,427
	MLE	0,50	0,823	0,532	0,015	-0,012	0,577	0,864	37,755	8,270
		0,75	0,857	0,596	-0,032	-0,004	0,524	0,789	39,569	9,193
		1,00	0,842	0,551	0,050	-0,064	0,543	0,839	39,586	6,084

When the table 1 is examined, the lowest correlation value in all conditions of 0.35 standard error-based termination is 0.733, this value is the highest (0.695) in the analysis based on 0.40 standard errors. The bias values produced similar results in conditions where 0.35 and 0.40 standard error-based termination rules are used. For all circumstances, the bias values were approached to zero for 0.35 standard error when their absolute values are averaged. RMSE values range from 0.393 to 0.801 error-based termination rule for 0.35; and 0.701 and 0.864 for 0.40 standard error-based termination rule. In all conditions, RMSE values are estimated close to zero.

The test lengths range from 35.4 to 42.1 for 0.35 standard error-based termination criterion and from 3.2 to 14.4 for 0.35 standard error-based termination criterion. In this direction, it can be said that the condition of 0.35 standard error is used perform better than the condition that 0.40 standard error is used.

When content balancing conditions are compared under similar conditions, the correlation values are found to be higher in 1 group condition where there is no limitation in the item pool. The bias values have been met approximately similar to all conditions. RMSE values are generally relatively close to zero in 1 group condition. When the ability estimation is MLE, the deducted values are closer to each other in content balancing groups.

In the case of a comparison between EAP and MLE estimation methods, when 2 groups are used as content balancing condition, the MLE method was found to have higher correlation values, similar bias value, RMSE values closer to zero and relatively shorter test length. In cases where 1 and 4 groups are used as the content balancing condition, the values obtained in the EAP and MLE methods are similar.

When the values obtained according to exposure rates are examined, the correlation values are seen to reduce as the exposure rates decrease. The bias values are mostly higher in the EAP method for 0.75 exposure rate, while the MLE method is higher in 0.5 and 1 exposure rates. RMSE values are relatively estimated in both MLE and EAP methods as closer to each other and zero.

In the second stage, 15 and 30-item fixed test length is used according to the termination rules, the analysis findings are given in Table 2.

Table 2.

Correlation, Bias, and RMSE Results by Fixed Test Length Termination Rule

Content Balancing	Estimation Method	Exposure Rate	Correlation		Bias		RMSE	
			15	30	15	30	15	30
1 Group	EAP	0,50	0,853	0,904	0,031	0,002	0,534	0,438
		0,75	0,872	0,895	-0,034	-0,003	0,509	0,436
		1,00	0,850	0,900	-0,009	0,015	0,532	0,432
	MLE	0,50	0,836	0,905	0,019	0,001	0,548	0,424
		0,75	0,838	0,908	-0,015	0,011	0,514	0,415
		1,00	0,857	0,892	0,019	-0,004	0,523	0,464
2 Group	EAP	0,50	0,823	0,838	0,018	-0,009	0,577	0,555
		0,75	0,790	0,798	-0,005	-0,011	0,608	0,645
		1,00	0,796	0,789	-0,022	-0,043	0,621	0,649
	MLE	0,50	0,770	0,823	-0,028	-0,005	0,633	0,611
		0,75	0,828	0,908	0,003	0,011	0,548	0,419
		1,00	0,859	0,859	0,003	0,019	0,517	0,507
4 Group	EAP	0,50	0,866	0,909	0,010	0,017	0,512	0,411
		0,75	0,837	0,869	-0,007	0,015	0,540	0,486
		1,00	0,858	0,894	-0,027	0,021	0,521	0,438
	MLE	0,50	0,839	0,889	-0,010	0,020	0,563	0,455
		0,75	0,840	0,880	0,016	-0,020	0,545	0,463
		1,00	0,853	0,899	0,044	0,027	0,516	0,435

When Table 2 is examined, except for the content balancing 2 group condition, the estimation method EAP and exposure rate 1 condition, in all conditions correlation values based on the termination rule with fixed test length are higher in 30-item constant testing lengths. According to test lengths, bias values did not show a specific pattern according to the content balancing, estimation method and exposure rate. The average of the absolute values of the bias values for all conditions were found closer to zero for 30-item testing. Except for the content balancing 2 group condition, the estimation method EAP and exposure rate 0.75 and 1 conditions; RMSE values were estimated to be smaller for 30-item testing. In this respect, it can be said that the 30-item condition based on the fixed test length is better performed than 15-item condition.

