International Journal of Assessment Tools in Education
2022, Vol. 9, No. 3, 682–696
https://doi.org/10.21449/ijate.1105769
Published at https://ijate.net/ https://dergipark.org.tr/en/pub/ijate *Research Article*

The Effect of ratio of items indicating differential item functioning on computer adaptive and multi-stage tests

Basak Erdem-Kara[1,*], **Nuri Dogan**[2]

[1]Anadolu University, Faculty of Education, Department of Educational Sciences, Eskisehir, Türkiye
[2]Hacettepe University, Faculty of Education, Department of Educational Sciences, Ankara, Türkiye

ARTICLE HISTORY

Received: Apr. 19, 2022
Revised: Aug. 08, 2022
Accepted: Aug. 19, 2022

Keywords:
Computer adaptive test, Multi-stage test, Differential item functioning.

Abstract: Recently, adaptive test approaches have become a viable alternative to traditional fixed-item tests. The main advantage of adaptive tests is that they reach desired measurement precision with fewer items. However, fewer items mean that each item has a more significant effect on ability estimation and therefore those tests are open to more consequential results from any flaw in an item. So, any items indicating differential item functioning (DIF) may play an important role in examinees' test scores. This study, therefore, aimed to investigate the effect of DIF items on the performance of computer adaptive and multi-stage tests. For this purpose, different test designs were tested under different test lengths and ratios of DIF items using Monte Carlo simulation. As a result, it was seen that computer adaptive test (CAT) designs had the best measurement precision over all conditions. When multi-stage test (MST) panel designs were compared, it was found that the 1-3-3 design had higher measurement precision in most of the conditions; however, the findings were not enough to say that 1-3-3 design performed better than the 1-2-4 design. Furthermore, CAT was found to be the least affected design by the increase of ratio of DIF items. MST designs were affected by that increment especially in the 10-item length test.

1. INTRODUCTION

Traditional linear tests that have been the milestone of educational assessment since the 1900's are generally administered using paper-pencil and have been a popular way to measure examinees' knowledge, skills, and abilities (Weiss & Kingsbury, 1984; Yan et al., 2014). However, especially over the past 40 years, computer-based tests have gained popularity over linear tests thanks to great advances in computer technology, thereby becoming a viable alternative to those paper-pencil tests (Keng, 2008; Luecht & Sireci, 2011; Magis et al., 2017; Yan et al., 2014). According to Yan et al. (2014) computer-based tests can be classified into three main groups; namely, linear, adaptive, or multi-stage.

Computer-based linear tests are the computerized version of traditional linear tests. As in linear tests, all individuals answer the same items in these tests, and the test length is fixed (Magis et al., 2017; Sarı, 2016; Yan et al., 2014). On the other hand, the primary purpose of computer adaptive tests (CAT) is to select items from the item pool so as to match the ability level of the

*CONTACT: Basak ERDEM KARA ✉ basakerdem@anadolu.edu.tr Anadolu University, Faculty of Education, Department of Educational Sciences, Eskisehir, Türkiye

e-ISSN: 2148-7456 /© **IJATE** 2022

individual and to ensure that test is neither too easy nor too difficult for the individual (Thai, 2015; Yan et al., 2014, Zheng & Chang, 2014). In the process, an item is administered and answered and the individual's ability level (theta) at that point is estimated according to the answer. Depending on that estimated theta, the next item is chosen from the pool and administered. Until the stopping criteria are met, the process goes on (Tay, 2015; Weiss & Kingsbury, 1984). Individuals only face items convenient to their ability levels and do not spend time with items which are too easy or too difficult for them. Thus, the main advantage of CAT over linear tests is that they reach the desired measurement precision with fewer items (Wainer, 2000; Wang, 2013; Wang, 2017).

The other type of computer-based tests which has become popular, especially in recent years, is multi-stage test (MST), which combines the many advantages of linear tests and CAT while minimizing their disadvantages (Hendrickson, 2007; Magis et al., 2017). MST which can be considered as a variation of adaptive testing differs from CAT in test adaptation level. While test adaptation occurs at item level in CAT, adaptation occurs at item set (module) level in multi-stage testing (Hendrickson, 2007; Yan, 2010). In MST, a set of items which is named as the module is administered to the examinee, examinees' ability is estimated based on his/her responses to that module, and s/he is routed to the next module at the next stage (Hendrickson, 2007; Wang, Lin, Chang, & Douglas, 2016). In MST, each module can be assembled so as to have desired contextual and statistical specifications; thus, test developers have more control over the construction of the desired test form when compared to CAT. Although, MST has less adaptation points than those of CAT, they provide more efficient test assembly and more controlled content balancing. Furthermore, MST allows some item review and change previous answers within each module. However, going back to previous stages and reviewing items in previous module/s are not allowed in MST. (Hendrickson, 2007; Sarı & Huggins-Manley, 2017; Wainer, 2000; Wang, 2017). In addition to their advantages, MST also has some disadvantages such as requiring more items to get the same measurement precision with CAT (Berger, Verschoor, Eggen & Moser, 2019). Besides, since MST modules are designed so as to be at optimum difficulty only at target ability levels (e.g., three levels at low, medium, and high proficiency), final ability estimations may not be as accurate as CAT designs (Rome, 2017).

The increase in computer-based testing application has brought some problems especially in test fairness issue (Chu & Lai, 2013; Gierl, Lai & Li, 2013; Zwick, 2010). Test fairness and equity issues are related with items presenting some bias towards a specific group of students. Non-bias items only measure ability of individuals that is intended to be measured without being affected by unrelated factors such as gender, socio-economic status, etc. On the other hand, bias items are affected by those factors which are not related with the characteristic which is intended to be measured. Because test results are used in critical decision-making situations that may affect individuals' future, test fairness becomes even more significant (Camilli & Shepard, 1994; Crocker & Algina, 1986; Hambleton & Swaminathan, 1991). Differential item functioning analyses are one of the most popular methods used to get information on the bias. Potentially problematic items are identified with DIF analyses and expert opinions are obtained on whether those items are really problematic or not (Zumbo, 1999).

