International Journal of Turkish Educational Studies,
2023, 11 (21), 549-587
https://doi.org/10.46778/goputeb.1267319

ISSN: 2148-2314

The Effect of Test Design on Misrouting in Computerized Multistage Testing*

Mahmut Sami Yiğiter[a] Nuri Doğan[b]

[a] *Lect. Dr., Social Sciences University of Ankara, Ankara, Türkiye, mahmutsamiyigiter@gmail.com*

[b] *Prof. Dr., Hacettepe University, Ankara, Türkiye, nuridogan2004@gmail.com*

ABSTRACT

Computerized Multistage Testing (MST) is an adaptive testing approach in which the test taker completes stages and modules on a pre-assembled panel according to his/her ability level. In MST, the test taker is routed to a module in the following stage based on his/her responses to the module in each stage. The test taker is expected to be routed to the module that fits his/her ability level best in the following stages. If the test taker is not routed to the module appropriate to his/her ability level, misrouting can be mentioned. Misrouting is thought to affect both measurement accuracy and the test taker's psychology. Although it is very difficult to completely eliminate misrouting, it is assumed that it can be reduced with the basic components of the MST design. The purpose of this study is to determine the level of misrouting according to different MST designs and to investigate the effects of changes in test design on the level of misrouting. The main components that are considered to affect misrouting are the MST test design [1-3, 1-2-3, 1-3-3], routing module design [Wide, Narrow], test length [12, 24, 36] and module length [L-S, M-M, S-L]. This study, which aims to reveal the current situation, is descriptive research and it was carried out by simulation method. The results of the study show that MST design and components can be effective in reducing misrouting. Three-stage MST designs offer lower misrouting and higher measurement accuracy than two-stage MST designs. Furthermore, increasing the test length and designing the routing module with a wide range of abilities reduce the misrouting rate. According to the measurement accuracy results, it can be stated that misrouting is not a significant problem in the MST in general, although the measurement accuracy of the misrouted test takers is low. It was concluded that the ability levels of the misrouted test takers were generally concentrated at the intersection points of the module information functions of the adjacent modules and generally in the middle of the ability scale.

Article Type
Research

Article Background
Received:
18.03.2023
Accepted:
11.04.2023

Keywords
Computerized Multistage Testing, Routing, Misrouting, Adaptive Testing, Measurement Accuracy, Adaptive Testing

To cite this article: Yiğiter, M. S. & Doğan, N. (2023). The effect of test design on misrouting in computerized multistage testing. *International Journal of Turkish Educational Sciences, 11* (21), 549-587.

Corresponding Author: Mahmut Sami Yiğiter, e-mail: mahmutsamiyigiter@gmail.com

* This study was presented verbally at the "8th International Congress on Measurement and Evaluation in Education and Psychology" organized by EPODDER in Izmir on September 21-23, 2022.

Introduction

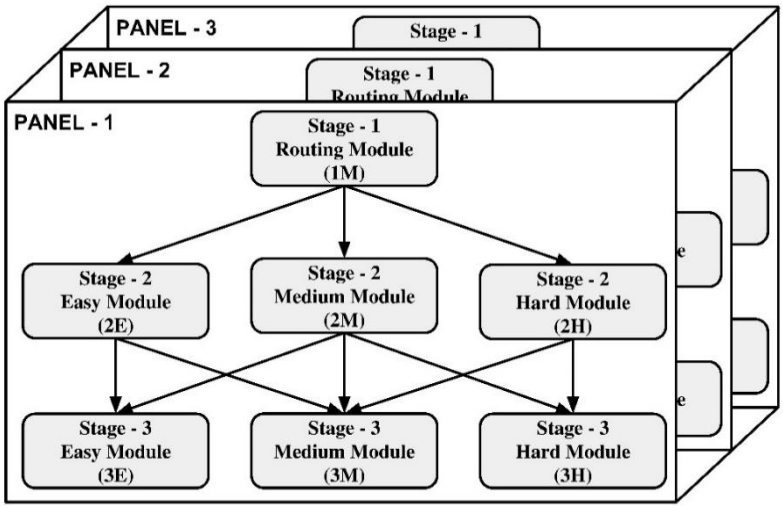
Linear Tests (LT), in which all test takers take the same test forms, have historically been the most frequently used method to measure students' knowledge, skills, and abilities in educational assessments. However, in the last half-century, adaptive tests have improved and gained popularity with the developments and models in computer software and hardware systems. Adaptive testing is a type of computerized test in which the difficulty level of each question is adjusted by an algorithm based on the test taker's response to the previous question (Wainer et al., 2001). Different Adaptive Tests have been developed according to the basic components of the test. Computerized Adaptive Testing (CAT), with shorter test length and more effective ability estimation,hase been adopted in many testing applications around the world (Khorramdel et al., 2020). Computerized Multistage Testing (MST), which has gained attention especially in recent years, is an adaptive testing approach that takes benefits from the advantages and reduces the disadvantages of LT and CAT (Zenisky et al., 2010).

In recent years, an MST has been used in many large-scale assessments due to its advantageous aspects. The Certified Public Accountants (CPA) Examination in the USA has adopted the MST since 2004 (Breithaupt et al., 2006). In 2011, a revised version of the Graduate Record Examinations (GRE) was administered with the MST. In 2011, the OECD used an adaptive approach with MST in a large-scale international test in the Programme for the International Assessment of Adult Competencies (PIAAC) (Kirsch & Lennon, 2017; Erdem Kara, 2022; Yiğiter & Doğan, 2023). Moreover, in the PISA 2018 and TIMSS 2023 cycles, an MST-like multi-stage testing approach is used. The main reason why MST has gained a place especially in international large-scale tests is that test forms like LT can be prepared and revised before test administration, and it is an adaptive test approach in which a test taker can progress based on his/her own ability level as in, CAT.

The modules and panels in the MST design are assembled from the items in the item pool before the test implementation. The items in the item pool are assembled according to their specific characteristics to form modules. The stages and panels are formed by assembling the modules on the CAT design. Figure 1 shows the 1-3-3 CAT design consisting of three panels.

Figure 1

1-3-3 MST Design



As can be seen in Figure 1, the 1-3-3 MST design consists of three panels with three stages in each panel. As a test taker progresses through a panel, to which he/she is usually randomly assigned, he/she first takes the top module called the routing module. Based on the test taker's responses to the module, his/her interim ability level is estimated, and the test taker is assigned to one of the three modules in the second stage at easy, medium, or hard level according to the routing rule. As in the first stage, the interim ability level is estimated again at the end of the second stage, taking into account the test taker's responses. The test taker is assigned to one of the modules in the third stage according to the estimated ability level. After the test taker answers the third module, the final ability level is estimated, and the test is ended.

Misrouting in Computerized Multistage Testing

While the CAT is an item-based adaptive test, MST is a module-based adaptive testing approach. As the CAT is an item-based adaptive approach, the ability level is updated after each item is administered to the test taker. In MST, the ability level is updated after each module. Therefore, simulation studies show that the MST offers slightly lower measurement accuracy than the CAT due to the lower number of adaptation points (Patsula, 1999; Wang, 2017). In addition, since the MST has fewer adaptation points, it is thought that the effect of possible misrouting between stages will be greater. In the MST, the assignment of a test taker to a module that is not appropriate for his/her ability level is called misrouting. For example, a test taker with a true ability level of -1.25 takes the routing module (1R) in the first stage of the 1-3 MST design. Then, in the second stage, he/she is expected to be assigned to the easy (2E) module among the easy (2E), medium (2M) and difficult (2D) modules. However, if the test taker is assigned to the medium (2M) or difficult (2D) module in the second stage according to his/her interim ability level estimated from his/her answers to the routing module, misrouting occurs. There may be two main reasons for misrouting. First, the test taker may have performed above or below his/her ability level in the routing module and his/her ability level may have been estimated higher or lower than his/her true ability level. The probability of this situation occurring is generally low. Second, components in the test design may have caused misrouting. For example, items in the routing module may not have successfully estimated the student's ability level or there may have been errors in the routing rules. As a result of the misrouting, it is expected that the test taker will be assigned to a module that is not appropriate for his/her level. Assigning the test taker to a module that is not appropriate for his/her ability level will lead to more errors in ability estimation due to the fact that the test taker has less information at the true ability level. In addition, since the test taker is routed to the wrong module, the questions in the module that are not at the test taker's true ability level (harder or easier questions) will affect the test taker psychologically (Kim & Moses, 2014).

One of the main problems in the evaluation of the data obtained from educational tests is the transformation of categorical data coded as true-false in the form of 1-0 into a continuous ability scale (Erkuş, 2013). Therefore, in the transition from categorical data to a continuous ability scale, many items are needed to estimate the ability accurately. When ability is estimated based on a few items, it is possible that the ability estimate may be lower or higher than the true ability level (Eroğlu & Kelecioğlu, 2015). As a result, it seems impossible to reduce the probability of misrouting to zero in CAT, MST or other adaptive testing approaches. However, it can be argued that it may be possible to reduce the probability of misrouting with some changes in the test design.

Since the MST is an adaptive testing approach, an ability is estimated after each module and the test taker is routed to one of the modules in the next stage according to this estimated ability level.

Therefore, the fewer the number of items in the applied module, the higher the standard error of the estimated ability will be and the estimated ability may be lower or higher than the true ability. In this case, it is stated that increasing the number of items in the routing module is a variable that reduces misrouting (Kim et al., 2015). On the other hand, according to the amount of item information, designing the routing module in a wide ability range may be a variable that reduces misrouting compared to designing it in a narrow ability range. Again, the fact that the MST design consists of more than one stage is considered as a variable that may reduce misrouting. Also, as the test length increases, the standard error of estimation decreases and measurement accuracy increases. Therefore, it is thought that misrouting will decrease as the test length increases. Finally, measurement accuracy should also be taken into account while evaluating the misrouting rate. For example, the misrouting rate in LT is zero (because there is no routing), but the measurement accuracy is also low compared to adaptive tests. Therefore, this study investigated the optimal design that offers both high measurement accuracy and low misrouting rate.

Purpose and Significance of the Study

The purpose of this study is to compare the misrouting rates of different MST designs. Adaptive testing approaches based on Item Response Theory are utilized in large-scale international assessments such as CPA, GRE, PIAAC, TIMSS and PISA. There are two critical effects of misrouting in CAT design: First, in the case of misrouting, item information is low as the module presented will be far from the test taker's true ability level. Therefore, misrouting will reduce measurement accuracy. Secondly, in the case of misrouting, the test taker will face questions that are harder or easier than his/her ability level. This will affect the test taker psychologically during the test (Kim & Moses, 2014). Considering these two critical effects, the importance of reducing misrouting emerges. Studies in the literature show that the CAT offers better measurement accuracy than the MST by a small amount (Patsula, 1999; Wang, 2017). The measurement accuracy of CAT is thought to be quite similar to that of MST, reducing the misrouting rate. The idea that MST can provide similar measurement accuracy results to CAT shows the importance of this study. In addition, it is seen that the studies on misrouting in the literature are quite limited (Kim & Moses, 2014; Kim et al. 2015; Karatoprak Erşen & Lee, 2023). Since the effects of different MST designs, different module lengths and different routing module designs on the misrouting rate have not been included in previous studies, it is thought that this study will provide important suggestions to researchers and practitioners.

Research Questions

In this study, it is aimed to investigate the following research questions :

1. To what extent do misrouted rates and measurement accuracy of misrouted ones vary according to different MST designs?
2. To what extent do misrouted rates and measurement accuracy of misrouted ones vary according to different test lengths?
3. To what extent do misrouted rates and measurement accuracy of the misrouted ones vary according to different module lengths?
4. To what extent do misrouted rates and measurement accuracy of misrouted ones vary according to different routing module designs?

5. What is the distribution of the ability levels of misrouted test takers?

Literature Review

Kim and Moses (2014) examined the effect of misrouting in a 1-3 MST design with one routing module and two stages. The results of this study indicate that misrouting is concentrated on the intersections of the module information functions and thus misrouting can have an impact on test takers and has a minimal impact on measurement accuracy.

Kim, Moses, and Yoo (2015) compared different ability estimation methods under 1-3 MST design. The results of this study report that as the length of the routing module increases, misrouting decreases.

Karatoprak Erşen and Lee (2023), on the other hand, compared different item calibration methods under 1-3 MST design and reported that the misrouting rate increased as the length of the routing module decreased.

Method

In this study, the effects of the design and components of the MST on test takers' routed to the wrong module were examined. The data in the study were generated by simulation method and comparisons were made based on different conditions. Simulation studies are known as computer experiments that involve the generation and analysis of data by random sampling with probability distributions. These experiments are conducted to compare the performance of statistical methods under certain scenarios and conditions. In the field of psychometrics, simulation studies have an important place in the evaluation of methods, development and comparison of new methods (Morris et al., 2019). Since it is very difficult to perform all the conditions considered in this study on real data, simulation method is preferred. Since this study aims to reveal the misrouting situation under different designs and components in MST, it can be claimed to be descriptive research (Fraenkel et al., 2012).

The simulation study was conducted by varying three different MST designs (1-3, 1-2-3 and 1-3-3), three different test lengths (12, 24 and 36), three different module length distributions (Long-Short [L-S], Medium-Medium [M-M] and Short-Long [S-L]) and two different Routing Module Designs (Wide Ability Range [-0.5, 0.0, 0.5] and Narrow Ability Range [0.0]). All conditions are crossed with each other. Therefore, 3x3x3x2=54 conditions were analyzed in this study. For each condition, 100 replications were made and analyzed. In the study, open-source R software was used to generate and analyze the data. "Rmst" (Luo, 2018) was used for automatic test assembly, "mstR" (Magis et al., 2017) was used for MST analysis, and codes written by the researchers were used to detect misrouting. Details about the simulation study are presented in the following sections.

Item and Ability Parameters

Within the scope of the research, the parameters of 200 items based on the 3PL model were generated by considering the distributions in the literature (Feinberg & Rubright, 2016; Mooney, 1997). Item discrimination parameters were obtained from log-normal distribution $a \sim lnN(0.3,\ 0.2)$, item difficulty parameters were obtained from normal distribution $b \sim N(0,\ 1)$, item guessing parameters were obtained from beta distribution $c \sim Beta(8,\ 32)$. Ability parameters were generated from the

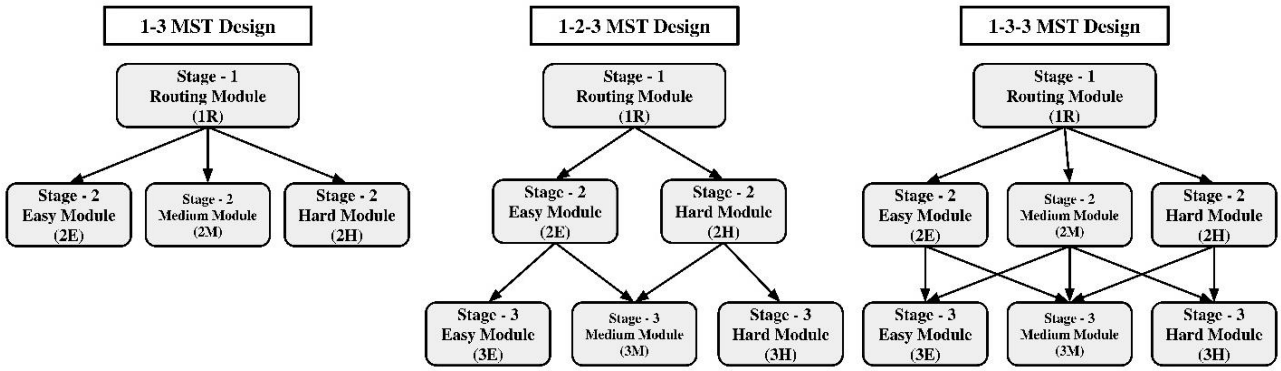
normal distribution $\theta \sim N(0, 1)$ for 1000 test takers. EAP method was used for ability estimation.