Under similar conditions, when content balancing conditions are compared, the correlation values were higher in the case of 4 group limitations in the item pool. The bias values were predicted as similar. RMSE values are relatively close to zero in conditions of the 1 group, 4 groups and 2 group limits, respectively. The predicted values are closer to each other in content balancing groups when the ability estimation is MLE.

When EAP and MLE estimation methods are compared, the correlation values, RMSE and bias values are found to have no significant differences. When the values obtained according to exposure rates are examined, the correlation and bias values are generally reduced as the exposure rate decreases in the MLE method; however, RMSE and test length increase.

In order to facilitate the comments and comparisons obtained according to all conditions, the values obtained by the averages of all other conditions are given in Table 3. When the test length averages are taken, the fixed test lengths are not included in the mean.

Table 3.

Correlation, Bias, RMSE, and Test Length Averages by Simulation Conditions

	Conditions	Correlation	Bias	RMSE	Test Lenght
Termination Rule	0,35	0,853	0,021	0,530	38,7
	0,40	0,571	0,026	0,813	9,1
	15	0,837	0,018	0,548	-
	30	0,875	0,014	0,482	-
Estimation Method	EAP	0,785	0,020	0,601	24,7
	MLE	0,783	0,019	0,585	23,1
Content Balancing	1	0,803	0,016	0,554	23,0
	2	0,757	0,020	0,642	24,3
	4	0,792	0,023	0,584	24,4
Exposure Rate	0,50	0,766	0,018	0,619	22,4
	0,75	0,788	0,016	0,590	24,5
	1,00	0,798	0,026	0,571	24,8

When Table 3 is examined, 0.853 in the standard error-based termination rule of 0.853, in 0.40 standard error-based termination rule 0.571; in 15-item fixed test length 0.837, and in 30-item fixed test length 0.875 average correlation value was obtained. The highest correlation value was obtained in the condition of 30-item constant test length. In the case of 30-item fixed test length, the bias value was 0.014 and the RMSE value was 0.482, and these values were found to be closer zero in all other conditions.

The average test length was estimated as 38.7 in standard error-based termination rule as 0.35, and 9.1 in standard error-based termination rule in 0.40.

When the values are examined according to the estimation methods, it is estimated that the average correlation value in EAP condition is 0.785 while it is 0.783 in MLE. These values are quite close; and it is similar for the bias, RMSE and the test length. The average test length is estimated in EAP condition as 24.7, and 23.1 in MLE condition.

For content balancing 1 group condition, the average correlation value is 0.803, 0.757 in 2 group condition and 0. 792 in 4 group condition. The highest mean correlation value is in content balancing 1 group condition and the value of bias (0.016) and RMSE (0.554) is closest to zero (0.016). The average test length was 23.0 in content balancing group 1, 24.3 in 2 group condition, and 24.3 in t4 group condition.

The average correlation values according to exposure rate are 0.766 in 0.50, 0.75 in 0.788 and 1 in 0.798 in 1. The highest average bias value (0.016) was obtained in exposure rate 0.75, while the average bias value (0.016) is the smallest in exposure rate 0.75. RMSE is predominantly predicted in exposure rate 1 (0.571) closer to zero. The average test length was estimated 22.4 in exposure rate 0.50, 24.5in condition 0.75; 24.8 in condition 1.

Discussion and Conclusion

In general, when the results were examined, a standard error rule of 0.35 as a stop rule has performed in terms of better in terms of RMSE, correlation, and bias. Similarly, in previously research where they compared .30 and .40 standard error-based termination rule, Özbaşı and Demirtaşlı (2015) determined .30 standard error-based termination rule performed better than .40. 15 and 30-items test length conditions based on fixed test length are added to standard error-based comparisons, 30 items test performed better in terms of RMSE, correlation and bias among the four different termination rules. These findings are supported by the studies in the literature (Eroğlu & Kelecioğlu 2015; Lee, 2014; Calender, 2011; Babcock & Weiss, 2012).