1.1. DIF and Adaptive Testing

The quality of adaptive testing applications largely depends on item pool quality (Han & Guo, 2011). Therefore, large item pools should be developed for those applications and each item in that pool should be checked in order to ensure that they satisfy the main fairness and equity issues (Gierl, Lai & Li, 2013). However, even if the item writing process is well planned and carefully designed, it is not easy to avoid the effects of DIF completely. Many factors which are not related with an item such as computer familiarity, testing environment, physical impairments, etc. may cause DIF (Birdsall, 2011). Independent of the item content, the context

in which the item is presented, for instance, item order, may also affect item parameters and may become a source for DIF (National Research Council, 1999). Besides, although items in the pool have no DIF initially, some may become DIF items over time. As a result of repeatedly usage of items over time, they may become known for other individuals prior to their administration. Even if this is not the case, the interaction between the item and test taker may change because of several reasons, which is known as item parameter drift. Therefore, the changing interaction between an item and a test taker may cause different item characteristics than initially calibrated item characteristics (Aksu-Dunya, 2017; Han & Guo, 2011). Parameter drift on items could be defined as a kind of DIF since items behave differently in groups which are involved in different testing applications (Aksu-Dunya, 2017; Babcock & Albano, 2012). Item parameter drift is a serious threat to validity and fairness (Han & Guo, 2011). DIF analyses may be more important for adaptive testing applications than they are for linear tests. Since the number of items administered in adaptive tests is fewer than in linear tests, each item has greater effect on final ability estimation. Therefore, any flaw in an item may cause more consequential results (Zwick, 2010; Gierl et al., 2013; Zwick & Bridgeman, 2014). So, DIF items may play an important role in examinees' test scores. Besides, performing the test application via computer may reveal some possible sources of DIF such as computer familiarity, anxiety, and environment that are not found in traditional tests (Zwick, 2010). These factors have increased the importance of DIF analyses in computer adaptive tests. Steinberg et al. (2000) stated that adaptive tests may be more sensitive to the effects of DIF on validity than linear tests. In addition, the presence of bias may affect the order of administration of the items because the next item/module in CAT and MST is determined according to the answers to the previous item/module (Zwick, 2010). It is important to note that concentration of biased items on certain modules for the MST may also pose a problem.

1.2. Purpose of the Study

Despite the importance of the existence of DIF items in adaptive testing is known, DIF studies in adaptive tests are limited to a few studies in the literature (Chu & Lai, 2013; Gierl et al., 2013; Lei, Chen & Yu, 2006; Piromsombat, 2014). Besides, those studies were limited to the investigation and comparison of DIF detection methods under different conditions (Chu & Lai, 2013; Gierl et al., 2013; Lei, Chen & Yu, 2006) and the investigation of the effect of DIF items on ability estimation on CAT (Piromsombat, 2014). No studies were found in the literature focusing on comparison of CAT and MST approaches in case of the presence of DIF items in the test. The current study aims to investigate the performances of two adaptive testing approaches, CAT and MST, in case of the presence of DIF items under different conditions. Therefore, the results of the study are likely to contribute to the literature focusing on DIF in adaptive testing applications. To this end an answer for the following research question is sought in the context of this research:

- *How does the test performance of CAT and MST change in case of the presence of DIF items on the test under different test lengths (10-20-30-40 item), test designs (CAT, 1-3-3 MST and 1-2-4 MST), and ratio of DIF items (10%, 20% and 30%)?*

2. METHOD

Within the scope of the research, it is aimed to examine the effect of the inclusion of items that have differential item functioning (DIF) in the test on the effectiveness of CAT and MST under different conditions. The data used in the research were generated by the simulation method and different test designs were compared under different conditions in a controlled manner. The related study is a Monte Carlo simulation study in which the data are simulated. Simulation data were preferred because it was difficult to meet all the conditions discussed in the study simultaneously in real data.

2.1. Research Design

In this study, test performances of three different adaptive test designs (CAT, 1-3-3 MST and 1-2-4 MST) were compared in case the test consists of DIF items. Those MST designs were some of the most popular ones. Two-stage test designs have only one adaptation point which may make them open for routing errors more (Yan, et al., 2014; Zenisky et. al, 2010). On the other hand, it was stated that more than three stages add little to the accuracy of ability estimations and increase the complexity of test designs (Yan et al., 2014). In general, maximum four modules in one stage and three stages were thought to be enough (Armstrong et al., 2004; Zenisky et al., 2010). The preferred test designs in this study were among the most popular ones used in the literature.

The manipulated factors were test length with four levels and ratio of DIF items in the test with three levels. Three different test designs (CAT, MST 1-2-4 and MST 1-3-3) were compared under three different DIF item ratio and four different test length conditions. All manipulated conditions were fully crossed within each of three test designs, which resulted in 36 conditions (Three Test Designs, Four Test Length and Three DIF Item Ratio). For each condition, 30 replications were performed and the whole simulation processes were performed by using R programming language (R Core Team, 2018). Detailed information on simulation processes is given as follows:

2.2. Data Generation

Five thousand examinees were randomly generated based on standard normal distribution and the same theta values were used for all test designs. Generated theta values were restricted to be generated between -3 and 3 in order to eliminate the effect of outliers. Besides, an item pool of 600 items was generated using the three-parameter logistic model. Discrimination, difficulty, and guessing parameters were randomly sampled from Uniform (0.5, 2.0), N (0, 1) and Uniform (0, 0.25) distributions, respectively. Difficulty parameters were restricted to be in the range of [-3, 3]. Descriptive statistics related to item pool are given in Table 1.

Table 1. *Descriptive statistics of item pool.*

	a parameter	b parameter	c parameters
N	600	600	600
Mean	1.268	-0.097	0.125
Standard Deviation	0.442	1.22	0.074
Minimum	0.501	-2.967	0.0002
Maximum	1.999	2.988	0.249

As can be seen in Table 1, the discrimination values (a parameter) had a minimum value of 0.501 and a maximum of 1.999 with a mean of 1.268. The item pool had items with a wide range of discrimination. Item difficulties ranged from -2.967 to 2.988 with a mean of -0.097 indicating that the item pool had items with a wide difficulty range in the specified range of [-3, 3]. Guessing parameter ranged from 0.0002 to 0.249 with a mean of 0.125. When the test information function of that item pool was examined, it was seen that items in the pool gave high information especially around the point where the ability level was 0 and covered the [-3, 3] ability range as intended.