MST Design

Three different MST designs were used in this study, namely 1-3, 1-2-3 and 1-3-3. These MST designs are shown in Figure 2.

Figure 2

1-3, 1-2-3 ve 1-3-3 MST Design



Test Length

In this study, three different test lengths, 12, 24 and 36, are included under the test length condition.

Module Length

In this study, there are three different module lengths as Short-Long [S-L], Medium-Medium [M-M] and Long-Short [L-S]. In the S-L module length, the modules in the first stage are short, while the subsequent modules are of long test length. In the M-M module length, all module lengths are evenly distributed. In the L-S module length, the routing module is long while the next module is short. Table 1 shows the distribution of the number of items in the modules according to the two- and three-stage MST designs and test lengths.

Table 1

Distribution of Module Lengths by Test Lengths

MST Design	Test Length	Long-Short [L-S]	Medium-Medium [M-M]	Short-Long [S-L]
1-3	12	8-4	6-6	4-8
	24	16-8	12-12	8-16
	36	24-12	18-18	12-24
1-2-3 and 1-3-3	12	6-3-3	4-4-4	3-3-6
	24	12-6-6	8-8-8	6-6-12
	36	18-9-9	12-12-12	9-9-18

Ability Distribution of Modules and Routing Module

Table 2 shows the ability levels at which the amount of module information can be maximized when assembling MST modules and panels.

Table 2

Ability Levels at which Module Information is Maximized

MST Design	Stage 1	Stage 2	Stage 3
1-3	Wide : θ = 0 Narrow : θ = (-0.5, 0, 0.5)	θ = (-1, 0, 1)	
1-2-3	Wide : θ = 0 Narrow : θ = (-0.5, 0, 0.5)	θ = (-0.5, 0.5)	θ = (-1, 0, 1)
1-3-3	Wide : θ = 0 Narrow : θ = (-0.5, 0, 0.5)	θ = (-0.75, 0, 0.75)	θ = (-1, 0, 1)

As seen in Table 2, there are two designs for the routing module: wide and narrow. In the wide routing module design, it is aimed to maximize the module information level in the range ϑ = (-0.5, 0.0, 0.5), while in the narrow routing module design, it is aimed to maximize the module information level at θ = 0.

When Table 2 is analyzed, it is seen that the module information levels of the last stages of the MST designs are designed to be the same in all MST designs with θ = (-1, 0, 1). The modules in the second stages of the three-stage 1-2-3 and 1-3-3 MST designs were designed with θ = (-0.5, 0.5) and θ = (-0.75, 0, 0.75), respectively.

Detection of Misrouting

In the stage following the routing module - according to the MFI routing rule - a test taker should be routed to the module that offers the maximum information according to his/her true ability level among the modules. Therefore, in this study, test takers who were not routed to the module that offered the maximum information based on their true ability level were labeled as misrouting. For example, in the 1-3 MST design, a test taker with an ability level of θ = 1.48 is expected to be routed to the difficult module after the routing module. However, if the test taker's ability estimation obtained from the test taker's responses to the routing module is low, and the test taker is routed to one of the easy or medium difficulty modules, this is labeled as misrouting. In this study, 1-3, 1-2-3 and 1-3-3 MST designs are considered. The detection of misrouting was performed only on the last stages of all MST designs. The limitation of this study is that the determination of misrouting is based only on the last stage. To detect misrouting, the intersection points of the item information functions of the modules in the last stage were determined and test takers who were not routed to the module with the highest item information according to these points were labeled as misrouting. The number of test takers who were misrouted out of 1000 test takers in all conditions was calculated and reported as percentage (%).

Data Analysis

To analyze the data obtained from different MST designs and components, the relationships between true and estimated ability parameters were interpreted by calculating Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) values. The formulas for RMSE and MAE values are given below.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(\widehat{\theta_i}-\theta_i)^2}{n}}$$

$$\text{MAE} = \frac{\sum_{i=1}^{n}|\widehat{\theta_i}-\theta_i|}{n}$$

In the study, RMSE and MAE values were calculated and presented separately for all participants, for the correctly routed test takers and for the misrouted test takers in order to determine the change in the measurement accuracy of the misrouted test takers. Finally, the percentage (%) of misrouted test takers was calculated and presented.

Results

In this section, firstly, all the results obtained from the 54 conditions analyzed in the study are presented in Table 3 and Table 4. Then, the findings based on the research questions are presented.

In Table 3 and Table 4, measurement accuracy findings obtained from all test takers are presented with RMSE and MAE values. In addition, RMSE and MAE values from only the misrouted test takers are presented in the columns of MR-RMSE and MR-MAE. Similarly, the RMSE and MAE values of the correctly routed test takers are presented in the CR-RMSE and CR-MAE columns.

Table 3

Results from All Conditions (1-3 MST Design)

Design	Test Length	Routing Module Design	Modul Length	RMSE	MR-RMSE	CR-RMSE	MAE	MR-MAE	CR-MAE	Rate of Misrouted (%)
1-3	12	Wide	L-S	0.356	0.399	0.343	0.274	0.320	0.261	21.4
			M-M	0.347	0.388	0.332	0.268	0.303	0.256	25.2
			S-L	0.338	0.391	0.316	0.261	0.299	0.246	27.2
		Narrow	L-S	0.356	0.394	0.344	0.273	0.311	0.262	22.6
			M-M	0.348	0.384	0.333	0.267	0.294	0.257	26.5
			S-L	0.350	0.406	0.324	0.267	0.303	0.252	29.1
	24	Wide	L-S	0.265	0.273	0.263	0.203	0.222	0.199	16.8
			M-M	0.254	0.277	0.248	0.195	0.218	0.190	18.6
			S-L	0.247	0.269	0.240	0.190	0.210	0.184	21.6
		Narrow	L-S	0.269	0.265	0.270	0.206	0.211	0.204	20.3
			M-M	0.258	0.264	0.256	0.198	0.207	0.196	21.7
			S-L	0.255	0.280	0.246	0.195	0.214	0.189	23.4
	36	Wide	L-S	0.230	0.220	0.232	0.175	0.178	0.174	15.3
			M-M	0.215	0.218	0.215	0.165	0.174	0.164	17.6
			S-L	0.211	0.222	0.208	0.162	0.176	0.159	19.0
		Narrow	L-S	0.233	0.217	0.236	0.177	0.173	0.178	17.6
			M-M	0.218	0.219	0.218	0.167	0.174	0.166	18.0
			S-L	0.214	0.225	0.211	0.165	0.176	0.162	21.4

* MR: Misrouting

* CR: Correctly Routing

Table 4

Results from All Conditions (1-2-3 and 1-3-3 MST Designs)

Design	Test Length	Routing Module Design	Modul Length	RMSE	MR-RMSE	CR-RMSE	MAE	MR-MAE	CR-MAE	Rate of Misrouted (%)
1-2-3	12	Wide	L-S	0.345	0.404	0.330	0.266	0.333	0.251	18.7
			M-M	0.341	0.405	0.322	0.263	0.324	0.248	20.1
			S-L	0.337	0.410	0.312	0.262	0.323	0.244	22.8
		Narrow	L-S	0.341	0.386	0.329	0.262	0.314	0.250	19.5
			M-M	0.336	0.401	0.318	0.258	0.320	0.243	19.5
			S-L	0.325	0.404	0.300	0.252	0.320	0.234	21.2
	24	Wide	L-S	0.259	0.285	0.255	0.199	0.235	0.193	14.2
			M-M	0.245	0.282	0.238	0.189	0.233	0.182	13.9
			S-L	0.244	0.292	0.234	0.189	0.235	0.180	16.0
		Narrow	L-S	0.258	0.288	0.252	0.198	0.239	0.191	14.2
			M-M	0.247	0.281	0.241	0.191	0.229	0.184	14.4
			S-L	0.247	0.293	0.237	0.191	0.234	0.183	16.6
	36	Wide	L-S	0.224	0.223	0.224	0.170	0.180	0.168	14.0
			M-M	0.210	0.223	0.207	0.161	0.183	0.158	12.4
			S-L	0.204	0.232	0.199	0.159	0.189	0.154	13.4
		Narrow	L-S	0.221	0.224	0.220	0.168	0.183	0.165	13.3
			M-M	0.211	0.228	0.209	0.162	0.188	0.158	12.5
			S-L	0.204	0.237	0.199	0.160	0.193	0.154	13.8
1-3-3	12	Wide	L-S	0.345	0.413	0.327	0.267	0.338	0.249	19.6
			M-M	0.336	0.390	0.321	0.260	0.315	0.245	20.3
			S-L	0.347	0.396	0.329	0.268	0.312	0.254	24.2
		Narrow	L-S	0.342	0.395	0.327	0.264	0.319	0.250	19.9
			M-M	0.331	0.373	0.319	0.255	0.298	0.244	21.1
			S-L	0.331	0.386	0.313	0.255	0.302	0.241	22.2
	24	Wide	L-S	0.262	0.273	0.260	0.200	0.226	0.195	15.0
			M-M	0.247	0.274	0.243	0.190	0.223	0.184	14.7
			S-L	0.245	0.267	0.240	0.188	0.214	0.183	17.4
		Narrow	L-S	0.262	0.281	0.258	0.200	0.229	0.194	15.4
			M-M	0.252	0.271	0.248	0.193	0.219	0.188	15.6
			S-L	0.249	0.279	0.242	0.190	0.220	0.184	16.9
	36	Wide	L-S	0.224	0.228	0.223	0.170	0.187	0.167	13.3
			M-M	0.211	0.224	0.209	0.161	0.184	0.158	12.8
			S-L	0.209	0.224	0.206	0.160	0.181	0.157	14.4
		Narrow	L-S	0.218	0.234	0.216	0.167	0.192	0.163	13.1
			M-M	0.211	0.227	0.209	0.162	0.186	0.158	13.6
			S-L	0.207	0.229	0.203	0.160	0.185	0.155	14.5

* MR: Misrouting

* CR: Correctly Routing

When Table 3 and Table 4 are examined, it can be seen that the RMSE value varies between 0.204 and 0.356 and MAE values vary between 0.161 and 0.274 in all conditions. MR-RMSE values vary between 0.217 and 0.413 and CR-RMSE values vary between 0.173 and 0.338. MR-MAE values range between 0.199 and 0.344; CR-MAE values range between 0.154 and 0.262. According to both RMSE and MAE values, it is seen that the measurement accuracy of the misrouted participants is lower than the correctly routed participants. Therefore, it can be stated that misrouting is a threat to MST designs in terms of measurement accuracy.

The misrouted rates in the tables are between 12.4% and 29.1%. This finding indicates that the changes to be made on the design and components of the MST are effective in reducing misrouted.

In this section of the study, the findings related to the research questions are presented. The tables in the research questions were obtained by averaging the conditions in Table 3 and Table 4.

Findings Related to the First Research Question

Table 5 shows the misrouting rate and measurement accuracy findings for different MST designs.

Table 5

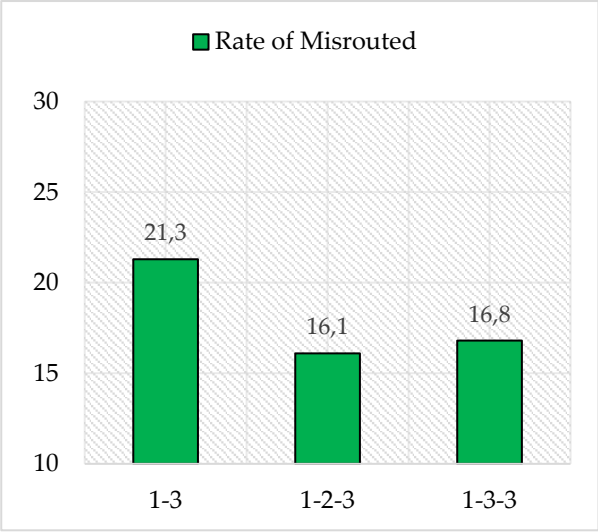
Findings According to Different MST Designs

MST Design	RMSE	MR-RMSE	CR-RMSE	MAE	MR-MAE	CR-MAE	Rate of Misrouted (%)
1-3	0.275	0.295	0.269	0.212	0.231	0.206	21.3
1-2-3	0.266	0.305	0.257	0.205	0.248	0.197	16.1
1-3-3	0.268	0.298	0.261	0.206	0.241	0.198	16.8

The misrouted rate findings in Table 5 are visualized and presented in Figure 3.

Figure 3

Misrouting Rates According to Different MST Designs



As seen in Figure 3, while the misrouting rate is 21.3% in the 1-3 MST design, it is 16.1% and 16.8% in the 1-2-3 and 1-3-3 MST designs, respectively. This finding shows that the misrouting rate decreases in three-stage MST designs compared to two-stage MST designs. It can also be claimed that the misrouting rate of the 1-2-3 MST design is slightly lower than the 1-3-3 MST design.

The RMSE and MAE measurement accuracy findings in Table 5 are visualized and presented in Figure 4.

Figure 4

Measurement Accuracy Results According to Different MST Designs

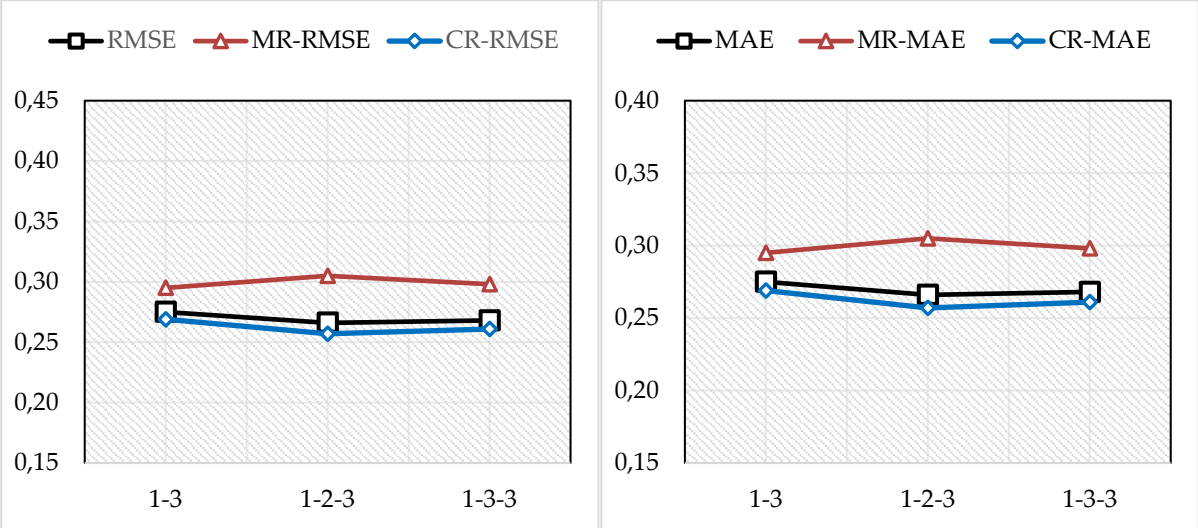


Figure 4 shows that while the RMSE value obtained from the 1-3 MST design is 0.275, the RMSE values obtained from the 1-2-3 and 1-3-3 MST designs are 0.266 and 0.268, respectively. Similarly, while the MAE value obtained from the 1-3 MST design is 0.212, the MAE values obtained from the 1-2-3 and 1-3-3 MST design are 0.205 and 0.206, respectively. These findings indicate that the measurement accuracy of the 1-2-3 and 1-3-3 MST designs is better than the 1-3 MST design. In addition, the measurement accuracy findings of the 1-2-3 and 1-3-3 MST designs are quite similar and it can be argued that the measurement accuracy of the 1-2-3 MST design is better by a smaller difference. In addition, it is seen that the RMSE and MAE values of the test takers who were misrouted in all MST designs were higher than those who were correctly routed. This finding shows that the measurement accuracy of the misrouted test takers is low in all MST designs.

Findings Related to the Second Research Question

Misrouting rate and measurement accuracy findings for different test lengths are shown in Table 6.