When the content balancing methods were compared within themselves, the highest correlation was observed when the number of groups was 1 and low when the number of groups was 4. It was evaluated that there was no interpretable relationship between the number of groups and the correlation, since the values were not ordered depending on the number of groups and the difference between the correlation values was small. Leung, Chang, and Hau (2003) conclude that content balancing does not have a systematic effect on the measurement accuracy is in parallel with the findings of this study. In addition to this, in the study mentioned, group 1 pattern, where there were not coverage balancing limitation, performed well similarly. The reason why the content balancing influences the measurement accuracy is the restrictions it brings to the item bank. It can be interpreted that if the item pool is large enough and a parameters are or high or the number of content groups are few; content balancing will not affect measurement accuracy significantly.

A noticeable difference was not detected between EAP and MLE ability estimation methods. Similarly, in the studies carried out by Kezer (2014), Malak and Kelecioğlu (2019), they have not found a difference between EAP and MLE approaches. In addition, the analysis where EAP method used took much time. It can be said that the use of the MLE method may be preferred in terms of the economic use of time.

In the exposure rate, it can be said that the 1.00 condition is better performed than 0.75 and 0.50 conditions. However, in terms of the item security, the values obtained for the exposure rate at 1.00 and 0.75 are close, a value between 0.75 and 1.00 can be chosen for the exposure rate. In the case of exposure rate conditions similar to content balancing conditions, the differences between the correlation values obtained by 1.00, 0.75 and 0.50 were small. The exposure rate is also related to the restricting the use of item bank such as content balancing. The fact that the difference between the conditions are small is due to a sufficiently large item pool, it is thought that other items selected from the wide pool can make similar estimations at certain points.

As a result, according to the results of the simulation, 0.35 standard error based 30 items fixed length based termination rules; 1 group content balancing and between 0.75 and 1.00 exposure rate conditions were seen to perform better in terms of RMSE, correlation and bias. In addition, in terms of test length, ability estimation methods, content balancing patterns and exposure rates are estimated approximately 22 to 25, while the standard error-based termination rules are predicted to be 39 for 0.35 standard error and 9 for 0.40 standard error.

The average correlation value obtained with a 15-item fixed termination rule is 0.837; This value is 0.875 for 30-item conditions. It was observed that there is only a 0.03 increase in the correlation value in terms of test length between these two finishing rules. In this respect, 15-item termination condition is considered more efficient.

In order to better observe the effects of content balancing and exposure rates, especially on the measurement, it is thought that similar studies can be carried out with smaller item pools, with a greater number of content groups, or lower exposure conditions. It is also thought that the content balancing and exposure ratio variables can interact with the item selection methods. Therefore, a similar study can also be performed by changing item selection methods.

In addition, because the low values of the standard error increase the measurement accuracy, practitioners can determine the appropriate standard error according to the vitality of the test. If the content balancing method is used, larger item pool is needed in each content area. Since there is no significant difference between the MLE and EAP methods in terms of measurement accuracy, the MLE method can provide to practitioners an advantage in terms of analysis time. In the large of item pool, the item exposure rate did not differ much between 1 and 0.75. Practitioners should use a large item pool if they want to use a item exposure rate.

Declarations

Author Contribution: Hüseyin YILDIZ: data analysis, conceptualization, investigation, methodology, visualization, writing - review & editing. Ceren TUNABOYLU DEMİR: conceptualization, investigation, writing - review & editing. Süleyman ÜLKÜ: conceptualization, investigation, writing - review & editing. Gamze GİRAY: conceptualization, investigation, writing - review & editing. Hülya KELECİOĞLU: writing-review & editing, supervision

Funding: The authors declare that no funds, grants, or other support were received during the preparation of this manuscript.

Ethical Approval: We declare that all ethical guidelines for authors have been followed by all authors. Ethical approval is not required as the data in this study were generated by a computer program.

Consent to Participate: All authors have given their consent to participate in submitting this manuscript to this journal.

Consent to Publish: Written consent was sought from each author to publish the manuscript. Competing Interests: No potential conflict of interest was reported by the authors.