2.2.1. *Generation of DIF items*

Item pool was developed to have 200 items on each difficulty level and 600 items in total. After that, 20% of the items on each difficulty level were randomly selected and rendered into DIF items. In order to make those items indicate DIF, +1 constant was added to the initial b

parameters and that value was considered as the focal group b parameter. For DIF items, the difference between b parameters of focal and reference groups was set as +1 (bfocal – breference = 1) and those items always worked in favor of the reference group, which means that all of them had uniform DIF (U-DIF). As a result, there were 40 U-DIF and 160 Non-DIF items in each level and 120 U-DIF and 480 Non-DIF items in total.

2.3. CAT and MST Simulations

After generation of item parameters, theta values and formation of DIF items, CAT, and MST environments were constructed. For CAT and MST simulations conducted on the same item pool and the same theta values, the commonly manipulated variable is the test length (10-20-30-40) and the rate of DIF items in the test (10% - 20% - 30%). Besides, panel designs (1-2-4 and 1-3-3) were manipulated within MST simulation. RMSE, bias, and correlation values were calculated and averaged across 30 replications and the performance of simulations was interpreted based on those values. In order to make better comparisons; maximum item exposure rates, IRT model, ability estimation method, and item/module selection method were fixed for both CAT and MST. Maximum Fisher Information (MFI) method was used in item selection and Expected A Posteriori (EAP) estimation method was used in ability estimation for both CAT and MST. MFI method is preferred since it provides the selection of the item that provides the maximum information each time. Although this method is quite popular, it has the disadvantage that items with high discrimination levels are chosen more because they provide more information and choosing those same items over and over leads to item exposure problem (Hambleton, Jac ve Pieters, 2000; van der Linden & Pashley, 2010; Wang, 2017); hence, this method should be used carefully. Controlling the item exposure can be an effective method to prevent this situation. As another method, Hambleton et al. (2000) suggested that the item be chosen randomly among items that provide maximum information at the relevant skill level. In our study, both of those methods were implemented. The maximum item use rate was fixed as 0.25 for CAT and four separate parallel panels were created for MST to ensure that the maximum item use rate was 0.25. Besides, 'randomesque' method was used and instead of choosing the most informative item, items were randomly selected from among the most informative ones at that ability level. As the ability estimation method, two most common methods are maximum likelihood (MLE) and Bayesian methods. However, MLE can be problematic since it cannot be estimated for individuals who answer all items correctly or incorrectly. This is particularly problematic for the early stages of computer adaptive tests and is not recommended when the test length is short. The use of MLE is not recommended until a true or false answer is received (Hambleton & Swaminathan, 1991; Wang, 2017). Bayesian methods are more consistent for short-length tests. The combination that is generally suggested for item selection and ability estimation in adaptive tests is the EAP estimation in ability estimation together with the maximum information method in item selection (van der Linden, 2008; van der Linden & Pashley, 2010). That is why, MFI with EAP combination was preferred in this study.

All simulation processes were carried out with the help of the catR and mstR packages that are conjugate of each other. Detailed explanations of simulations are given as follows:

2.3.1. *CAT simulations*

CAT environment was created via catR package and 12 different conditions in total (4 test lengths x 3 ratios of items with DIF), including four different test lengths (10-20-30-40) and three different ratios of items with the DIF (10 % - 20% - 30%) were examined as specified earlier. As a starting rule, the initial ability level was set to 0 and this value was used for each condition. According to this rule, the initial ability levels of individuals were accepted as '0' (zero) and the first item that the individual would encounter was determined accordingly.

2.3.2. *MST simulations*

To create MST environment, xxIRT and mstR pckages were used. In MST simulation, 24 different conditions in total, including four different test lengths, three different ratios of items with DIF, and two other test designs (1-2-4 and 1-3-3) were examined. In the 1-3-3 MST design, a single module was used in the first stage, while there were three modules each in the second and third stages. For that 1-3-3 design, which included 7 modules in total, a single module common to all individuals was created at the first stage, and the difficulty level of this module was determined as medium. The three modules in the second and third stages had three different difficulty levels (easy, medium, and hard). Each individual answered three modules, each one from a different stage, in total. Similarly, in the 1-2-4 panel design, individuals responded to a total of three modules. Individuals who answered a single module in the first stage were directed to one of the two different modules in the second stage according to the ability estimations obtained from the first module. After completing this second stage, they were directed to one of the four modules in the third stage, considering the abilities estimated at the end of the second stage. The number of items in the modules and the number of items required to form a panel differed according to test lengths and panel design and are presented in detail in Table 2.

Table 2. *Number of items in modules and panels.*

Panel Design		Test Length			
		10	20	30	40
1-3-3	Module length	3-3-4	6-7-7	10-10-10	13-13-14
	Number of items used in panel	24	48	70	94
1-2-4	Module length	3-3-4	6-7-7	10-10-10	13-13-14
	Number of items used in panel	25	48	70	95

Since the modules in a panel are at different ability levels, the number of items used while creating the related panel is more than the test length, e.g, in the 1-3-3 design, in the condition that the test length was 40, individuals answered a total of 40 items, 13 each in the first two stages, and 14 in the last stage. However, while 13 items were needed in the first stage, 39 and 42 items were needed in the second and third stages for the modules at three different levels, respectively. As a result, a total of 94 items were used. For both panel designs, 10%, 20%, and 30% of the items in the modules in the second stage were selected among the items with DIF, e.g., in the case where the test length was 10, under the condition that the rate of items with DIF is 20%, 8 of the items were selected among the items that did not show DIF and 2 of them showed DIF. It was ensured that the selected 2 DIF items were included in the modules of the second stage.

Within the scope of the study, four different panels were created, so that the maximum panel, module, and item exposure became comparable with the CAT. Four different panels were obtained through an open source "mixed integer linear programming solver" (lp_solve 5.5) included in the xxIRT package, and it was ensured that the items used in one panel were not included in the other panels. “Bottom-up” method was used in the creation of the panels. In this method, firstly, four different parallel forms were created for each module. In order to ensure that the modules were parallel, information function targets were determined at the module level and the modules were structured to meet those targets. The items in the modules were chosen to provide maximum information at the specified skill levels. After the construction of four parallel forms for each module, those modules were assigned to the panels randomly and parallel panels were obtained. Thanks to the parallelism of the constructed modules, these modules could be used alternately between the panels (Yan et al., 2014).