Table 6

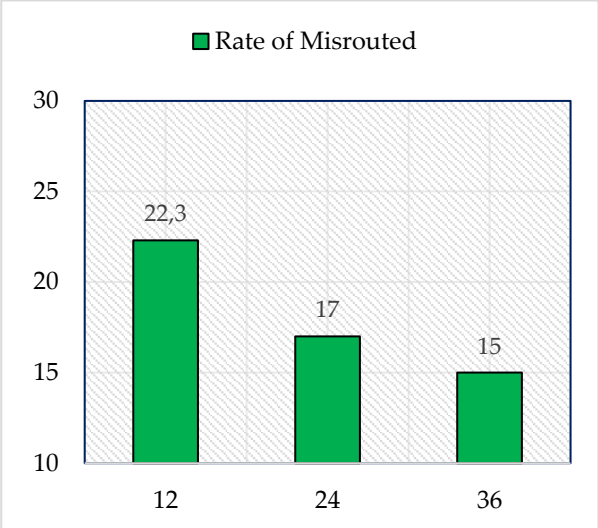
Findings According to Different Test Length

Test Length	RMSE	MR-RMSE	CR-RMSE	MAE	MR-MAE	CR-MAE	Rate of Misrouted (%)
12	0.342	0.396	0.324	0.263	0.314	0.249	22.3
24	0.254	0.277	0.248	0.195	0.223	0.189	17.0
36	0.215	0.225	0.214	0.165	0.182	0.162	15.0

The misrouting rate findings in Table 6 are visualized and presented in Figure 5.

Figure 5

Misrouting Rates for Different Test Lengths



As can be seen in Figure 5, while the misrouting rate was 22.3% in 12 test lengths, it was 17.00% and 15.00% in 24 and 36 test lengths. This finding shows that the misrouting rate decreases as the test length increases. In addition, while the misrouting rate decreased by 5.3% units from 12 to 24 test lengths, it decreased by 2.0% units from 24 to 36 test lengths. Therefore, it should be noted that the misrouting rate does not decrease linearly as the test length increases.

The RMSE and MAE measurement accuracy findings in Table 6 are visualized and presented in Figure 6.

Figure 6

Measurement Accuracy Findings for Different Test Lengths

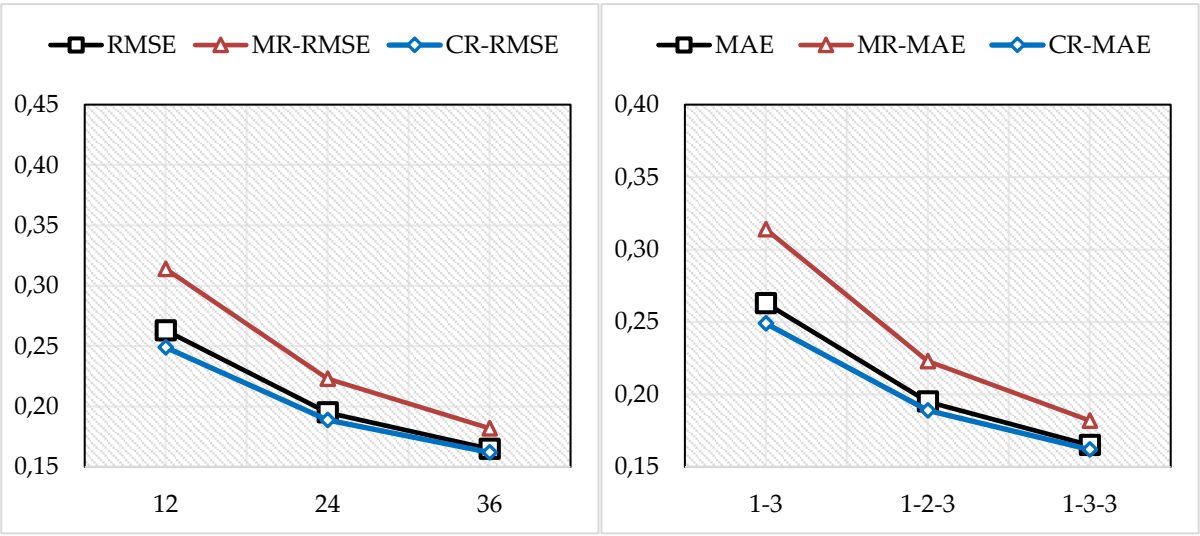


Figure 6 shows that while the RMSE value obtained from 12 test lengths is 0.342, the RMSE values obtained from 1-2-3 and 1-3-3 MST designs are 0.254 and 0.215, respectively. Similarly, while the MAE value obtained from 12 test lengths is 0.263, the MAE values obtained from 24 and 36 test lengths are 0.195 and 0.165, respectively. These findings show that measurement accuracy increases as test length increases.

As seen in Figure 6, while the difference between RMSE-RMSE and RMSE-RMSE is 0.072 (0.396-0.324) at test length 12, this difference is 0.019 (0.277-0.248) and 0.014 (0.225-0.214) at test length 24 and 36, respectively. This shows that as the test length increases, the measurement accuracy of the misrouted test takers increases significantly. In other words, increasing test length decreases both the proportion of misrouted test takers and the ability estimation errors of misrouted test takers.

Findings Related to the Third Research Question

Measurement accuracy and misrouting findings according to different module lengths are shown in Table 7.

Table 7

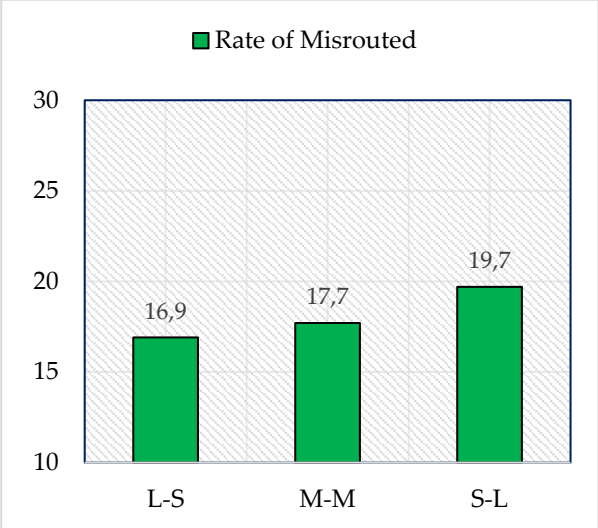
Findings According to Different Module Lengths

Modul Length	RMSE	MR-RMSE	CR-RMSE	MAE	MR-MAE	CR-MAE	Rate of Misrouted (%)
L-S	0.278	0.300	0.273	0.213	0.244	0.206	16.9
M-M	0.268	0.296	0.260	0.206	0.237	0.199	17.7
S-L	0.265	0.302	0.253	0.204	0.238	0.195	19.7

The misrouting rate findings in Table 7 are visualized and presented in Figure 7.

Figure 7

Misrouting Rates According to Different Module Lengths



As can be seen in Figure 7, while the misrouting rate is 16.9% for the L-S module length, it is 17.7% and 19.7% for the M-M and S-L module lengths, respectively. This finding shows that the misrouting rate decreases as the length of the routing module increases.

The RMSE and MAE measurement accuracy findings in Table 7 are visualized and presented in Figure 8.

Figure 8

Measurement Accuracy Results for Different Module Lengths

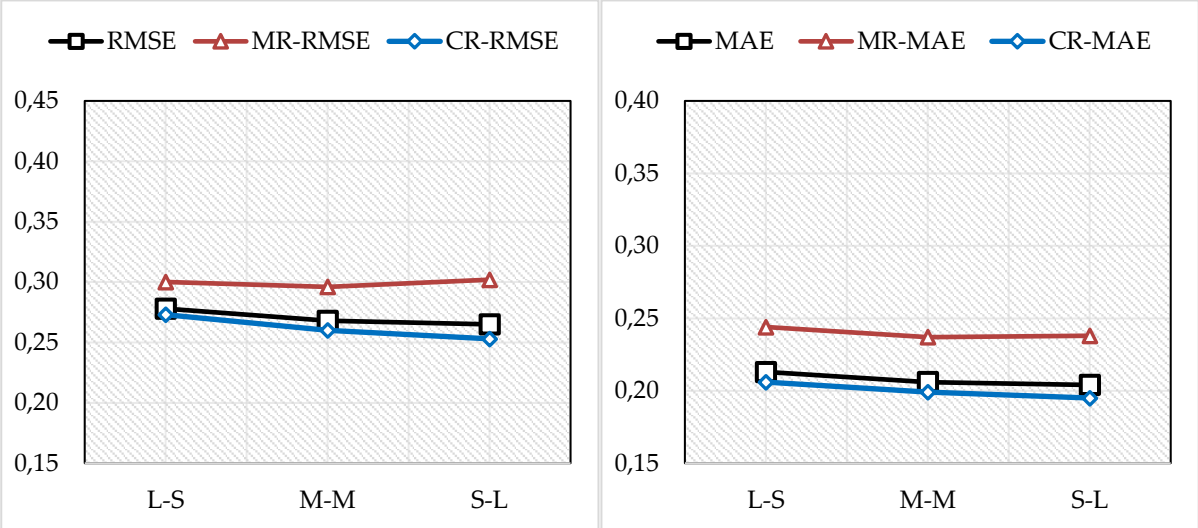


Figure 8 shows that the RMSE value obtained from the L-S module length is 0.278, while the RMSE values obtained from the M-M and S-L module lengths are 0.268 and 0.205, respectively. Similarly, the MAE value obtained from the L-S module length is 0.213, while the MAE values obtained from the M-M and S-L module lengths are 0.268 and 0.265, respectively. These findings show that the measurement accuracy decreases slightly as the routing module length increases. Therefore, as the routing module length increases, misrouting decreases and measurement accuracy decreases at the same time. At this point, it is clear that the module length should be decided by taking into account the measurement accuracy as well as the misrouting.

As seen in Figure 8, it can be stated that the measurement accuracy of the test takers who were misrouted in L-S, M-M and S-L module lengths was lower than the test takers who were correctly routed.

Findings Related to the Fourth Research Question

Table 8 shows the measurement accuracy and misrouting findings according to different routing module designs.

Table 8

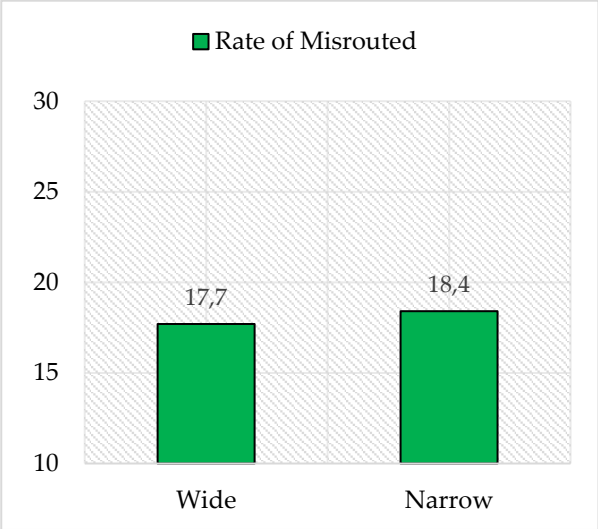
Findings According to Different Routing Module Designs

Routing Module Design	RMSE	MR-RMSE	CR-RMSE	MAE	MR-MAE	CR-MAE	Rate of Misrouted (%)
Wide	0.270	0.300	0.262	0.208	0.241	0.200	17.7
Narrow	0.270	0.299	0.262	0.208	0.238	0.200	18.4

The misrouting rate findings in Table 8 are visualized and presented in Figure 9.

Figure 9

Misrouting Rates According to Different Routing Module Designs



As seen in Figure 9, the misrouting rate of the wide routing module design is 17.7%, while the misrouting rate of the narrow routing module design is 18.4%. This finding shows that the wide routing module has a lower misrouting rate than the narrow routing module. In other words, the wide routing design reduces the misrouting rate compared to the narrow routing design.

The RMSE and MAE measurement accuracy findings in Table 8 are visualized and presented in Figure 10.

Figure 10

Measurement Accuracy Findings According to Different Routing Designs

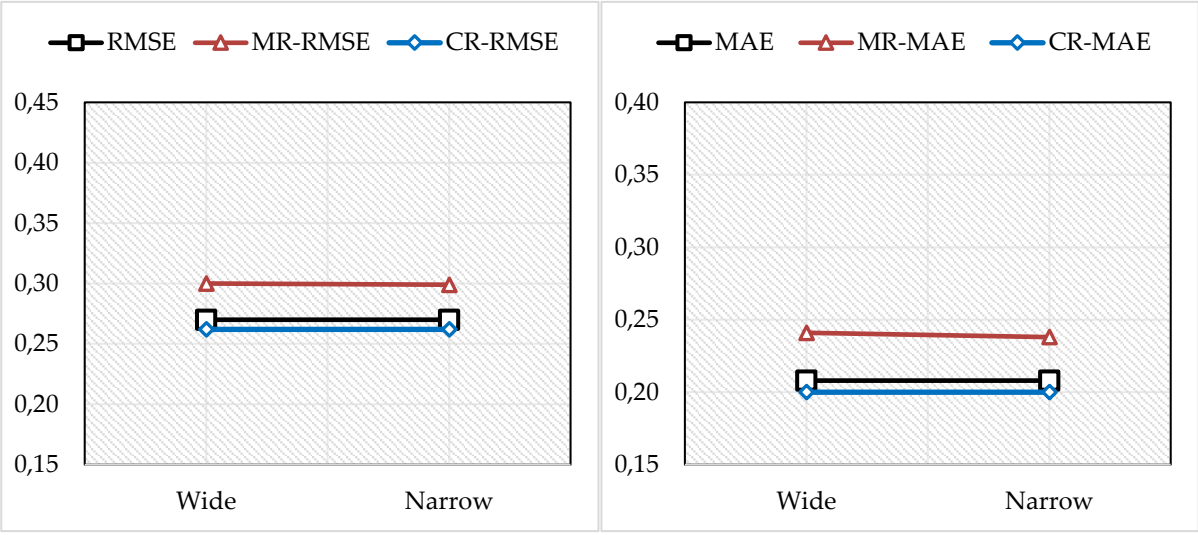


Figure 10 shows that the RMSE value obtained from the wide module design is 0.270, while the RMSE value obtained from the narrow module design is 0.270. Similarly, the MAE value obtained from the wide module design is 0.208, while the MAE value obtained from the narrow module design is 0.208. These findings show that the routing module design does not change the measurement accuracy.