References

Chen, L-Y. (2010). *An investigation of the optimal test design for multi-stage test using the generalized partial credit model.* [Doctoral dissertation, The University of Texas]. UT Electronic Theses and Dissertations. https://repositories.lib.utexas.edu/handle/2152/ETD-UT-2010-12-344

Choi, Y. J., & Asilkalkan, A. (2019). R packages for item response theory analysis: Descriptions and features. *Measurement: Interdisciplinary Research and Perspectives*, *17*(3), 168-175. https://doi.org/10.1080/15366367.2019.1586404

Aybek, E., & Çıkrıkçı, R. (2018). Kendini Değerlendirme Envanteri'nin Bilgisayar Ortamında Bireye Uyarlanmış Test Olarak Uygulanabilirliği [Applicability of the self assessment inventory as a computerized adaptive test] . Türk Psikolojik Danışma ve Rehberlik Dergisi, 117-141. https://dergipark.org.tr/en/pub/tpdrd/issue/40299/481364

Aybek, E. C., Şahin, D. M., Eriş, H. M., Şimşek, A. S., & Köse, M. (2014). Kağıt-kalem ve bilgisayar formunda uygulanan testlerde öğrenci başarısının karşılaştırıldığı çalışmaların meta-analizi [Meta-analysis of comparative studies of student achievement on paper-pencil and computer-based test]. Asya Öğretim Dergisi, 2(2), 18-26. https://dergipark.org.tr/en/pub/aji/issue/1539/18831

Babcock, B. & Weiss, D.J. (2012). Termination criteria in Computerized Adaptive Tests: do variable-length CAT's provide efficient and effective measurement? International Association for Computerized Adaptive Testing, 1, 1-18. https://doi.org/10.7333/1212-0101001

Boyd, M. A. (2003). Strategies for Controlling Testlet Exposure Rates in Computerized Adaptive Testing Systems. Unpublished Doctoral Thesis, The University of Texas, Austin.

Boyd, A. M., Dodd, B., & Fitzpatrick, S. (2013). A Comparison of Exposure Control Procedures in CAT Systems Based on Different Measurement Models for Testlets. Applied Measurement in Education, 113–135. https://doi.org/10.1080/08957347.2013.765434

Chen, J.-H., Chao, H.-Y., & Chen, S.-Y. (2019). A Dynamic Stratification Method for Improving Trait Estimation in Computerized Adaptive Testing Under Item Exposure Control. Applied Psychological Measurement, 1-15. https://doi.org/10.1177/0146621619843820

Davis, L. L. (2002). Strategies for Controlling Item Exposure in Computerized Adaptive Testing with Polytomously Scored Items. Unpublished Doctoral Thesis, The University of Texas, Austin.

Davis, L. L., & Dodd, B. G. (2008). Strategies for Controlling Item Exposure in Computerized Adaptive Testing with Partial Credit Model. Pearson Educational Measurement, , 9(1), 1. https://doi.org/10.1177/0146621604264133

Demir, S. (2018). Çok kategorili bireyselleştirilmiş bilgisayarlı test uygulamalarının farklı madde seçim yöntemlerinde sonlandırma kuralları açısından incelenmesi [Investigation of Different Item Selection Methods in Terms of Stopping Rules in Polytomous Computerized Adaptive Testing]. [Unpublished Doctoral Thesis]. Hacettepe University, Ankara.

Demir, S. (2019). Bireyselleştirilmiş Bilgisayarlı Sınıflama Testlerinde Sınıflama Doğruluğunun İncelenmesi [Investigation of Classification Accuracy at Computerized Adaptive Classification Tests]. [Unpublished Doctoral Thesis]. Hacettepe University, Ankara.

Choe, E., Kern, J., & Chang, H.-H. (2017). Optimizing the Use of Response Times for Item Selection in Computerized Adaptive Testing. Journal of Educational and Behavioral Statistics, 1-24. https://doi.org/10.3102/1076998617723642

Embretson, S. E., & Reise, S. P. (2000). Item Response Theory for Psychologists. Taylor & Francis.

Eroğlu, M. G., &Kelecioğlu, H. (2012). Bireyselleştirilmiş bilgisayarli test uygulamalarinda farklı sonlandırma kurallarının ölçme kesinliği ve test uzunluğu açısından karşılaştırılması [Comparison of different test termination rules in terms of measurement precision and test length in computerized adaptive testing]. Uludağ Üniversitesi Eğitim Fakültesi Dergisi, 28(1), 31-52. https://doi.org/10.19171/uuefd.87973

Feinberg, R. A., & Rubright, J. D. (2016). Conducting Simulation Studies in Psychometrics. Educational Measurement: Issues and Practice, 35(2), 36-49. https://doi.org/10.1111/emip.12111

Wainer, H., Dorans, N. J., Flaugher, R., Green, B. F., Mislevy, R. J. (2000). Computerized Adaptive Testing: A Primer Second Edition (s. 37-59). Lawrence Erlbaum Associates, Publishers. https://doi.org/10.4324/9781410605931

Fraenkel, J., & Wallen, N. (2011). How to design and evaluate research in education (6th ed.). McGraw-Hill, Inc.