2.4. Data Analysis

In the analysis of data, Root Mean Square Error (RMSE), bias, and correlation (ρ) values between estimated and true ability parameters were used to evaluate the results obtained from CAT and MST. $\hat{\theta}_j$ represents the estimated ability parameter, θ_j represents the true ability parameter, and N represents the total number of individuals. RMSE and bias values were calculated with the help of the following formulas:

$$RMSE = \sqrt{\frac{\sum_{j=1}^{N}(\hat{\theta}_j - \theta_j)^2}{N}} \qquad Bias = \frac{\sum_{j=1}^{N}|(\hat{\theta}_j - \theta_j)|}{N} \tag{1}$$

The correlation value was obtained by the following formula, with the standard error values of the estimated ($\sigma_{\hat{\theta}_j}$) and true (σ_{θ_j}) ability parameters.

$$\rho_{\hat{\theta}_j,\theta_j} = \frac{cov(\hat{\theta}_j, \theta_j)}{\sigma_{\hat{\theta}_j}\sigma_{\theta_j}} \tag{2}$$

RMSE, bias, and correlation values were calculated for each of the 30 replications and interpretations were made based on the average of those values. Based on the calculated values, evaluations were made as to which of the two MST and one CAT application gave higher measurement accuracy than the others under different conditions. After those evaluations, whether the differences between the test designs (CAT, MST 1-3-3 and MST 1-2-4) reached a significant level were examined by ANOVA analysis. Post-Hoc analyses were made for the design groups that differed significantly from each other and the results were interpreted.

3. RESULTS

In this section, results of CAT and MST simulations are presented in detail. Findings are evaluated for each condition under different DIF item ratio.

3.1. Results of the Condition that the Ratio of DIF Items is 10%

In this part, ratio of DIF items in the test was fixed at 10% and the performance of three test design was examined under different test lengths. RMSE, bias and correlation values are presented in Table 3.

Table 3. *RMSE, bias and correlation values of test designs under test length and DIF item ratio.*

DIF Item Ratio	Test Length	RMSE			Bias			Correlation		
		CAT	MST (1-3-3)	MST (1-2-4)	CAT	MST (1-3-3)	MST (1-2-4)	CAT	MST (1-3-3)	MST (1-2-4)
10%	10 items	0.269	0.382	0.389	0.212	0.299	0.307	0.963	0.924	0.921
	20 items	0.192	0.282	0.306	0.151	0.221	0.240	0.982	0.963	0.953
	30 items	0.164	0.246	0.253	0.130	0.193	0.200	0.987	0.974	0.969
	40 items	0.155	0.236	0.238	0.122	0.184	0.187	0.989	0.977	0.974
20%	10 items	0.268	0.408	0.400	0.211	0.318	0.312	0.963	0.913	0.916
	20 items	0.192	0.284	0.320	0.151	0.223	0.252	0.982	0.962	0.949
	30 items	0.164	0.252	0.270	0.130	0.197	0.213	0.987	0.971	0.964
	40 items	0.153	0.242	0.241	0.121	0.190	0.191	0.989	0.975	0.973
30%	10 items	0.269	0.448	0.451	0.212	0.348	0.352	0.963	0.894	0.892
	20 items	0.194	0.301	0.303	0.153	0.237	0.238	0.981	0.955	0.954
	30 items	0.166	0.276	0.269	0.131	0.218	0.212	0.987	0.962	0.965
	40 items	0.153	0.246	0.259	0.121	0.194	0.204	0.989	0.970	0.967

As indicated in Table 3, RMSE values range from [0.155, 0.269] for CAT design, [0.236, 0.382] for MST 1-3-3 design, and [0.238, 0.389] for MST 1-2-4 design. The bias values ranged between [0.122 - 0.212] for the CAT design, [0.184-0.299] for the MST 1-3-3 design, and [0.187-0.307] for the MST 1-2-4 design. It was seen that CAT application had the lowest and MST 1-2-4 application had the highest RMSE and bias values for all test lengths. However, in MST 1-3-3 and 1-2-4 designs, those values seemed to be quite close to each other throughout all test lengths. In addition, it was observed that as the number of items increased, the bias values decreased and the difference between the designs decreased. Finally, when the correlation values were examined, it was observed that those values varied between the range of [0.963-0.989] for CAT, [0.924-0.977] for the MST 1-3-3 design and [0.921- 0.974] for the MST 1-2-4 designs. Looking at the correlation values in Table 3, it was determined that the design with the highest correlation value throughout all test lengths was CAT and the design with the lowest correlation value was the MST 1-2-4 design. Correlation values increased as the number of items increased for all test designs and got closer to each other.

3.2. Results of the Condition that the Ratio of DIF Items is 20%

In this part, ratio of DIF items in the test was fixed at 20% and the performance of three test design was examined under different test lengths. As indicated in Table 3, RMSE values range from [0.153, 0.268] for CAT, [0.242, 0.408] for MST 1-3-3, and [0.241, 0.400] for MST 1-2-4 design. The lowest RMSE value for all test lengths was obtained in CAT, while the highest RMSE value was found in the MST 1-3-3 design for the 10 and 40-item tests, and the MST 1-2-4 design for the 20 and 30-item tests. When the bias values were examined, it was seen that the CAT design had values in the range of [0.121, 0.211], the MST 1-3-3 design was in the range of [0.190, 0.318], and the MST 1-2-4 design was in the range of [0.191, 0.312]. As is seen in Table 3, the lowest bias values belong to CAT whereas the highest bias is in the MST 1-3-3 design for 10 items, and in the MST 1-2-4 design for other test lengths. When the test designs were compared in terms of correlation, the design with the highest correlation across all test lengths was CAT [0.963, 0.989], the lowest correlation was in the MST 1-3-3 design for the 10-item test, and the MST 1-2-4 design for the other test lengths. As the number of items increased for all designs, the correlation values increased and got closer to each other.