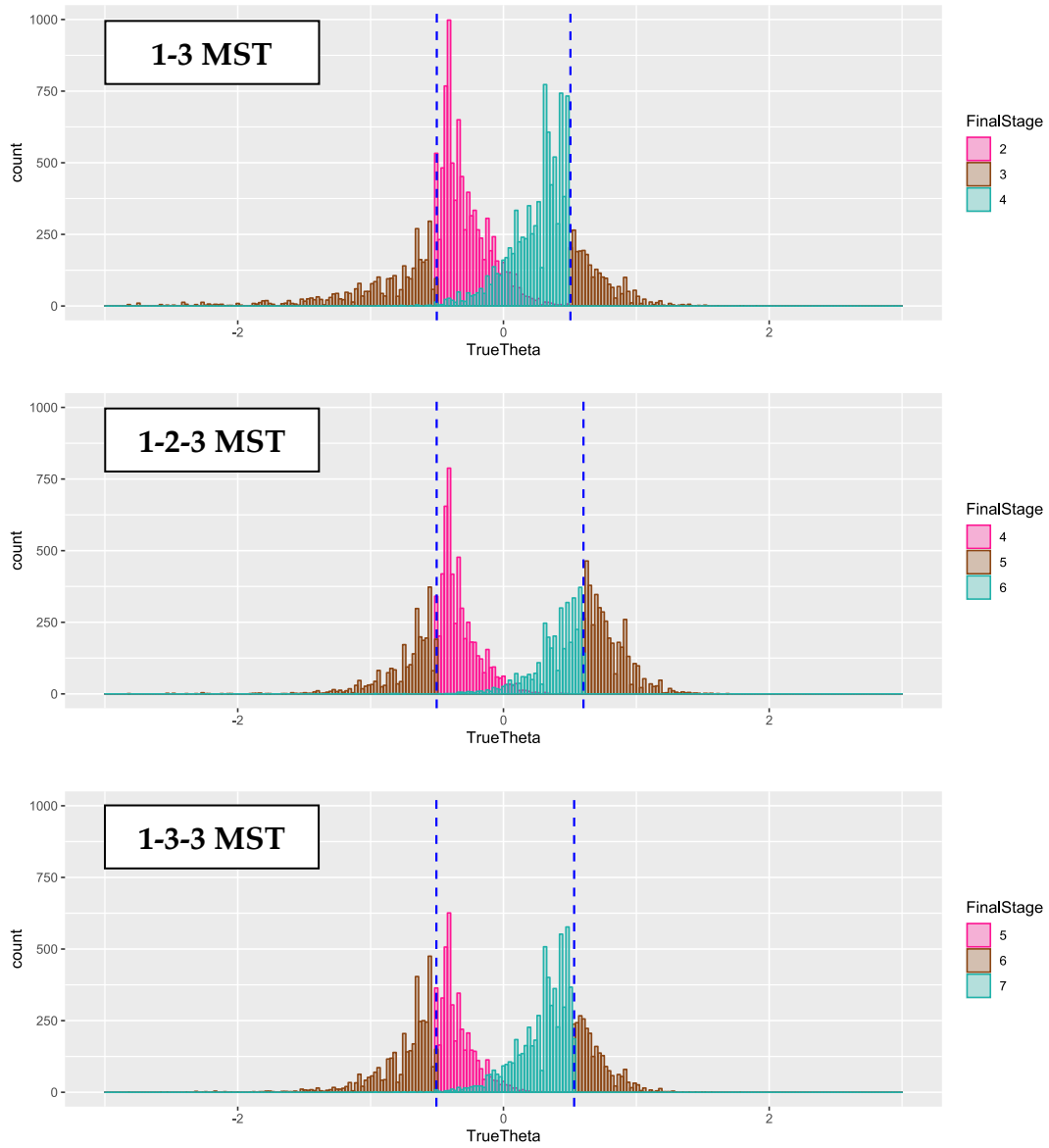
As seen in Figure 10, it can be stated that the measurement accuracy of the misrouted ones is lower than the measurement accuracy of the correctly routed ones in both wide and narrow routing module design.

Findings Related to the Fifth Research Question

The ability distribution of misrouted test takers is presented in Figure 11 below.

Figure 11

Ability Levels of the Misrouted



In Figure 11, the modules to which the misrouted test takers were routed in the last stage of the MST design according to their true ability levels are colored on the histogram graph. The blue dashed lines indicate the intersection points of the module information functions at the last stage. For example, the point -0.502 in the "1-3 MST" graph represents the intersection point of the information functions of easy (2E) and medium (2M) difficulty modules. The colors of the histogram bars indicate to which module the test takers at the true ability level were misrouted. For example, in the turquoise columns in the 1-3 MST design, misrouted test takers at the true ability level were redirected to module 4 (the difficult module [2H] in the second stage) in the second stage. Similarly, in the 1-3-3

MST design, the position of the pink columns indicates that test takers with the true ability level were directed to the easy module (3K) in the third stage.

As seen in Figure 11, the true ability level of misrouted test takers is concentrated at the intersection of module information functions and decreases towards the extremes of the ability scale. It can be said that the fact that the ability levels of the misrouted test takers are located in the middle part of the ability scale reduces the effect of misrouting on measurement accuracy.

Discussion

Computerized adaptive tests have displayed promising results in terms of higher test-taker engagement, higher motivation, and lower anxiety levels (Ling et al., 2017; Martin & Ladenzic, 2018). These advantages are based on the fact that adaptive tests provide test takers with items appropriate for their ability level (Şenel, 2021). Similarly, in MST, routing test takers to the most appropriate module for their ability level is important in terms of both measurement accuracy and test psychology.

In this study, the effects of different MST designs and different core components on misrouting were examined. According to the results, three-stage MST designs have lower misrouting rates than two-stage MST designs.

In terms of both measurement accuracy and misrouting rate, it was observed that the 1-2-3 and 1-3-3 MST designs offered better measurement accuracy, while the 1-3-3 MST design had lower measurement accuracy. There are many studies in the literature indicating that three-stage MST designs offer higher measurement accuracy than two-stage MST designs (Sari & Huggins-Manley, 2017; Zenisky et al., 2010). In addition, in this study, the 1-2-3 MST design was found to offer better measurement accuracy and lower misrouting rate than the 1-3-3 MST design.

According to the test length findings, it was observed that the misrouting rate decreased as the test length increased. In addition to the decrease in the misrouting rate, the increase in test length also significantly increased the measurement accuracy of the misrouted. However, the increase in test length did not linearly increase the measurement accuracy and similarly did not linearly decrease the misrouting rate. Therefore, to determine the test length, the targeted efficiency may not be obtained from high test lengths. In a similar to the findings of this study, there are many studies in the literature indicating that increasing test length will increase measurement accuracy (Thompson & Weiss, 2011; Şahin, 2020; Erdem Kara & Doğan, 2022).

According to the module lengths, it was concluded that the misrouting rate decreased as the length of the routing module increased. However, increasing the length of the routing module decreases the length of the other stages, which are adaptive stages, which leads to a decrease in measurement accuracy. In this respect, it is recommended to use the most appropriate module length by considering both misrouting rates and measurement accuracy together. The findings of the current study indicating that increasing the length of the routing module reduces the misrouting are similar to the findings of the study by Kim et al. (2015) and Karatoprak Erşen and Lee (2023).

It was concluded that the design of the routing module in the wide ability range reduced the misrouting rate compared to the narrow ability range. In addition, it was seen that designing the

routing module in the wide and narrow ability range did not make a difference in terms of measurement accuracy. Therefore, it is recommended to design the routing module in a wide range. Similar to the results of the current study, Cai et al. (2021) state that misrouting decreases when the amount of module information of the routing module increases. In addition, Şahin (2020) concluded that the change in the difference between the b parameters does not make a significant difference on the measurement accuracy. This finding of Şahin (2020) supports the results of the current study.

It is found that the ability levels of the misrouted participants are concentrated around the intersection points of the module information functions. Besides, the measurement accuracy of the misrouted participants is lower than that of the correctly routed participants, but since the measurement accuracy is not excessively different, it can be said that misrouting does not pose a significant threat. The reason why the measurement accuracy of the misrouted participants was not excessively different may be that these participants were concentrated in the middle points of the ability scale. In a similar vein the results of this study, Kim and Moses (2014) state that misrouted participants are concentrated on the cut-off points of cross-module information functions, so the effect of misrouted participants on measurement accuracy is minimal. Similarly, Karamese (2022) states that misrouted participants are concentrated around the cut-off score.

The results of this study show that misrouting in the MST slightly decreases measurement accuracy. In order to reduce the misrouting rate, practitioners can make changes to the design and basic components of the MST. According to the results of this study, increasing the number of stages of the MST design, increasing the test length, and designing a large routing module will reduce the misrouting rate. Regarding module lengths, it can be suggested that module lengths should be equal in terms of both misrouting rate and measurement accuracy. In future studies, it may be recommended to design studies on the effects of routing rules, ability estimation methods and item pool characteristics on the misrouting rate.

Ethics Committee Approval: The data in the study were generated in a computer simulation environment and no data were collected from any human or live participant.

Author Contributions: The authors contributed equally to all steps of the research process of this study.

Conflict of Interest: The authors declare that there are no potential conflicts of interest.

Uluslararası Türk Eğitim Bilimleri Dergisi,
2023, 11 (21), 549-587
https://doi.org/10.46778/goputeb.1267319

ISSN: 2148-2314

Bireyselleştirilmiş Çok Aşamalı Testlerde Test Tasarımının Yanlış Yönlendirmeye Etkisi*

Mahmut Sami Yiğiter[a] Nuri Doğan[b]

[a] *Öğr. Gör. Dr., Ankara Sosyal Bilimler Üniversitesi, Ankara, Türkiye, mahmutsamiyigiter@gmail.com*

[b] *Prof. Dr., Hacettepe Üniversitesi, Ankara, Türkiye, nuridogan2004@gmail.com*

ÖZET

Bireyselleştirilmiş Çok Aşamalı Testler (BÇAT), test katılımcısının yetenek düzeyine göre önceden birleştirilmiş panel üzerinde aşama ve modülleri tamamlayarak ilerlediği bireyselleştirilmiş bir test yaklaşımıdır. Test katılımcısı, her aşamada modüle verdiği yanıtlara göre ileriki aşamada bir modüle yönlendirilir. Katılımcının ilerleyen aşamalarda yetenek düzeyine en uygun modüle yönlendirilmesi beklenir. Eğer ki katılımcı yetenek düzeyine uygun modüle yönlendirilmiyorsa yanlış yönlendirmeden bahsedilebilir. Yanlış yönlendirmenin hem ölçme kesinliğini hem de katılımcının sınav psikolojisini etkilediği düşünülmektedir. Yanlış yönlendirmeden tamamıyla kurtulmak çok güç olsa da BÇAT tasarımının temel bileşenleri ile azaltılabileceği varsayılmaktadır. Bu çalışmanın amacı, farklı BÇAT tasarımlarına göre yanlış yönlendirme düzeylerinin belirlenmesi ve test tasarımında yapılacak değişimlerin yanlış yönlendirme düzeyine etkilerinin araştırılmasıdır. Yanlış yönlendirmeyi etkileyeceği düşünülen temel bileşenler olarak BÇAT test tasarımı [1-3, 1-2-3, 1-3-3], yönlendirme modülü tasarımı [Dar, Geniş], test uzunluğu [12, 24, 36] ve modül uzunluğu [U-K, O-O, K-U] değişimlenmiştir. Mevcut durumu ortaya koymayı hedefleyen bu çalışma, betimsel araştırma türünde olup simülasyon yöntemi ile gerçekleştirilmiştir. Araştırma sonuçları, BÇAT tasarım ve bileşenlerinin yanlış yönlendirmeyi azaltmada etkili olabileceğini göstermektedir. Üç aşamalı BÇAT tasarımları iki aşamalı BÇAT tasarımına göre daha düşük yanlış yönlendirme ve daha yüksek ölçme kesinliği sunmaktadır. Ayrıca, test uzunluğunun arttırılması, yönlendirme modülünün geniş yetenek aralığında tasarlanması yanlış yönlendirme oranını düşürmektedir. Ölçme kesinliği sonuçlarına göre yanlış yönlendirilenlerin ölçme kesinliğinin düşük olmasının yanında, yanlış yönlendirmenin genel olarak BÇAT'ta önemli bir sorun oluşturmadığı ifade edilebilir. Yanlış yönlendirilenlerin yetenek düzeylerinin bitişik modüllerin modül bilgi fonksiyonlarının kesişim noktalarında ve genellikle yetenek ölçeğinin ortalarında yoğunlaştığı sonucuna ulaşılmıştır.

MAKALE BİLGİSİ

Makale Türü
Araştırma

Makale Geçmişi
Gönderim tarihi: 18.03.2023
Kabul tarihi: 11.04.2023

Anahtar Kelimeler
Bireyselleştirilmiş Çok Aşamalı Testler, Yönlendirme, Yanlış Yönlendirme, Uyarlanabilir Testler, Ölçme Kesinliği

Atıf Bilgisi: Yiğiter, M. S. ve Doğan, N. (2023). Bireyselleştirilmiş çok aşamalı testlerde test tasarımının yanlış yönlendirmeye etkisi. *Uluslararası Türk Eğitim Bilimleri Dergisi, 11* (21), 549-587.

Sorumlu yazar: Mahmut Sami Yiğiter, e-posta: mahmutsamiyigiter@gmail.com

* Bu çalışma, EPODDER'in 21-23 Eylül 2022 tarihlerinde İzmir'de düzenlediği 8. Uluslararası Eğitimde ve Psikoloji'de Ölçme ve Değerlendirme Kongresi'nde sözlü olarak sunulmuştur.

Giriş

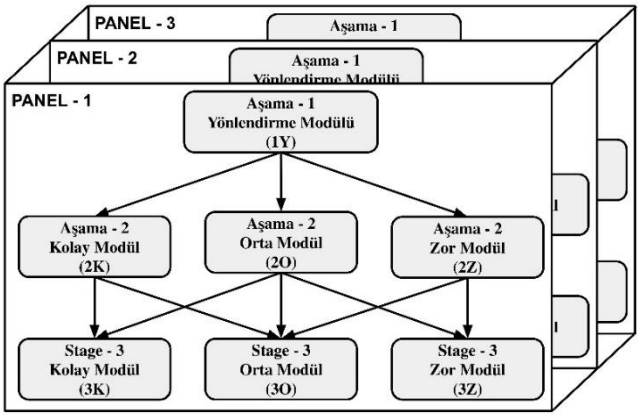
Eğitim alanında yapılan ölçmelerde öğrencilerin bilgi, beceri ve yeteneklerini ölçmede tüm sınav katılımcılarının aynı test formlarını aldığı Doğrusal Testler (DT) tarihsel süreçte en sık kullanılan yöntem olmuştur. Fakat son yarım yüzyıldır bilgisayar yazılım ve donanımlarındaki gelişmeler ve geliştirilen modellerle birlikte uyarlanabilir testler gelişim göstermiş ve popülerlik kazanmıştır. Uyarlanabilir testler, bir testte test katılımcısının bir önceki soruya verdiği yanıta göre her sorunun zorluk seviyesinin algoritma tarafından ayarlandığı bilgisayarlı bir test türüdür (Wainer ve diğerleri, 2001). Testin temel bileşenlerine göre farklı Uyarlanabilir Testler geliştirilmiştir. Daha kısa test uzunluğu ve daha etkili yetenek kestirimi ile Bireyselleştirilmiş Bilgisayarlı Testler (BBT), Dünya'da pek çok test uygulamasında benimsenmiştir (Khorramdel ve diğerleri, 2020). Özellikle son yıllarda dikkat çeken Bireyselleştirilmiş Çok Aşamalı Testler (BÇAT) ise DT ve BBT'nin avantajlı yönlerinden faydalanan ve dezavantajlarını azaltan uyarlanabilir bir test yaklaşımıdır (Zenisky ve diğerleri, 2010).

BÇAT, özellikle son yıllarda avantajlı yönlerinden dolayı pek çok geniş ölçekli sınavda kullanılmaktadır. ABD'de gerçekleştirilen Yeminli Mali Müşavirler (Certified Public Accountants - CPA) Sınavı, 2004'ten beri BÇAT'ı benimsemiştir (Breithaupt ve diğerleri, 2006). Lisansüstü Kayıt Sınavları'nın (Graduate Record Examinations - GRE) 2011'de gözden geçirilmiş bir versiyonunu BÇAT ile uygulamıştır. OECD, 2011 yılında Uluslararası Yetişkin Yeterliliklerini Değerlendirme Programında (PIAAC) BÇAT ile uluslararası geniş ölçekli bir testte uyarlanabilir bir yaklaşım kullanmıştır (Kirsch ve Lennon, 2017; Erdem Kara, 2022; Yiğiter ve Doğan, 2023). Ayrıca PISA 2018 ve TIMSS 2023 döngülerinde BÇAT benzeri çok aşamalı uyarlanabilir test yaklaşımı kullanılmaktadır. BÇAT'ın özellikle uluslararası geniş ölçekli testlerde kendisine yer edinmesinin temel sebebi DT gibi test uygulamasından önce test formlarının hazırlanabilir ve gözden geçirilebilir olması ve BBT gibi test katılımcısının kendi yetenek düzeyinde ilerleyebildiği uyarlanabilir bir test yaklaşımı olmasıdır.