Gorin, J. S., Dodd, B. G., Fitzpatrick, S. J., & Shieh, Y. Y. (2005). Computerized adaptive testing with the partial credit model: Estimation procedures, population distributions, and item pool characteristics. Applied Psychological Measurement, 29(6), 433-456. https://doi.org/10.1177/0146621605280072

Guzmán, E., & Conejo, R. (2004, August). A model for student knowledge diagnosis through adaptive testing. In International Conference on Intelligent Tutoring Systems (pp. 12-21). Springer. https://doi.org/10.1007/978-3-540-30139-4_2

Hambleton, R. K., & Swaminathan, H. (1985). Item response theory. Springer.

He, W., Diao, Q., & Hauser, C. (2014). A comparison of four item-selection methods for severely constrained CATs. Educational and Psychological Measurement, 74(4), 677-696. https://doi.org/10.1177/0013164413517503

Kalender, İ. (2011). Effects of Different Computerized Adaptive Testing Strategied on Recovery of Abilitiy. Unpublished Doctoral Thesis, Middle East Technical University, Ankara.

Kara, B. E. (2019). Computer adaptive testing simulations in R. International Journal of Assessment Tools in Education, 6(5), 44-56. https://doi.org/10.21449/ijate.621157

Kezer, F. (2013). Bilgisayar Ortamında Bireye Uyarlanmış Test Stratejilerinin Karşılaştırılması [Comparision of The Computerized Adaptive Testing Strategies].[Unpublished Doctoral Thesis]. Ankara University, Ankara.

Lee, M. (2014). Application of Higher-Order IRT Models And Hierarchical IRT Models To Computerized Adaptive Testing. Unpublished Doctoral Thesis, University of California, Los Angeles.

Lin, C. (2011). Item selection criteria with practical constraints for computerized classification testing. Applied Psychological Measurement 71(1), 20-36. https://doi.org/10.1177/0013164410387336

Magis, D., & Raiche, G. (2012). Random Generation of Response Patterns under Computerized Adaptive Testing with the R package catR. Journal of Statistical Software, 1-31. https://doi.org/10.18637/jss.v048.i08

Özbaşı, D., & Demirtaşlı, N. (2015). Bilgisayar okuryazarlığı testinin bilgisayar ortamında bireye uyarlanmış test olarak geliştirilmesi [Development of computer literacy test as computerized adaptive testing]. Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi, 6(2), 218-237. https://doi.org/10.21031/epod.79491

Pastor, D. A., Dodd, B. G., & Chang, H.-H. (2002). A Comparison of Item Selection Techniques and Exposure Control Mechanisms in CATs Using the Generalized Partial Credit Model. Applied Psychological Measurement , 147-163. https://doi.org/10.1177/01421602026002003

Reckase, M. D. (2009). Designing item pools to optimize the functioning of a computerized adaptive test. Psychological Test and Assessment Modeling, 127-141.

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, 315-326. https://doi.org/10.21031/epod.530528

Thompson, N. A., & Weiss, D. J. (2011). A Framework for the Development of Computerized Adaptive Tests. Practical Assessment, Research & Evaluation., 1- 9. https://scholarworks.umass.edu/pare/vol16/iss1/1/

Yao, L. (2013). Comparing the Performance of Five Multidimensional CAT Selection Procedures With Different Stopping Rules. Applied Psychological Measurement, 3-23. https://doi.org/10.1177/0146621612455687

Weiss, D. J. (2004). Computerized Adaptive Testing for Effective and Efficient Measurement in Counseling and Education. Measurement and Evaluation in Counseling and Development, 70-84. https://doi.org/10.1080/07481756.2004.11909751

Wainer, H. (2000). Introduction and history. In H. Wainer (Ed.), Computerized adaptive testing: A primer. Lawrence Erlbaum Associates.

Wilcox, R. R. (1997). Simulation as a research technique. In J. P. Reeves (Ed.), Educational research, methodology, and measurement: An international handbook (pp. 150–153). Pergamon.

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. https://doi.org/10.3102/10769986022002203