3.3. Results of the Condition that the Ratio of DIF Items is 30%

Finally, values were examined for the condition that DIF item ratio was fixed at 30%. As indicated in Table 3, considering the RMSE values, the lowest RMSE value for all test lengths was obtained in the CAT design. The RMSE values for the MST 1-3-3 and 1-2-4 designs appear to be quite close to each other for all test lengths. When the bias values were examined, it was observed that the lowest bias values were calculated in the CAT design at all test lengths and MST designs gave very close results to each other. When the test designs were compared in terms of correlation values, the design with the highest correlation value across all test lengths is the CAT design [0.963-0.989] (Table 3). As the test length increased, the bias value for all designs decreased and correlation values increased as the number of items increased for all designs.

3.4. ANOVA Analysis

After the interpretation of RMSE, bias and correlation values, separate one-way ANOVA tests were conducted in order to observe whether those values differ significantly between test designs. Three separate one-way ANOVA analyses were performed for each DIF item ratio (10%, 20% and 30%), in which the RMSE, bias, and correlation values were taken as the dependent variable and the test design as the independent variable, and the findings were analyzed separately for each test length. While the assumption of normal distribution was provided in the analyses, the assumption of homogeneity of variances was violated in some

cases. In cases where that assumption was violated, the Welch test was used and in other cases the data in the ANOVA table (Table 3) were interpreted. ANOVA results are given in detail for each DIF item ratio condition as follows:

3.4.1. *Ratio of DIF items is 10%*

As a result of ANOVA analysis, it was seen that RMSE, bias, and correlation values significantly differed between test designs at each test length (p<.05). According to the results of the Post-Hoc comparison, the difference in RMSE, bias, and correlation values reached a significant level among all designs at all test lengths. In short, the lowest values for RMSE and bias were obtained in CAT design at all test lengths and those values differed significantly from the values of the MST designs. The highest RMSE and bias values were observed in the MST 1-2-4 design at all test lengths and differed significantly from others. For the correlation, the highest values were obtained in CAT and the lowest values were obtained in the MST 1-2-4 design along all test lengths. The difference in correlation values between the designs was significant over all test lengths. When all the results were considered together, it was concluded that the CAT design that had the lowest RMSE, bias, and the highest correlation values provided the highest measurement precision. On the other hand, the MST 1-2-4 design, which had the highest RMSE and bias and the lowest correlation values, was the design with the lowest measurement precision.

3.4.2. *Ratio of DIF items is 20%*

As a result of ANOVA analyses made for the condition that the DIF item ratio was 20%, RMSE, bias, and correlation value differences among designs were found to be significant for all cases. Therefore, it was concluded that CAT was the test that provided the highest measurement accuracy among the three designs. The design with the highest RMSE was MST 1-3-3 for the 10 and 40-item tests, and MST 1-2-4 for the 20 and 30-item tests. The difference between the MST designs reached a significant level in the 10, 20, and 30-item tests. The highest values of the bias were in the MST 1-3-3 design for 10 items, and in the MST 1-2-4 design for the other test lengths. The lowest correlation was obtained from the MST 1-3-3 design in the 10-item test and the MST 1-2-4 design in the other test lengths. When all the results were considered together, it was concluded that the CAT design with the lowest RMSE and bias and the highest correlation values provided the highest measurement precision. Besides, the lowest measurement precision was obtained in the MST 1-3-3 design in the 10-item test and in the MST 1-2-4 design for other test lengths.

3.4.3. *Ratio of DIF items is 30%*

It was seen that RMSE, bias, and correlation values significantly differed between test designs at each test length (p<.05) as in the previous DIF item ratios. When the Post-Hoc comparison results were analyzed in terms of the RMSE variable, there was a significant difference between all designs for the 10, 30, and 40-item tests; however, in the 20-item test, it was seen that the mean difference of .001 between the MST 1-3-3 and MST 1-2-4 designs could not reach a significant level. When the mean differences of the bias values of the designs were examined, it was concluded that all three designs differed significantly from each other for all test lengths. Finally, for the correlation values, Post-Hoc results were examined and it was concluded that there was a significant difference between all designs for all test lengths. Therefore, the highest measurement accuracy was obtained for the CAT design as this measurement precision was maintained over all test lengths and it was significantly higher than the precision of other designs. It can be seen that the values of the MST designs were very close to each other. The lowest measurement precision for the 30-item test was observed in the MST 1-3-3 design, and for the other conditions it was in the MST 1-2-4 design.

Finally, in order to descriptively see the effects of the increase in the rate of items with DIF in the test on RMSE, bias and correlation values, graphs were formed and presented in Figure 1. Looking at the RMSE graph, RMSE values for CAT were quite close to each other at different DIF ratios; however, it was observed that an increase in the DIF ratio increased the RMSE value in the MST designs, especially in the 10-item test where the test length was the lowest. For MST designs, the effect of the increase in DIF ratio on the RMSE decreased as the number of items increased. Similarly, the bias values were close to each other at different DMF ratios for CAT. In the MST designs, the increase in the DIF ratio in the 10-item test affected the bias values considerably, and this effect decreased as the test length increased. Looking at the correlation graph in Figure 1c, similar comments can be made to the comments made for RMSE and bias. It was found that the correlation values decreased as the DIF ratio increased for the CAT designs.

It was determined that the increase in DIF item ratio indicated the most serious effect for the 10-item test. For CAT, the increase in the DIF item ratio did not have a great effect. Those findings showed that CAT was the least affected test design by the increase in the ratio of items with DIF in the test. Two MST designs generally indicated parallel findings which were especially affected by the change in DIF item ratio in the 10-item test, and this effect decreased as the test length increased.