BÇAT tasarımında yer alan modül ve paneller, madde havuzunda yer alan maddelerden test uygulamasından önce tasarlanıp birleştirilir. Madde havuzunda yer alan maddeler belirli özelliklerine göre birleştirilerek modüller oluşturulur. Modüllerin BÇAT tasarımı üzerinde birleşimi ile de aşamalar ve paneller oluşur. Şekil 1'de üç panelden oluşan 1-3-3 BÇAT tasarımı yer almaktadır.

Şekil 1

1-3-3 BÇAT Tasarımı



Şekil 1'de görüldüğü üzere 1-3-3 BÇAT tasarımı üç panelden oluşmakta ve panellerin her birinde üç aşama yer almaktadır. Sınava giren bir katılımcı, genellikle rastgele atandığı bir panel üzerinde ilerlerken öncelikle yönlendirme modülü adı verilen en üstteki modülü alır. Katılımcının modüle verdiği yanıtlara göre geçici yetenek düzeyi kestirilir ve katılımcı, yönlendirme kuralına göre ikinci aşamada yer alan kolay, orta ya da zor güçlük düzeyindeki üç modülden birine atanır. İlk aşamada olduğu gibi ikinci aşamanın sonunda da katılımcının verdiği yanıtlar da dikkate alınarak tekrar geçici yetenek düzeyi kestirilir ve kestirilen yetenek düzeyine göre katılımcı, üçüncü aşamadaki modüllerden birine atanır. Katılımcının üçüncü modülü yanıtlamasının ardından nihai yetenek düzeyi kestirilir ve test sonlandırılır.

Bireyselleştirilmiş Çok Aşamalı Testlerde Yanlış Yönlendirme

BBT, madde bazında bir uyarlanabilir test iken; BÇAT, modül bazında bir uyarlanabilir test yaklaşımıdır. BBT'nin madde bazında uyarlanabilir bir yaklaşım olması ile katılımcıya uygulanan her maddeden sonra yetenek düzeyi güncellenir. BÇAT'ta ise her modülden sonra yetenek düzeyi güncellenmektedir. Dolayısıyla uyarlama noktasının daha az olması sebebiyle yapılan simülasyon çalışmaları BÇAT'ın BBT'den bir miktar daha düşük ölçme kesinliği sunduğunu göstermektedir (Patsula, 1999; Wang, 2017). Bunun yanında BÇAT'ın uyarlama noktası sayısının daha az olması nedeniyle aşamalar arasındaki olası yanlış yönlendirmelerin etkisinin daha büyük olacağı düşünülmektedir. BÇAT'ta bir sınav katılımcısının yetenek düzeyine uygun olmayan bir modüle atanması yanlış yönlendirme (misrouting) olarak adlandırılmaktadır. Örneğin gerçek yetenek düzeyi -1.25 olan bir sınav katılımcısı 1-3 BÇAT tasarımı üzerinde birinci aşamada yönlendirme modülünü (1Y) alır. Ardından ikinci aşamada kolay (2K), orta (2O) ve zor (2Z) güçlükteki modüllerden kolay (2K) modüle atanması beklenir. Fakat katılımcı, yönlendirme modülüne verdiği yanıtlardan kestirilen geçici yetenek düzeyine göre ikinci aşamada orta (2O) ya da zor (2Z) modüle atanırsa yanlış yönlendirme gerçekleşir. Yanlış yönlendirmenin iki temel sebebi olabilir. Birincisi katılımcı, yönlendirme modülünde yetenek düzeyinin üzerinde veya altında bir performans sergilemiştir ve yetenek düzeyi gerçek yetenek düzeyinden yüksek veya düşük kestirilmiş olabilir. Bu durum genellikle olasılıksal olarak düşüktür. İkincisi ise test tasarımında yer alan bileşenler yanlış yönlendirmeye neden olmuş olabilir. Örneğin yönlendirme modülünde yer alan maddeler öğrencinin yetenek düzeyini başarılı bir şekilde kestirememiş olabilir veya yönlendirme kurallarında hatalar olabilir. Ortaya çıkan yanlış yönlendirme sonucunda sınav katılımcısının yetenek düzeyine uygun olmayan bir modüle atanması ile yanlış atanan modülün katılımcının gerçek yetenek düzeyinde daha az bilgi vermesi nedeniyle yeteneğin hatalı kestirilmesine neden olması beklenir. Ayrıca katılımcının yanlış modüle yönlendirilmesinden dolayı modülde yer alan sorular öğrencinin yetenek düzeyinde olmayacağından (daha zor veya daha kolay sorular) psikolojik olarak katılımcıyı etkileyecektir (Kim ve Moses, 2014).

Eğitsel testlerden elde edilen verinin değerlendirilmesinde temel sorunlardan biri doğru-yanlış olarak 1-0 şeklinde kodlanan kategorik verinin sürekli olan bir yetenek ölçeğine dönüştürülmesidir (Erkuş, 2013). Dolayısıyla kategorik veriden sürekli olan yetenek ölçeğine geçişte yeteneğin hassas biçimde kestirilmesi için oldukça fazla sayıda maddeye ihtiyaç vardır. Az sayıda madde ile yeteneğin kestirilmesi sonucunda yetenek kestiriminin gerçek yetenek düzeyinden düşük ya da yüksek olması muhtemeldir (Eroğlu ve Kelecioğlu, 2015). Sonuç olarak, BBT, BÇAT veya diğer uyarlanabilir test yaklaşımlarında yanlış yönlendirme olasılığını sıfıra indirmek imkânsız görünmektedir. Fakat test tasarımında yapılacak bazı değişiklikler ile yanlış yönlendirme olasılığını azaltmanın mümkün olabileceği düşünülmektedir.

BÇAT, uyarlanabilir bir test yaklaşımı olduğundan her modülden sonra yetenek kestirimi yapılır ve bu kestirilen yetenek düzeyine göre test katılımcısı sıradaki aşamadaki modüllerden birine yönlendirilir. Dolayısıyla uygulanan modüldeki madde sayısı ne kadar az ise kestirilen yeteneğin standart hatası o düzeyde artacak ve kestirilen yetenek, gerçek yetenekten düşük veya yüksek olabilecektir. Bu durumda yönlendirme modülünde yer alan madde sayısının arttırılmasının yanlış yönlendirmeyi azaltan bir değişken olduğu ifade edilmektedir (Kim ve diğerleri, 2015). Diğer taraftan, madde bilgi miktarına göre yönlendirme modülünün geniş yetenek aralığında tasarlanması, dar yetenek aralığında tasarlanmasına göre yanlış yönlendirmeyi azaltan bir değişken olabilir. Yine, BÇAT tasarımının birden fazla aşamadan oluşması yanlış yönlendirmeyi azaltacak bir değişken olarak düşünülmektedir. Son olarak, test uzunluğu arttıkça kestirimlerin standart hatası azalmakta ve ölçme kesinliği artmaktadır. Dolayısıyla test uzunluğu arttıkça yanlış yönlendirmenin azalacağı düşünülmektedir. Son olarak yanlış yönlendirme oranı değerlendirilirken ölçme kesinliği de dikkate alınmalıdır. Örneğin DT'de yanlış yönlendirme oranı sıfırdır (çünkü yönlendirme yok), fakat ölçme kesinliği de uyarlanabilir testlere göre düşüktür. Dolayısıyla bu araştırmada hem yüksek ölçme kesinliği hem de düşük yanlış yönlendirme oranı sunan optimal tasarım araştırılmıştır.

Araştırmanın Amacı ve Önemi

Bu araştırmanın amacı, farklı BÇAT tasarımlarının yanlış yönlendirme oranlarının karşılaştırılmasıdır. Dünya'da CPA, GRE, PIAAC, TIMSS ve PISA gibi uluslararası geniş ölçekli sınavlarda Madde Tepki Kuramı'na dayalı uyarlanabilir test yaklaşımlarından faydalanılmaktadır. BÇAT tasarımında yanlış yönlendirmenin iki kritik etkisi vardır: Birincisi, yanlış yönlendirme durumunda sunulan modül, katılımcının gerçek yetenek düzeyinin uzağında olacağından madde bilgisi düşüktür. Dolayısıyla yanlış yönlendirme ölçme kesinliğini azaltacaktır. İkinci olarak ise yanlış yönlendirme durumunda katılımcı yetenek düzeyinden daha zor ya da daha kolay sorular ile karşılaşacaktır. Bu durum katılımcıyı sınav esnasında psikolojik olarak etkileyecektir (Kim ve Moses, 2014). Bu iki kritik etki göz önüne alındığında yanlış yönlendirmenin azaltılmasının önemi ortaya çıkmaktadır. Literatürde yapılan çalışmalar BBT'nin BÇAT'tan az bir miktar fark ile daha iyi ölçme kesinliği sunduğunu göstermektedir (Patsula, 1999; Wang, 2017). Bu çalışmanın sonuçları ile yanlış yönlendirme oranı azaltılarak BÇAT'tan BBT'ye oldukça benzer ölçme kesinliği sonuçları elde edilebileceği düşüncesi bu çalışmanın önemini ortaya koymaktadır. Ayrıca yanlış yönlendirme üzerine literatürde yapılan çalışmaların oldukça kısıtlı olduğu görülmektedir (Kim ve Moses, 2014; Kim ve diğerleri, 2015; Karatoprak Erşen ve Lee, 2023). Bu çalışmada ele alınan farklı BÇAT tasarımlarının, farklı modül uzunluklarının ve farklı yönlendirme modülü tasarımlarının yanlış yönlendirme oranına etkilerine daha önce herhangi bir çalışmada yer verilmediğinden; bu çalışmanın araştırmacılara ve uygulayıcılara önemli öneriler sunacağı düşünülmektedir.

Araştırma Soruları

Araştırmada incelenen koşullar altında aşağıda yer alan beş araştırma sorusuna cevap aranmıştır:

1. Farklı BÇAT tasarımlarına göre yanlış yönlendirme oranları ve yanlış yönlendirilenlerin ölçme kesinliği ne düzeyde değişim göstermektedir?

2. Farklı test uzunluklarına göre yanlış yönlendirme oranları ve yanlış yönlendirilenlerin ölçme kesinliği ne düzeyde değişim göstermektedir?

3. Farklı modül uzunluklarına göre yanlış yönlendirme oranları ve yanlış yönlendirilenlerin

ölçme kesinliği ne düzeyde değişim göstermektedir?

4. Farklı yönlendirme modülü tasarımlarına göre yanlış yönlendirme oranları ve yanlış yönlendirilenlerin ölçme kesinliği ne düzeyde değişim göstermektedir?

5. Yanlış yönlendirilen katılımcıların yetenek düzeylerinin dağılımı nasıldır?

İlgili Araştırmalar

Kim ve Moses (2014), bir yönlendirme modülü ve iki aşama içeren 1-3 BÇAT tasarımında yanlış yönlendirmenin etkisini incelemiştir. Bu araştırmanın sonuçları, yanlış yönlendirmenin modül bilgi fonksiyonlarının kesişim bölgelerinde yoğunlaştığını ve dolayısıyla yanlış yönlendirmenin sınav katılımcılara etkisinin olabileceğini ve ölçme kesinliğin ise minimum düzeyde etkilediğini belirtmektedir.

Kim, Moses ve Yoo (2015), 1-3 BÇAT tasarımı altında farklı yetenek kestirimi yöntemlerini karşılaştırmışlardır. Bu araştırmanın sonuçları, yönlendirme modülünün uzunluğu arttıkça yanlış yönlendirmenin azaldığını raporlamaktadır.

Karatoprak Erşen ve Lee (2023) ise 1-3 BÇAT tasarımı altında farklı madde kalibrasyon yöntemlerini karşılaştırdığı çalışmasında, yönlendirme modülü uzunluğu kısaldıkça yanlış yönlendirme oranının arttığını bildirmektedir.

Yöntem

Bu çalışmada BÇAT tasarım ve bileşenlerinin test katılımcılarının yanlış modüle yönlendirilmesine etkisi incelenmiştir. Araştırmada yer alan veriler simülasyon yöntemi ile üretilmiş ve farklı koşullar altında karşılaştırmalar yapılmıştır. Simülasyon çalışması, olasılık dağılımlarıyla rastgele örnekleme yapılarak verilerin üretilmesini ve analizin içeren bilgisayar deneyleri olarak bilinir. Bu deneyler, belirli senaryo ve koşullar altında istatistiksel yöntemlerin performanslarını karşılaştırmak amacıyla yapılmaktadır. Psikometri alanında simülasyon çalışmaları; yöntemlerin değerlendirilmesi, yeni yöntemlerin geliştirilmesi ve karşılaştırılması noktasında önemli bir yer tutmaktadır (Morris ve diğerleri, 2019). Bu çalışmada ele alınan tüm koşulların gerçek veri üzerinde aynı anda ele alınmasının çok zor olması nedeniyle simülasyon yöntemi tercih edilmiştir. Bu çalışma, BÇAT'ta farklı tasarım ve bileşenler altında yanlış yönlendirme durumunun ortaya konulmasını amaçladığından betimsel araştırma olduğu söylenebilir (Fraenkel ve diğerleri, 2012).

Araştırmada üç farklı BÇAT tasarımı (1-3, 1-2-3 ve 1-3-3), üç farklı test uzunluğu (12, 24 ve 36), üç farklı modül uzunluğu dağılımı (Uzun-Kısa [L-S], Orta-Orta [M-M] ve Kısa-Uzun [S-L]) ve iki farklı Yönlendirme Modülü Tasarımı (Geniş Yetenek Aralığı [-0.5, 0.0, 0.5] ve Dar Yetenek Aralığı [0.0]) koşulları değişimlenerek simülasyon çalışması gerçekleştirilmiştir. Tüm koşullar birbirleri ile çaprazlanmıştır. Dolayısıyla bu çalışmada 3x3x3x2=54 koşul incelenmiştir. Her koşul için 100 replikasyon yapılarak analizler gerçekleştirilmiştir. Çalışmada verilerin üretilmesi ve analiz edilmesi için açık kaynak kodlu R yazılımından faydalanılmıştır. Otomatik test birleştirme için "Rmst" (Luo, 2018), BÇAT analizleri için "mstR" (Magis ve diğerleri, 2017) paketleri ve yanlış yönlendirmenin tespit edilmesi için araştırmacılar tarafından yazılan kodlar kullanılmıştır. Simülasyon çalışmasına dair ayrıntılar aşağıda yer alan başlıklar altında sunulmuştur.

Madde ve Yetenek Parametreleri

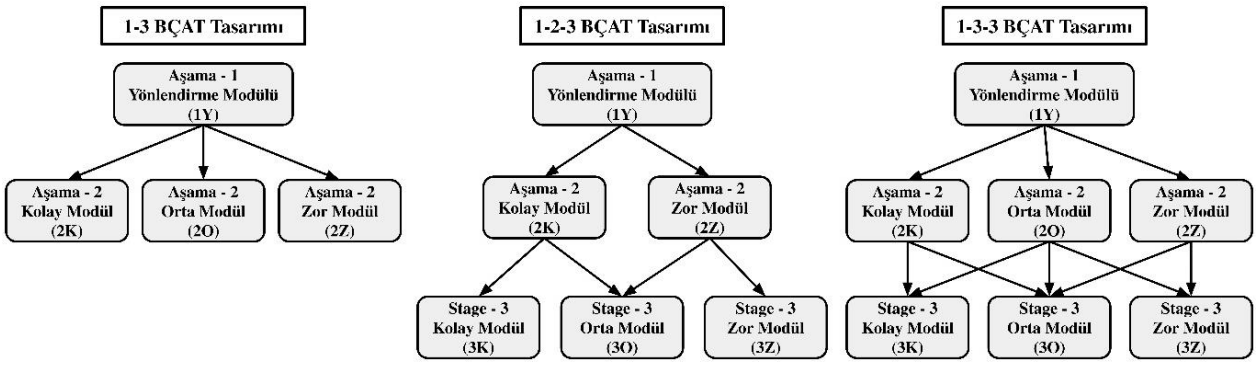
Araştırma kapsamında 3PL modele dayalı olarak 200 maddenin parametreleri literatürde yer alan dağılımlar göz önüne alınarak üretilmiştir (Feinberg ve Rubright, 2016; Mooney, 1997). Madde ayırt edicilik parametreleri log-normal dağılımdan $a \sim lnN(0.3, 0.2)$, madde güçlük parametreleri normal dağılımdan $b \sim N(0, 1)$, madde şans (guessing) parametreleri beta dağılımından $c \sim Beta(8, 32)$ elde edilmiştir. Yetenek parametreleri ise 1000 test katılımcısı olacak şekilde normal dağılımdan $\theta \sim N(0, 1)$ üretilmiştir. Yetenek kestiriminde EAP yöntemi kullanılmıştır.

BÇAT Tasarımı

Bu çalışmada 1-3, 1-2-3 ve 1-3-3 olmak üzere üç farklı BÇAT tasarımı kullanılmıştır. Bu BÇAT tasarımları Şekil 2'te gösterilmektedir.

Şekil 2

1-3, 1-2-3 ve 1-3-3 BÇAT Tasarımları



Test Uzunluğu

Bu çalışmada test uzunluğu koşulu altında 12, 24 ve 36 olmak üzere üç farklı test uzunluğuna yer verilmiştir.

Modül Uzunluğu

Bu çalışmada modül uzunluğu olarak Kısa-Uzun [K-U], Orta-Orta [O-O] ve Uzun-Kısa [U-K] olmak üzere üç farklı modül uzunluğu yer almaktadır. K-U modül uzunluğunda öncelikle ilk aşamadaki modüller kısa iken, sonraki modüller uzun test uzunluğundadır. O-O modül uzunluğunda ise tüm modül uzunlukları eşit bir şekilde dağıtılmıştır. U-K modül uzunluğunda da yönlendirme modülü uzun iken sonraki modül kısa uzunluktadır. Tablo 1'de iki ve üç aşamalı BÇAT tasarımlarına ve test uzunluklarına göre modüllerde yer alan madde sayılarının dağılımı gösterilmektedir.

Tablo 1

Test Uzunluklarına Göre Modül Uzunlukları Dağılımı

BÇAT Tasarımları	Test Uzunluğu	Uzun-Kısa [L-S]	Orta-Orta [M-M]	Kısa-Uzun [S-L]
1-3	12	8-4	6-6	4-8
	24	16-8	12-12	8-16
	36	24-12	18-18	12-24
1-2-3 ve 1-3-3	12	6-3-3	4-4-4	3-3-6
	24	12-6-6	8-8-8	6-6-12
	36	18-9-9	12-12-12	9-9-18

Modüllerin Yetenek Dağılımları ve Yönlendirme Modülü

BÇAT modül ve panellerinin birleştirilmesinde modül bilgi miktarının maksimum düzeye çıkarılacağı yetenek düzeyleri Tablo 2'de gösterilmektedir.

Tablo 2

Modül Bilgisinin Maksimuma Ulaştığı Yetenek Düzeyleri

BÇAT Desenleri	Aşama 1	Aşama 2	Aşama 3
1-3	Dar : $\theta = 0$ Geniş : $\theta = (-0.5, 0, 0.5)$	$\theta = (-1, 0, 1)$	
1-2-3	Dar : $\theta = 0$ Geniş : $\theta = (-0.5, 0, 0.5)$	$\theta = (-0.5, 0.5)$	$\theta = (-1, 0, 1)$
1-3-3	Dar : $\theta = 0$ Geniş : $\theta = (-0.5, 0, 0.5)$	$\theta = (-0.75, 0, 0.75)$	$\theta = (-1, 0, 1)$

Tablo 2'de görüldüğü üzere yönlendirme modülü için geniş ve dar olmak üzere iki tasarım yer almaktadır. Geniş yönlendirme modülü tasarımında $\vartheta = (-0.5, 0.0, 0.5)$ aralığında modül bilgi düzeyinin maksimum düzeye çıkarılması hedeflenirken, dar yönlendirme modülü tasarımında ise $\theta = 0$ düzeyinde modül bilgi düzeyinde modül bilgisinin maksimuma çıkarılması hedeflenmiştir.

Tablo 2 incelendiğinde BÇAT tasarımlarının son aşamalarının modül bilgi düzeylerinin $\theta = (-1, 0, 1)$ olacak şekilde tüm BÇAT tasarımlarında aynı olacak şekilde tasarlanmıştır. Üç aşamalı 1-2-3 ve 1-3-3 BÇAT tasarımlarının ikinci aşamalarında yer alan modüller ise sırasıyla $\theta = (-0.5, 0.5)$ ve $\theta = (-0.75, 0, 0.75)$ olacak şekilde tasarlanmıştır.

Yanlış Yönlendirmenin Tespiti

Bir katılımcı, yönlendirme modülünün ardından gelen aşamada - MFI yönlendirme kuralına göre- modüller arasında gerçek yetenek düzeyine göre maksimum bilgi sunan modüle yönlendirilmelidir. Dolayısıyla bu çalışmada gerçek yetenek düzeyine göre maksimum bilgi sunan modüle yönlendirilmeyen katılımcılar yanlış yönlendirme olarak etiketlenmiştir. Örneğin 1-3 BÇAT tasarımında $\theta = 1.48$ yetenek düzeyindeki bir katılımcının yönlendirme modülünün ardından zor modüle yönlendirilmesi beklenir. Fakat katılımcının yönlendirme modülüne verdiği yanıtlardan elde edilen yetenek kestiriminin düşük olması durumunda bu katılımcı kolay veya orta güçlükteki modüllerden birine yönlendiriliyorsa bu durum yanlış yönlendirme olarak değerlendirilmektedir. Bu çalışmada 1-3, 1-2-3 ve 1-3-3 BÇAT tasarımları ele alınmıştır. Yanlış yönlendirmenin tespiti yalnızca tüm BÇAT tasarımlarının son aşamaları üzerinden gerçekleştirilmiştir. Yanlış yönlendirmenin tespitinin yalnızca son aşamaya göre belirlenmesi bu araştırmanın sınırlılığıdır.

Yanlış yönlendirmenin belirlenmesinde son aşamadaki modüllerin madde bilgi fonksiyonlarının kesiştiği noktalar belirlenmiş ve bu noktalara göre en yüksek madde bilgisi sunan modüle yönlendirilmeyen katılımcılar yanlış yönlendirme olarak etiketlenmiştir. Tüm koşullarda 1000 katılımcından kaçının yanlış yönlendirildiği hesaplanmış ve yüzde (%) üzerinden raporlanmıştır.

Verilerin Analizi

Verilerin analizinde farklı BÇAT tasarım ve bileşenlerinden elde edilen verilerin değerlendirilmesi için gerçek ve kestirilen birey parametreleri arasındaki ilişkiler Hata Kareleri Ortalamasının Karekökü (Root Mean Square Error – RMSE) ve ortalama mutlak hata (OMH) değerleri hesaplanarak yorumlanmıştır. RMSE ve OMH değerlerinin formülleri aşağıda yer almaktadır.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(\widehat{\theta_i}-\theta_i)^2}{n}}$$

$$\text{OMH} = \frac{\sum_{i=1}^{n}|\widehat{\theta_i}-\theta_i|}{n}$$

Çalışmada yanlış yönlendirilenlerin ölçme kesinliğindeki değişimin tespit edilebilmesi amacıyla hem tüm katılımcılar için hem de doğru yönlendirilen ve yanlış yönlendirilen katılımcılar için RMSE ve OMH değerleri ayrı ayrı hesaplanarak sunulmuştur. Son olarak, yanlış yönlendirilenlerin yüzdesi (%) hesaplanarak sunulmuştur.

Bulgular

Bu bölümde öncelikle araştırmada incelenen 54 koşuldan elde edilen tüm sonuçlar Tablo 3 ve Tablo 4'te sunulmuştur. Ardından araştırma soruları bazında bulgulara yer verilmiştir.

Tablo 3 ve Tablo 4'te tüm katılımcılardan elde edilen ölçme kesinliği bulguları RMSE ve OMH değerleri ile sunulmuştur. Ayrıca sadece yanlış yönlendirilen katılımcıların RMSE ve OMH değerleri YY-RMSE ve YY-OMH sütunlarında yer almaktadır. Benzer şekilde doğru yönlendirilenlerin RMSE ve OMH değerleri ise DY-RMSE ve DY-OMH sütunlarında sunulmuştur.

Tablo 3

Tüm Koşullardan Elde Edilen Sonuçlar (1-3 BÇAT Tasarımı)

Tasarım	Test Uzunluğu	Yönlendirme Modülü Tasarımı	Modül Uzunluğu	RMSE	YY-RMSE	DY-RMSE	OMH	YY-OMH	DY-OMH	Yanlış Yönlendirilen Oranı (%)
1-3	12	Geniş	U-K	0.356	0.399	0.343	0.274	0.320	0.261	21.4
			O-O	0.347	0.388	0.332	0.268	0.303	0.256	25.2
			K-U	0.338	0.391	0.316	0.261	0.299	0.246	27.2
		Dar	U-K	0.356	0.394	0.344	0.273	0.311	0.262	22.6
			O-O	0.348	0.384	0.333	0.267	0.294	0.257	26.5
			K-U	0.350	0.406	0.324	0.267	0.303	0.252	29.1
	24	Geniş	U-K	0.265	0.273	0.263	0.203	0.222	0.199	16.8
			O-O	0.254	0.277	0.248	0.195	0.218	0.190	18.6
			K-U	0.247	0.269	0.240	0.190	0.210	0.184	21.6
		Dar	U-K	0.269	0.265	0.270	0.206	0.211	0.204	20.3
			O-O	0.258	0.264	0.256	0.198	0.207	0.196	21.7
			K-U	0.255	0.280	0.246	0.195	0.214	0.189	23.4
	36	Geniş	U-K	0.230	0.220	0.232	0.175	0.178	0.174	15.3
			O-O	0.215	0.218	0.215	0.165	0.174	0.164	17.6
			K-U	0.211	0.222	0.208	0.162	0.176	0.159	19.0
		Dar	U-K	0.233	0.217	0.236	0.177	0.173	0.178	17.6
			O-O	0.218	0.219	0.218	0.167	0.174	0.166	18.0
			K-U	0.214	0.225	0.211	0.165	0.176	0.162	21.4

* YY: Yanlış Yönlendirme

* DY: Doğru Yönlendirme

Tablo 4

Tüm Koşullardan Elde Edilen Sonuçlar (1-2-3 ve 1-3-3 BÇAT Tasarımları)

Tasarım	Test Uzunluğu	Yönlendirme Modülü Tasarımı	Modül Uzunluğu	RMSE	YY-RMSE	DY-RMSE	OMH	YY-OMH	DY-OMH	Yanlış Yönlendirilen Oranı (‰)
1-2-3	12	Geniş	U-K	0.345	0.404	0.330	0.266	0.333	0.251	18.7
			O-O	0.341	0.405	0.322	0.263	0.324	0.248	20.1
			K-U	0.337	0.410	0.312	0.262	0.323	0.244	22.8
		Dar	U-K	0.341	0.386	0.329	0.262	0.314	0.250	19.5
			O-O	0.336	0.401	0.318	0.258	0.320	0.243	19.5
			K-U	0.325	0.404	0.300	0.252	0.320	0.234	21.2
	24	Geniş	U-K	0.259	0.285	0.255	0.199	0.235	0.193	14.2
			O-O	0.245	0.282	0.238	0.189	0.233	0.182	13.9
			K-U	0.244	0.292	0.234	0.189	0.235	0.180	16.0
		Dar	U-K	0.258	0.288	0.252	0.198	0.239	0.191	14.2
			O-O	0.247	0.281	0.241	0.191	0.229	0.184	14.4
			K-U	0.247	0.293	0.237	0.191	0.234	0.183	16.6
	36	Geniş	U-K	0.224	0.223	0.224	0.170	0.180	0.168	14.0
			O-O	0.210	0.223	0.207	0.161	0.183	0.158	12.4
			K-U	0.204	0.232	0.199	0.159	0.189	0.154	13.4
		Dar	U-K	0.221	0.224	0.220	0.168	0.183	0.165	13.3
			O-O	0.211	0.228	0.209	0.162	0.188	0.158	12.5
			K-U	0.204	0.237	0.199	0.160	0.193	0.154	13.8
1-3-3	12	Geniş	U-K	0.345	0.413	0.327	0.267	0.338	0.249	19.6
			O-O	0.336	0.390	0.321	0.260	0.315	0.245	20.3
			K-U	0.347	0.396	0.329	0.268	0.312	0.254	24.2
		Dar	U-K	0.342	0.395	0.327	0.264	0.319	0.250	19.9
			O-O	0.331	0.373	0.319	0.255	0.298	0.244	21.1
			K-U	0.331	0.386	0.313	0.255	0.302	0.241	22.2
	24	Geniş	U-K	0.262	0.273	0.260	0.200	0.226	0.195	15.0
			O-O	0.247	0.274	0.243	0.190	0.223	0.184	14.7
			K-U	0.245	0.267	0.240	0.188	0.214	0.183	17.4
		Dar	U-K	0.262	0.281	0.258	0.200	0.229	0.194	15.4
			O-O	0.252	0.271	0.248	0.193	0.219	0.188	15.6
			K-U	0.249	0.279	0.242	0.190	0.220	0.184	16.9
	36	Geniş	U-K	0.224	0.228	0.223	0.170	0.187	0.167	13.3
			O-O	0.211	0.224	0.209	0.161	0.184	0.158	12.8
			K-U	0.209	0.224	0.206	0.160	0.181	0.157	14.4
		Dar	U-K	0.218	0.234	0.216	0.167	0.192	0.163	13.1
			O-O	0.211	0.227	0.209	0.162	0.186	0.158	13.6
			K-U	0.207	0.229	0.203	0.160	0.185	0.155	14.5

* YY: Yanlış Yönlendirme

* DY: Doğru Yönlendirme

Tablo 3 ve Tablo 4 incelendiğinde tüm koşullarda RMSE değerinin 0.204 ile 0.356, OMH değerlerinin ise 0.161 ile 0.274 aralığında değiştiği görülmektedir. YY-RMSE değerleri 0.217 ile 0.413; DY-RMSE değerleri ise 0.173 ile 0.338 aralığında değişmektedir. YY-OMH değerleri 0.199 ile 0.344; DY-OMH

değerleri ise 0.154 ile 0.262 aralığında değişmektedir. Hem RMSE hem de OMH değerlerine göre genel olarak bakıldığında yanlış yönlendirilenlerin ölçme kesinliğinin doğru yönlendirilenlere göre daha düşük olduğu görülmektedir. Dolayısıyla yanlış yönlendirmenin ölçme kesinliği açısından BÇAT tasarımları için bir tehdit oluşturduğu ifade edilebilir.

Tablolarda yer alan yanlış yönlendirilen oranları ise %12.4 ile %29.1 aralığındadır. Bu bulgu ise BÇAT'ta tasarım ve bileşenler üzerinde yapılacak değişimlerin yanlış yönlendirmeyi azaltmada etkili olduğuna işaret etmektedir.

Araştırmanın bu bölümünde araştırma sorularına ilişkin bulgulara yer verilmiştir. Araştırma sorularında yer alan tablolar Tablo 3 ve Tablo 4'teki koşulların ortalaması alınarak elde edilmiştir.

Birinci Araştırma Sorusuna İlişkin Bulgular

Farklı BÇAT tasarımlarına göre yanlış yönlendirme oranı ve ölçme kesinliği bulguları Tablo 5'te gösterilmektedir.