Figure 1. *Change of RMSE, bias, and correlation values with the increase of DIF item ratio.*

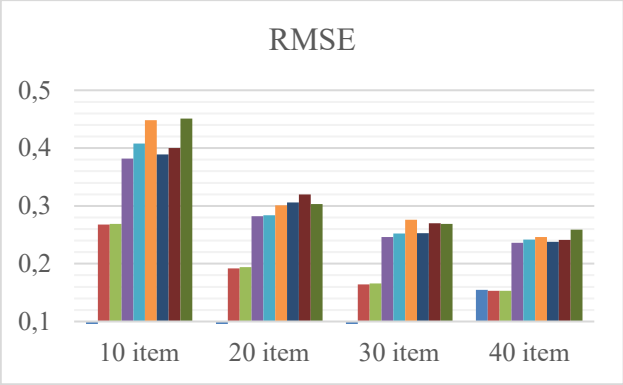

Figure 1a. *RMSE values.*

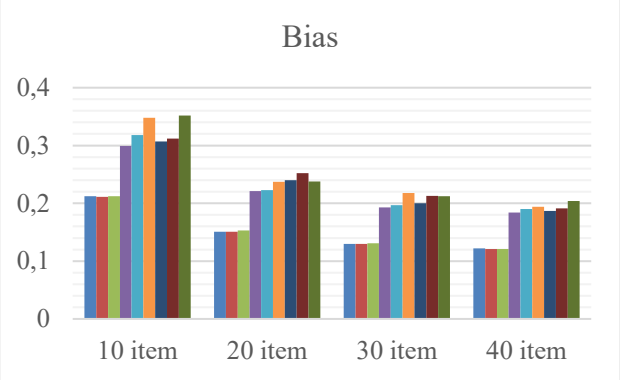

Figure 1b. *Bias values.*

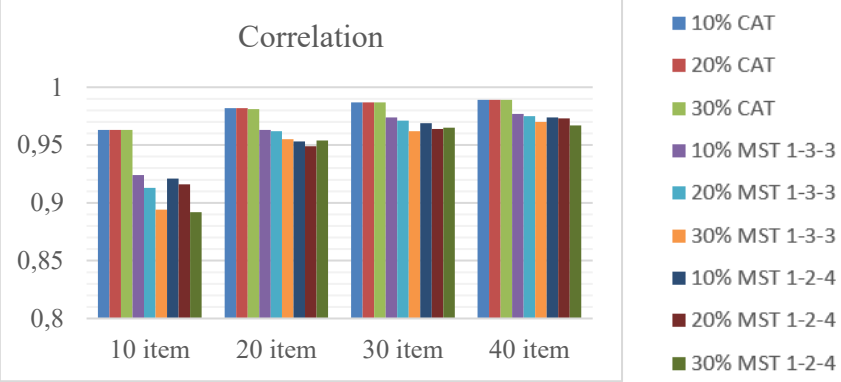

Figure 1c. *Correlation values.*

4. DISCUSSION and CONCLUSION

Within the scope of this study, it was aimed to examine the effect of the inclusion of items that have differential item functioning (DIF) in the test on the effectiveness of computer adaptive test (CAT) and multi-stage test (MST) under different conditions. For this purpose, data were generated by the simulation method and the performances of different test designs (CAT, MST 1-3-3 and MST 1-2-4) were compared under different test lengths and DIF-item ratios.

In order to evaluate test performances, RMSE, bias, and correlation values were considered together. When the obtained results were analyzed in terms of RMSE and bias, it was seen that the CAT design had the lowest values for all conditions. When the MST 1-3-3 and 1-2-4 designs were compared, a general interpretation couldn't be made. While the RMSE value of the 1-2-4 design was significantly higher than that of the 1-3-3 design throughout all test lengths in the condition that the DIF rate was 10%. When it was 20%, the RMSE of MST 1-3-3 was higher in the 10 and 40-item test length, and it was higher for the 30-item test when the ratio was 30%. Similar to the RMSE, in the condition that the DIF item rate was 10%, while the bias value of the 1-2-4 design was significantly higher than that of the 1-3-3 design throughout all test lengths; the 1-3-3 design gave higher bias values in the 10 and 40-item test when it was 20%, and in the 30-item test when it was 30%. Finally, the findings obtained from the correlation values indicate that CAT had the highest correlation value in all conditions. The lowest correlation values were obtained for 1-3-3 design in 20% DIF-10 items, 30% DIF-30 item conditions, and 1-2-4 design in all other conditions.

In addition, when we looked at how the increase in the DIF ratio affected the performance of the test designs, it was observed that the CAT gave similar results in terms of RMSE, bias and correlation, regardless of the ratio of items with DIF (Figure 1). However, the same was not the case for MST designs. The increase in the DIF ratio in the MST designs generally led to an increase in the RMSE and bias values and a decrease in the correlation values. In particular, the 10-item tests were more affected by the increase in the DIF item ratio than that in the CAT, and this effect decreased as the number of items increased.

When the information given above is interpreted, it can be concluded that CAT provided better measurement accuracy compared to the other two MST designs under all test length and DIF item ratio conditions. In addition, the design that was least affected by the increase in the ratio of items with DIF was CAT. Therefore, it can be interpreted that CAT could reduce the effect of DIF more than other designs. When the two MST designs were compared, it was seen that the 1-3-3 design offered higher measurement accuracy in most conditions. However, those findings were not sufficient to say that the 1-3-3 design outperformed the 1-2-4 design.

The main finding from this study is that the CAT was the design that minimized the effect of DIF throughout all test lengths. The finding that CAT can regulate the effect of DIF is in line with the findings obtained from the study of Piromsombat (2014). Piromsombat examined the effect of DIF items in the test on ability estimation on CAT and revealed that CAT can modulate the effect of DIF if it comes early in the test, especially when the DIF level is moderate. In other cases, CAT reduced the effect of DIF. Besides, that the number of adaptation points in CAT is higher than that in MST can result in higher CAT measurement accuracy (Sarı, 2016; Thai, 2015). For example, while the 1-3-3 panel design has only two adaptation points, regardless of the number of items, there are 19 adaptation points in a 20-item CAT. This finding may be the result of this fact. Since CAT has more adaptation points than MST desings have, CAT may control the DIF effect in a better way. Another finding obtained from this specific study is that the effect of increase in DIF item ratio on CAT performance is lower compared to the effect on MST designs. MST designs were highly affected by the increase in DIF item rate, especially when the number of items was 10, which is also thought to be relevant with the number of adaptation points. As stated before, CAT has more adaptation points than MST has, regulating

the DIF effect better. Therefore, it is expected that CAT offers better measurement accuracy compared to MST designs and is less affected by the DIF item ratio in the presence of an item with DIF in the test. Since no other studies examining the effect of DIF items on adaptive tests have been found in the literature, the discussion on this finding has been limited.

Apart from the DIF effect, studies that CAT and MST designs were compared in the literature were also examined. Kim and Plake (1993) examined measurement precision of CAT and MST in terms of first stage module length (10, 15 and 20 items), total test length (40, 45 and 50 items), number of second stage modules (6, 7, 8 modules), and item difficulty distribution in the first stage module. It has been revealed that CAT gives better results in terms of measurement accuracy than MST does. In the study conducted by Patsula (1999), the accuracy of the ability estimations obtained from different CAT designs, paper-and-pencil tests, and MST designs (number of stages, number of modules in each stage, and number of items in each module) were compared and it was determined that CAT produced the most accurate ability estimation and that the increase in the number of modules in each stage affected the measurement precision and effectiveness. In another study, Sarı (2016) investigated the precision of the results obtained from CAT and MST, while the number of content areas varied in tests of different lengths. The main finding of the study was that CAT gave better results than the other two MST designs for all conditions and the two MST designs offered comparable results. In addition, Tay (2015) stated that CAT has more adaptation points than those of MST, therefore they are more effective designs. The common result obtained from the studies in the literature is that CAT gives better results than MST does in different studies and under different conditions. This inference based on those studies shows parallelism with the finding that the CAT performance obtained as a result of the study is higher than the MST performance.

The last finding to state, not related with DIF again, was that when available findings were examined, it was seen that the RMSE and bias values decreased and the correlation values increased as the test length increased for all designs. Therefore, it can be concluded that increasing the test length increases the measurement accuracy. Similar to this finding, Sarı (2016) also revealed in his study that increasing the test length resulted in a decrease in the RMSE and bias value and an increase in correlation for both CAT and MST. Another finding obtained as a result of the research was that regarding the comparison of MST designs among themselves, the 1-3-3 design offered high measurement accuracy in a larger number of conditions, but the available findings were not sufficient to say that the 1-3-3 design outperforms the 1-2-4 design. There is no study in the literature comparing those two designs. Findings from different studies are needed to make a discussion about the relevant finding. Based upon the results of our particular study, some recommendations for practitioners are stated as follows. Firstly, it has been seen that CAT gives better results compared to MST for situations where items with DIF are present in the test. In cases with similar conditions to this study, the use of CAT may be recommended. Secondly, MST designs were more affected by items with DIF than those with CAT. Both MST designs used could not regulate the effect of DMF, and the measurement accuracy was more negatively affected compared to that of CAT. If MST is to be used, DIF analyses must be performed. Lastly, especially when the test is 10-items length, the increase in the DIF rate negatively affects the measurement accuracy of the MST. In those cases, the use of MST should not be preferred or should be used very carefully. When RMSE, bias, and correlation values were carefully examined, it can be said that values were getting closer with the increment of test length and they were very close especially after 30 items for all designs. Therefore, a test with at least 30 items can be recommended to use in cases where the presence of DIF is suspected. Those findings were thought to make a significant contribution to the literature since there were no studies found in the literature focusing on comparing CAT and MST approaches in case of the presence of DIF items in the test.

4.1. Further Research

The data set used in the research is limited to simulation data and the item pool used in the study is limited to the item parameters determined by the researcher. It can be recommended to work with real data set in future research. An item pool can also be created with different item parameter distributions and values and the study can be repeated. Besides, only dichotomously scored (1-0) items were taken into account within the scope of the study. Similar studies can be done with polytomously scored items. On the other hand, items were produced to show only uniform DIF when generating items with DIF. Similar studies can be done by adding items indicating non-uniform DIF. The study can be replicated by changing the effect size of the generated DIF items. In addition, since only fixed length was used as test termination rule in this study, the research can be repeated by using different test termination rules. Another limitation of the study is that only two designs were used for the MST. Therefore, the study can be repeated with different MST designs.

Acknowledgments

This research article was produced from the doctoral dissertation of first author under the supervision of second author.

Declaration of Conflicting Interests and Ethics

The authors declare no conflict of interest. This research study complies with research publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the authors.

Authorship Contribution Statement

Başak Erdem Kara: Investigation, Methodology, Visualization, Software, Formal Analysis, and Writing-original draft. **Nuri Dogan**: Investigation, Methodology, Supervision, and Validation.

Orcid

Basak ERDEM KARA https://orcid.org/0000-0003-3066-2892
Nuri DOGAN https://orcid.org/0000-0001-6274-2016

REFERENCES

Aksu-Dunya, B. (2017). *Item parameter drift in computer adaptive testing due to lack of content knowledge within sub-populations* (Publication No. 10708515) [Doctoral Dissertation, University of Illinois]. ProQuest Dissertations & Theses.

Armstrong, R.D., Jones, D.H., Koppel, N.B., & Pashley, P.J. (2004). Computerized adaptive testing with multiple-form structures. *Applied Psychological Measurement*, *28*(3), 147–164. https://doi.org/10.1177/0146621604263652

Babcock, B., & Albano, A.D. (2012). Rasch scale stability in the presence of item parameter and trait drift. *Applied Psychological Measurement*, *36*(7), 565-580. https://dx.doi.org/10.1177/0146621612455090

Berger, S., Verschoor, A.J., Eggen, T.J.H.M., & Moser, U. (2019). Improvement of measurement efficiency in multistage tests by targeted assignment. *Frontiers in Education*, *4*(1), 1–18. https://doi.org/10.3389/feduc.2019.00001

Birdsall, M. (2011). *Implementing computer adaptive testing to improve achievement opportunities*. Office of Qualifications and Examinations Regulation Report. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/606023/0411_MichaelBirdsall_implementing-computer-testing-_Final_April_2011_With_Copyright.pdf

Camilli, G., & Shepard, L.A. (1994). *Methods for identifying biased test items* (4th ed.). Sage Publications, Inc.

Chu, M.W., & Lai, H. (2013). Detecting biased items using CATSIB to increase fairness in computer adaptive tests. *Alberta Journal of Educational Research*, *59*(4), 630–643. https://doi.org/10.11575/ajer.v59i4.55750

Crocker, L. & Algina, J. (1986). *Introduction to classical and modern test theory*. Wadsworth Group/Thomson Learning.