Tablo 5

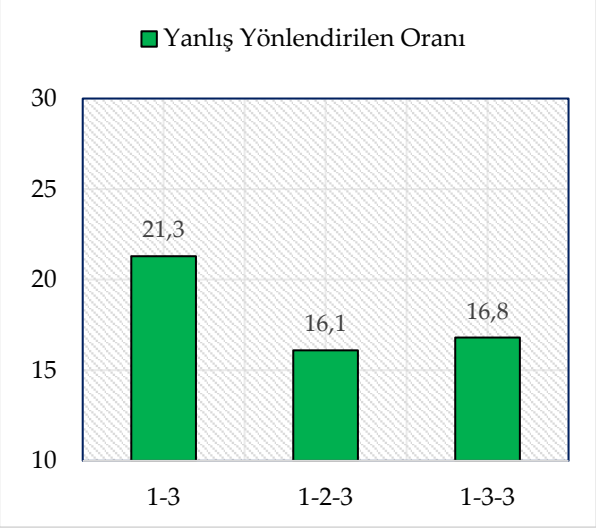
Farklı BÇAT Tasarımlarına Göre Bulgular

BÇAT Design	RMSE	YY-RMSE	DY-RMSE	OMH	YY-OMH	DY-OMH	Yanlış Yönlendirme Oranı (%)
1-3	0.275	0.295	0.269	0.212	0.231	0.206	21.3
1-2-3	0.266	0.305	0.257	0.205	0.248	0.197	16.1
1-3-3	0.268	0.298	0.261	0.206	0.241	0.198	16.8

Tablo 5'te yer alan yanlış yönlendirme oranı bulguları Şekil 3'de görselleştirilerek sunulmuştur.

Şekil 3

Farklı BÇAT Tasarımlarına Göre Yanlış Yönlendirme Oranları



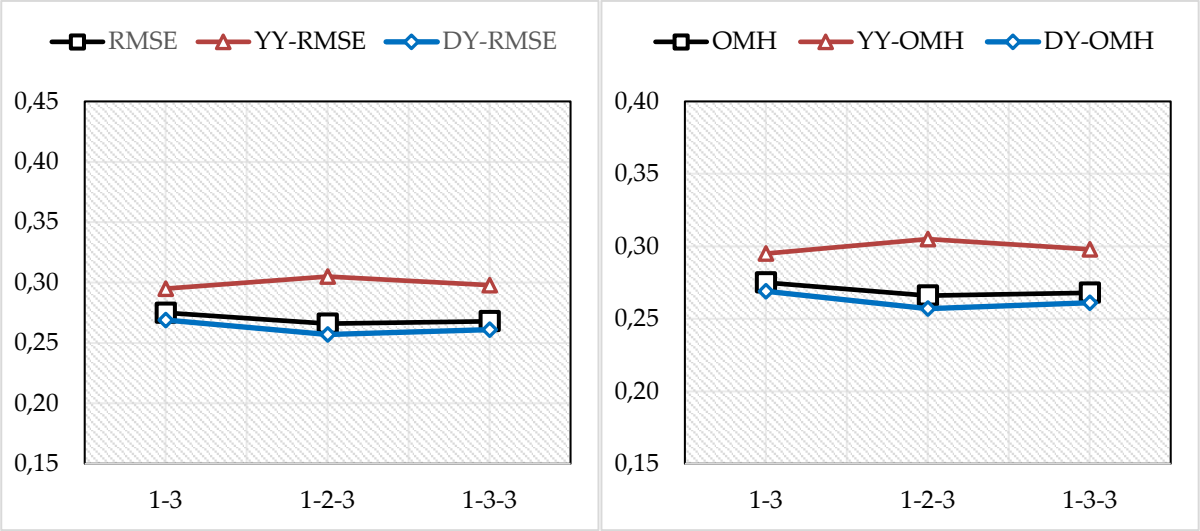
Şekil 3'de görüldüğü üzere 1-3 BÇAT tasarımında yanlış yönlendirme oranı %21.3 iken, 1-2-3 ve 1-3-3 BÇAT tasarımlarında sırasıyla %16.1 ve 16.8 olduğu görülmektedir. Bu bulgu üç aşamalı BÇAT tasarımlarında iki aşamalı BÇAT tasarımlarına göre yanlış yönlendirmenin azaldığını göstermektedir. Ayrıca 1-2-3 BÇAT tasarımının yanlış yönlendirme oranının 1-3-3 BÇAT

tasarımından az da olsa düşük olduğu ifade edilebilir.

Tablo 5'te yer alan RMSE ve OMH ölçme kesinliği bulguları Şekil 4'te görselleştirilerek sunulmuştur.

Şekil 4

Farklı BÇAT Tasarımlarına Göre Ölçme Kesinliği Bulguları



Şekil 4 incelendiğinde 1-3 BÇAT tasarımından elde edilen RMSE değeri 0.275 iken, 1-2-3 ve 1-3-3 BÇAT tasarımından elde edilen RMSE değerleri sırasıyla 0.266 ve 0.268'dir. Benzer şekilde, 1-3 BÇAT tasarımından elde edilen OMH değeri 0.212 iken, 1-2-3 ve 1-3-3 BÇAT tasarımından elde edilen OMH değerleri sırasıyla 0.205 ve 0.206'dır. Bu bulgular 1-2-3 ve 1-3-3 BÇAT tasarımlarının ölçme kesinliğinin 1-3 BÇAT tasarımından daha iyi olduğunu göstermektedir. Ayrıca 1-2-3 ve 1-3-3 BÇAT tasarımlarının ölçme kesinliği bulguları oldukça benzer olup 1-2-3 BÇAT tasarımının ölçme kesinliğinin daha az bir fark ile daha iyi olduğu söylenebilir. Ayrıca tüm BÇAT tasarımlarında yanlış yönlendirilen katılımcıların RMSE ve OMH değerlerinin doğru yönlendirilenlere göre daha yüksek olduğu görülmektedir. Bu bulgu tüm BÇAT tasarımlarında yanlış yönlendirilenlerin ölçme kesinliğinin az olduğunu göstermektedir.

İkinci Araştırma Sorusuna İlişkin Bulgular

Farklı test uzunluklarına göre yanlış yönlendirme oranı ve ölçme kesinliği bulguları Tablo 6'da gösterilmektedir.

Tablo 6

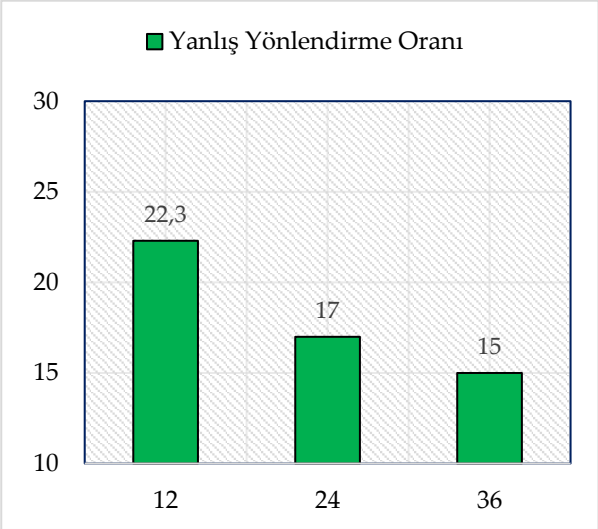
Farklı Test Uzunluklarına Göre Bulgular

Test Uzunluğu	RMSE	YY-RMSE	DY-RMSE	OMH	YY-OMH	DY-OMH	Yanlış Yönlendirme Oranı (%)
12	0.342	0.396	0.324	0.263	0.314	0.249	22.3
24	0.254	0.277	0.248	0.195	0.223	0.189	17.0
36	0.215	0.225	0.214	0.165	0.182	0.162	15.0

Tablo 6'da yer alan yanlış yönlendirme oranı bulguları Şekil 5'te görselleştirilerek sunulmuştur.

Şekil 5

Farklı Test Uzunluklarına Göre Yanlış Yönlendirme Oranları

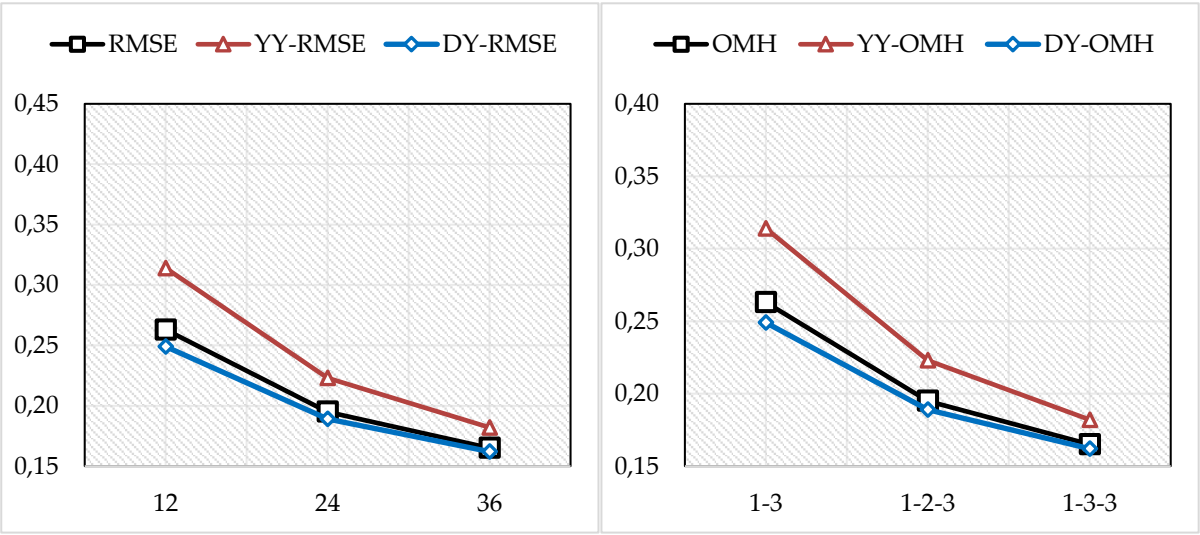


Şekil 5'te görüldüğü üzere 12 test uzunluğunda yanlış yönlendirme oranı %22.3 iken, 24 ve 36 test uzunluklarında %17.00 ve %15.00 olduğu görülmektedir. Bu bulgu test uzunluğu arttıkça yanlış yönlendirme oranının azaldığını göstermektedir. Ayrıca 12 test uzunluğundan 24 test uzunluğuna yanlış yönlendirme oranı %5.3 birim azalırken, 24 test uzunluğundan 36 test uzunluğuna %2.0 birim azalmıştır. Dolayısıyla test uzunluğu arttıkça yanlış yönlendirme oranının doğrusal bir şekilde azalmadığı belirtilmelidir.

Tablo 6'da yer alan RMSE ve OMH ölçme kesinliği bulguları Şekil 6'da görselleştirilerek sunulmuştur.

Şekil 6

Farklı Test Uzunluklarına Göre Ölçme Kesinliği Bulguları



Şekil 6 incelendiğinde 12 test uzunluğundan elde edilen RMSE değeri 0.342 iken, 1-2-3 ve 1-3-3 BÇAT tasarımından elde edilen RMSE değerleri sırasıyla 0.254 ve 0.215'dir. Benzer şekilde, 12 test uzunluğundan elde edilen OMH değeri 0.263 iken, 24 ve 36 test uzunluğundan elde edilen OMH değerleri sırasıyla 0.195 ve 0.165'dır. Bu bulgular test uzunluğu arttıkça ölçme kesinliğinin arttığını

göstermektedir.

Şekil 6'da görüldüğü üzere 12 test uzunluğunda YY-RMSE ve DY-RMSE farkı 0.072 (0.396-0.324) iken 24 ve 36 test uzunluklarında bu fark sırasıyla 0.019 (0.277-0.248) ve 0.014 (0.225-0.214) olmaktadır. Bu durum, test uzunluğu arttıkça yanlış yönlendirilen katılımcıların ölçme kesinliğinin de belirgin bir şekilde arttığını göstermektedir. Diğer bir deyişle, test uzunluğunun artması hem yanlış yönlendirilenlerin oranını hem de yanlış yönlendirilenlerin yetenek kestirimi hatalarını azaltmaktadır.

Üçüncü Araştırma Sorusuna İlişkin Bulgular

Farklı modül uzunluklarına göre ölçme kesinliği ve yanlış yönlendirme bulguları Tablo 7'de gösterilmektedir.

Tablo 7

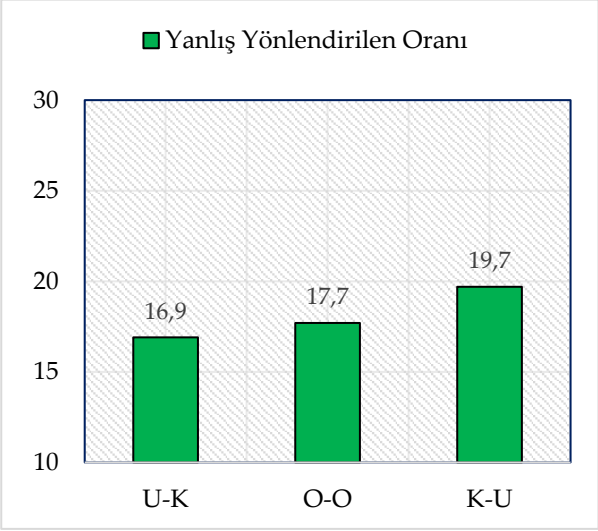
Farklı Modül Uzunluklarına Göre Bulgular

Modül Uzunluğu	RMSE	YY-RMSE	DY-RMSE	OMH	YY-OMH	DY-OMH	Yanlış Yönlendirme Oranı (%)
U-K	0.278	0.300	0.273	0.213	0.244	0.206	16.9
O-O	0.268	0.296	0.260	0.206	0.237	0.199	17.7
K-U	0.265	0.302	0.253	0.204	0.238	0.195	19.7

Tablo 7'de yer alan yanlış yönlendirme oranı bulguları Şekil 7'de görselleştirilerek sunulmuştur.

Şekil 7

Farklı Modül Uzunluklarına Göre Yanlış Yönlendirme Oranları

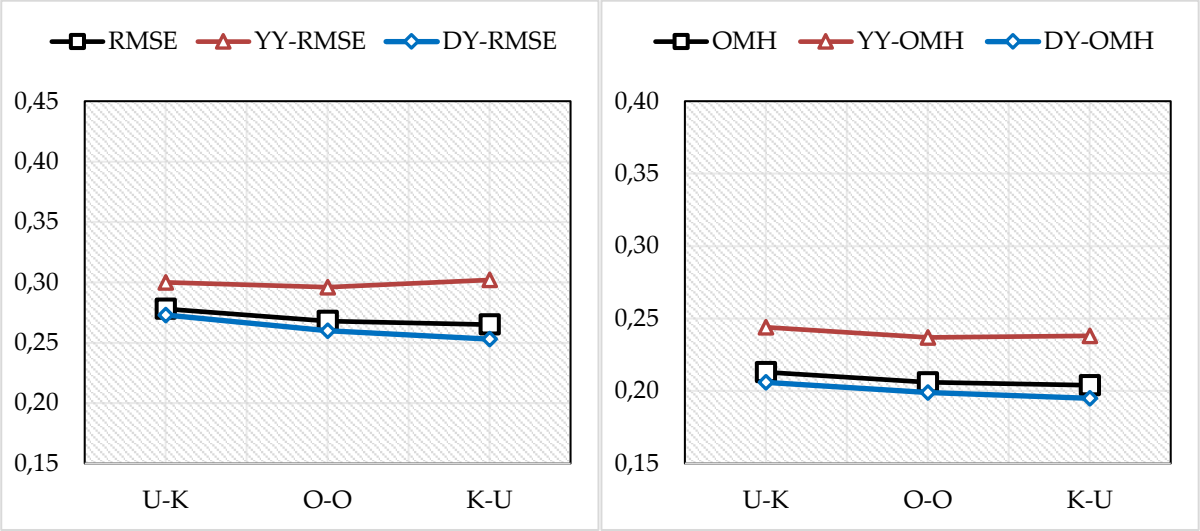


Şekil 7'de görüldüğü üzere U-K modül uzunluğunda yanlış yönlendirme oranı %16.9 iken, O-O ve K-U modül uzunluğu oranlarında sırasıyla %17.7 ve %19.7 olduğu görülmektedir. Bu bulgu, yönlendirme modülünün uzunluğu arttıkça yanlış yönlendirme oranının azaldığını göstermektedir.