Gierl, M.J., Lai, H., & Li, J. (2013). Identifying differential item functioning in multi-stage computer adaptive testing. *Educational Research and Evaluation*, *19*(2-3), 188–203. https://www.tandfonline.com/doi/full/10.1080/13803611.2013.767622

Hambleton, R.K., & Swaminathan, H. (1991*). Item response theory: Principles and applications*. Springer.

Hambleton, R.K., Jac, N.Z., & Pieters, J.P.M. (2000). Computerized adaptive testing: Theory, applications and standards. In R.K. Hambleton, & J.N. Zaal (Eds.), *Advances in educational and psychological testing: Theory and application*s (4th ed., pp. 341–366). Springer.

Han, K.T., & Guo, F. (2011). *Potential impact of item parameter drift due to practice and curriculum change on item calibration in computerized adaptive testing* (Report No. RR-11-02). Graduate Management Admission Council (GMAC) Research Reports. https://www.gmac.com/~/media/Files/gmac/Research/research-report-series/rr1102_itemcalibration.pdf

Hendrickson, A. (2007). An NCME instructional module on multistage testing. *Educational Measurement: Issues and Practice*, Summer 2007, 44-52. https://doi.org/10.1111/j.1745-3992.2007.00093.x

Keng, L. (2008). *A comparison of the performance of testlet-based computer adaptive tests and multistage tests* (Publication No. 3315089) [Doctoral Dissertation, University of Texas]. ProQuest Dissertations & Theses.

Lei, P.W., Chen, S.Y., & Yu, L. (2006). Comparing methods of assessing differential item functioning in a computerized adaptive testing environment. *Journal of Educational Measurement*, *43*(3), 245-264. http://dx.doi.org/10.1111/j.1745-3984.2006.00015.x

Luecht, R.M., & Sireci, S.G. (2011). *A review of models for computer-based testing* (Report No. 2011-12). College Board Research Report. https://files.eric.ed.gov/fulltext/ED562580.pdf

Magis, D., Yan, D., & von-Davier, A. (Eds.). (2017). *Computerized adaptive and multistage testing with R: Using packages catR and mstR*. Springer.

National Research Council (1999). *Designing mathematics or science curriculum programs: A guide for using mathematics and science education standards*. National Academies Press. https://www.nap.edu/catalog/9658.html

Piromsombat, C. (2014). *Differential item functioning in computerized adaptive testing: Can cat self-adjust enough*? (Publication No. 3620715) [Doctoral Dissertation, University of Minnesota]. ProQuest Dissertations & Theses.

Sari, H.I. (2016). *Examining content control in adaptive tests: Computerized adaptive testing vs. Computerized multistage testing* (Publication No. 403003) [Doctoral Dissertation, University of Florida]. The Council of Higher Education National Thesis Center.

Sari, H.I., & Huggins-Manley, A.C. (2017). Examining content control in adaptive tests: Computerized adaptive testing vs. computerized adaptive multistage testing. *Educational Sciences: Theory and Practice*, 17, 1759-1781. http://doi:10.12738/estp.2017.5.0484

Steinberg, L., Thissen, D., & Wainer, H. (2000). Validity. In H. Wainer (Ed.), *Computerized adaptive testing: A primer* (2. ed., p. 185–229). Routledge.

Tay, P.H. (2015). *On-the-fly assembled multistage adaptive testing* (Publication No. 3740572). [Doctoral Dissertation, University of Illinois]. ProQuest Dissertations & Theses.

van der Linden, W.J. (2008). Using response times for item selection in adaptive testing. *Journal of Educational and Behavioral Statistics*, *33*(1), 5-20. https://doi.org/10.3102/1076998607302626

van der Linden, W.J., & Pashley, P.J. (2010). Item selection and ability estimation in adaptive testing. In W. J. van der Linden & C.A.W. Glas (Eds.), *Elements of adaptive testing*. Springer.

Wainer, H. (2000). Introduction and history. In H. Wainer (Ed.), *Computerized adaptive testing: A primer* (2nd ed., p. 1–22). Lawrence Erlbaum Associates.

Wang, K. (2017). *A fair comparison of the performance of computerized adaptive testing and multistage adaptive testing* (Publication No. 10273809). [Doctoral Dissertation, Michigan State University]. ProQuest Dissertations & Theses.

Wang, S., Haiyan, L., Chang, H.H., & Douglas, J. (2016). Hybrid computerized adaptive testing: From group sequential design to fully sequential design. *Journal of Educational Measurement*, *53*(1), 45–62. https://doi.org/10.1111/jedm.12100

Wang, X. (2013). *An investigation on computer-adaptive multistage testing panels for multidimensional assessment* (Publication No. 3609605). [Doctoral Dissertation, University of North Carolina]. ProQuest Dissertations & Theses.

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

Yan, D. (2010). *Investigation of optimal design and scoring for adaptive multi-stage testing: A tree-based regression approach* (Publication No. 3452799). [Master Thesis, Fordham University]. ProQuest Dissertations & Theses.

Yan, D., von-Davier, A.A., & Lewis, C. (2014). Overview of computerized multistage tests. In D. Yan, A.A. von-Davier, & C. Lewis (Eds.), *Computerized multistage testing* (p. 3–20). CRC Press; Taylor & Francis Group.

Zheng, Y., & Chang, H.H. (2014). On-the-fly assembled multistage adaptive testing. *Applied Psychological Measurement*, 39 (2), 104-118. https://doi.org/10.1177/0146621614544519

Zumbo, B.D. (1999). *A handbook on the theory and methods of differential item functioning* (DIF). Headquarters of National Defense.

Zwick, R. (2010). The investigation of differential item functioning in adaptive tests. In W.J. van der Linden and C.A.W. Glas (Eds.), *Elements of adaptive testing*. Springer

Zwick, R., & Bridgeman, B. (2014). Evaluating validity, fairness, and differential item functioning in multistage testing. In D. Yan, A.A. von-Davier, & C. Lewis (Eds.), *Computerized multistage testing*. CRC Press; Taylor&Francis Group.