Tablo 7'de yer alan RMSE ve OMH ölçme kesinliği bulguları Şekil 8'de görselleştirilerek sunulmuştur.

Şekil 8

Farklı Modül Uzunluklarına Göre Ölçme Kesinliği Bulguları



Şekil 8 incelendiğinde U-K modül uzunluğundan elde edilen RMSE değeri 0.278 iken, O-O ve K-U modül uzunluklarından elde edilen RMSE değerleri sırasıyla 0.268 ve 0.205'tir. Benzer şekilde, U-K modül uzunluğundan elde edilen OMH değeri 0.213 iken, O-O ve K-U modül uzunluğundan elde edilen OMH değerleri sırasıyla 0.268 ve 0.265'tir. Bu bulgular yönlendirme modülü uzunluğu arttıkça ölçme kesinliğinin bir miktar azaldığını göstermektedir. Dolayısıyla yönlendirme modülü uzunluğu arttıkça yanlış yönlendirme azalırken, aynı zamanda ölçme kesinliği de azalmaktadır. Bu noktada yanlış yönlendirmenin yanında ölçme kesinliği de dikkate alınarak modül uzunluğuna karar vermek gerektiği anlaşılmaktadır.

Şekil 8'de görüldüğü üzere U-K, O-O ve K-U modül uzunluklarında yanlış yönlendirilen katılımcıların ölçme kesinliğinin doğru yönlendirilenlerden daha düşük olduğu belirtilebilir.

Dördüncü Araştırma Sorusuna İlişkin Bulgular

Farklı yönlendirme modülü tasarımlarına göre ölçme kesinliği ve yanlış yönlendirme bulguları Tablo 8'de gösterilmektedir.

Tablo 8

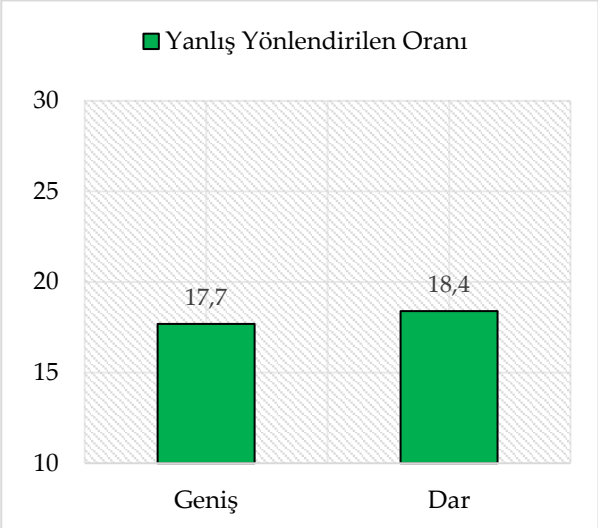
Farklı Yönlendirme Modülü Tasarımlarına Göre Bulgular

Yönlendirme Modülü Tasarımı	RMSE	YY-RMSE	DY-RMSE	OMH	YY-OMH	DY-OMH	Yanlış Yönlendirme Oranı (%)
Geniş	0.270	0.300	0.262	0.208	0.241	0.200	17.7
Dar	0.270	0.299	0.262	0.208	0.238	0.200	18.4

Tablo 8'de yer alan yanlış yönlendirme oranı bulguları Şekil 9'da görselleştirilerek sunulmuştur.

Şekil 9

Farklı Yönlendirme Modülü Tasarımlarına Göre Yanlış Yönlendirme Oranları

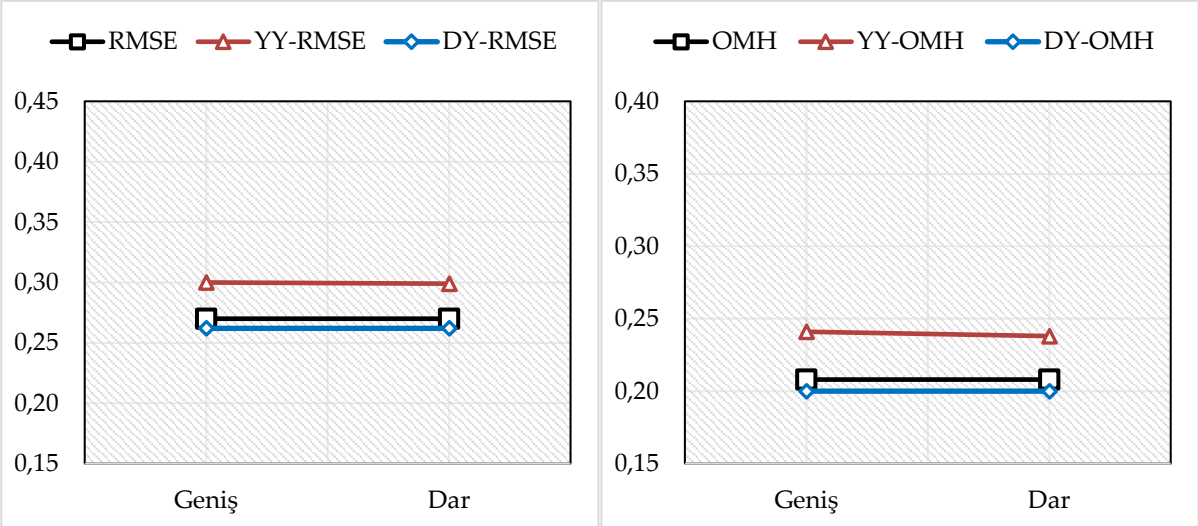


Şekil 9'da görüldüğü üzere geniş yönlendirme modülü tasarımının yanlış yönlendirme oranı %17.7 iken, dar yönlendirme modülü tasarımının yanlış yönlendirme oranı %18,4'tür. Bu bulgu, geniş yönlendirme modülünün, dar yönlendirme modülüne göre daha düşük yanlış yönlendirme oranına sahip olduğunu göstermektedir. Diğer bir deyişle, geniş yönlendirme tasarımı, dar yönlendirme tasarımına göre yanlış yönlendirme oranını azaltmaktadır.

Tablo 8'de yer alan RMSE ve OMH ölçme kesinliği bulguları Şekil 10'da görselleştirilerek sunulmuştur.

Şekil 10

Farklı Yönlendirme Tasarımlarına Göre Ölçme Kesinliği Bulguları



Şekil 10 incelendiğinde geniş modül tasarımından RMSE değeri 0.270 iken, dar modül tasarımından elde edilen RMSE değeri 0.270'tir. Benzer şekilde, geniş modül tasarımından elde edilen OMH değeri 0.208 iken, dar modül tasarımından elde edilen OMH değeri 0.208'dir. Bu bulgular yönlendirme modülü tasarımının ölçme kesinliği değiştirmediğini göstermektedir.

Şekil 10'da görüldüğü üzere hem geniş hem de dar yönlendirme modülü tasarımında yanlış

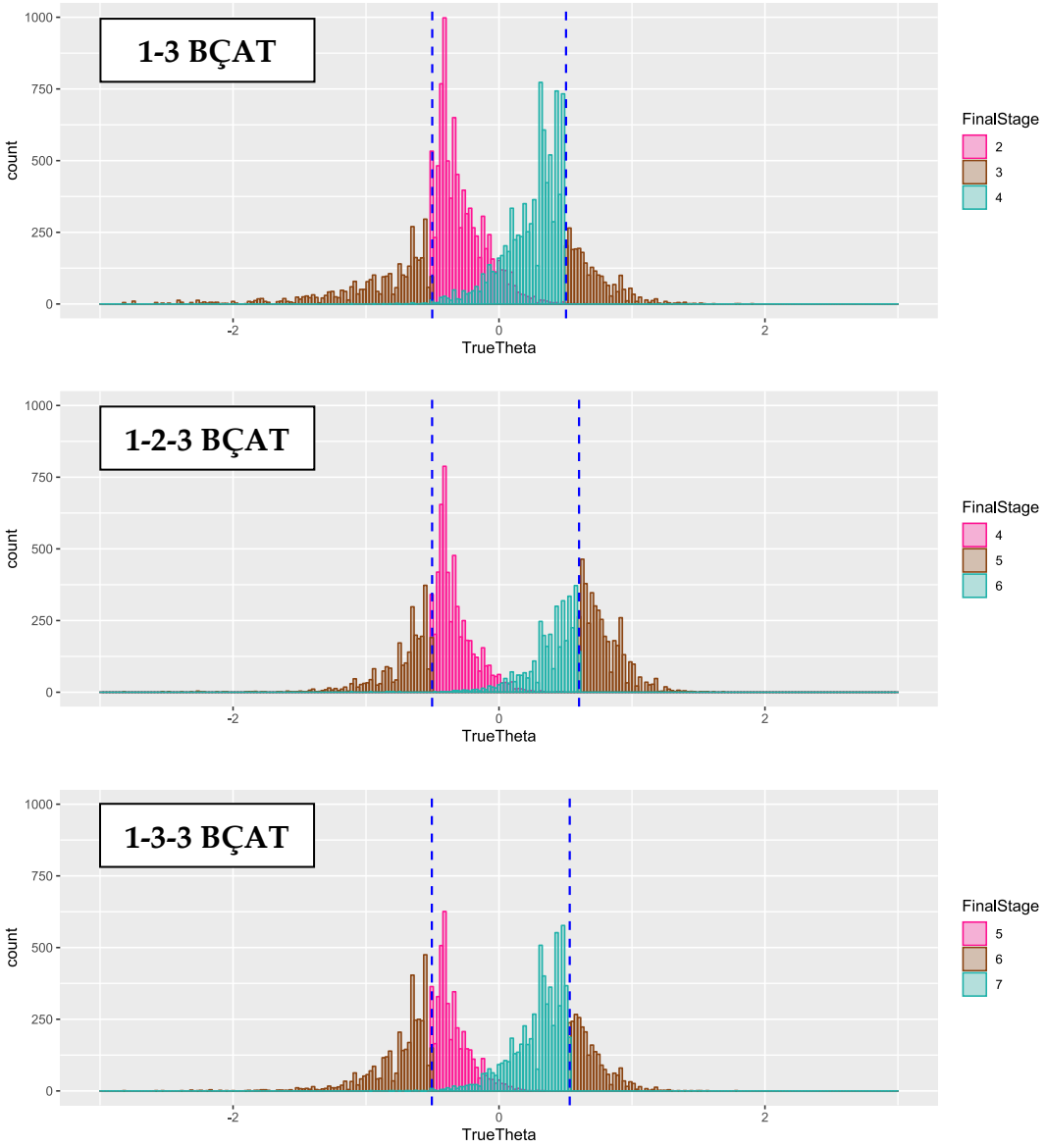
yönlendirilenlerin ölçme kesinliğinin doğru yönlendirilenlerin ölçme kesinliğinden daha düşük olduğu belirtilebilir.

Beşinci Araştırma Sorusuna İlişkin Bulgular

Yanlış yönlendirilen test katılımcılarının yetenek dağılımı aşağıdaki Şekil 11'te sunulmuştur.

Şekil 11

Yanlış Yönlendirilenlerin Yetenek Düzeyleri



Şekil 11'de yanlış yönlendirilen test katılımcılarının gerçek yetenek düzeylerine göre BÇAT tasarımının son aşamasında yönlendirildiği modüller histogram grafiği üzerinde renklendirilerek gösterilmiştir. Mavi renkli kesikli çizgiler, son aşamadaki modül bilgi fonksiyonlarının kesişim noktalarını göstermektedir. Örneğin 1-3 BÇAT grafiğinde yer alan -0.502 noktası kolay (2K) ve orta (2O) güçlükteki modüllerin bilgi fonksiyonlarının kesişim noktasını ifade etmektedir. Histogram çubuklarının renkleri ise gerçek yetenek düzeyindeki katılımcıların yanlış yönlendirme ile hangi modüle yönlendirildiklerini göstermektedir. Örneğin 1-3 BÇAT tasarımındaki turkuaz sütunlarda gerçek yetenek düzeyine sahip olan yanlış yönlendirilen katılımcılar, ikinci aşamada 4 numaralı

modüle (ikinci aşamadaki zor modüle [2K]) yönlendirildiğini ifade etmektedir. Benzer şekilde 1-3-3 BÇAT tasarımında pembe sütunların konumunda gerçek yetenek düzeyine sahip olan katılımcıların üçüncü aşamada kolay modüle (3K) yönlendirildiği anlaşılmaktadır.

Şekil 11'de görüldüğü üzere yanlış yönlendirilen test katılımcılarının gerçek yetenek düzeyi modül bilgi fonksiyonlarının kesişim noktasında yoğunlaşmakta olup yetenek ölçeğinin uç noktalarına doğru azalmaktadır. Yanlış yönlendirilen katılımcıların yetenek düzeylerinin yetenek ölçeğinin orta bölümünde yer almasının yanlış yönlendirmenin ölçme kesinliğine olan etkisini azaltmakta olduğu söylenebilir.

Tartışma

Bireyselleştirilmiş uyarlanabilir testler, test katılımcılarının daha yüksek katılım, daha yüksek motivasyon ve daha düşük kaygı düzeyi açısından umut vadeden sonuçlara sahiptir (Ling ve diğerleri, 2017; Martin ve Ladenzic, 2018). Bu avantajların temelinde uyarlanabilir testlerde katılımcılara yetenek düzeyine uygun maddeler sunulması yatmaktadır (Şenel, 2021). Benzer şekilde, BÇAT'ta da katılımcıların yetenek düzeyine en uygun modüle yönlendirilmesi hem ölçme kesinliği hem de sınav psikolojisi açısından önemlidir.

Bu çalışmada farklı BÇAT tasarımları ve farklı temel bileşenlerin yanlış yönlendirmeye etkisi incelenmiştir. Araştırma sonuçlarına göre, üç aşamalı BÇAT tasarımları, iki aşamalı BÇAT tasarımına göre daha düşük yanlış yönlendirme oranlarına sahiptir.

Araştırmada hem ölçme kesinliği hem de yanlış yönlendirme oranı açısından 1-2-3 ve 1-3-3 BÇAT tasarımlarının daha iyi ölçme kesinliği sunduğu, 1-3 BÇAT tasarımının ise ölçme kesinliğinin daha düşük olduğu görülmüştür. Üç aşamalı BÇAT tasarımlarının iki aşamalı BÇAT tasarımlarına göre daha yüksek ölçme kesinliği sunduğunu belirten literatürde pek çok çalışma bulunmaktadır (Sari ve Huggins-Manley, 2017; Zenisky ve diğerleri, 2010). Ayrıca, bu çalışmada 1-2-3 BÇAT tasarımının 1-3-3 BÇAT tasarımına göre daha iyi ölçme kesinliği ve daha düşük yanlış yönlendirme oranı sunduğu görülmüştür.

Test uzunluğu bulgularına göre, test uzunluğu arttıkça yanlış yönlendirme oranının azaldığı görülmüştür. Ayrıca test uzunluğunun artması, yanlış yönlendirme oranının düşmesinin yanında, yanlış yönlendirilenlerin ölçme kesinliğini de oldukça arttırmıştır. Fakat test uzunluğunun artışının ölçme kesinliğini doğrusal bir şekilde arttırmamış ve benzer şekilde yanlış yönlendirme oranını doğrusal olarak azaltmamıştır. Dolayısıyla test uzunluğunun belirlenmesinde yüksek test uzunluklarından hedeflenen verim alınamayabilir. Bu çalışmanın bulgularına benzer şekilde literatürde test uzunluğunu arttırmanın ölçme kesinliğini arttıracağını belirten pek çok çalışma yer almaktadır (Thompson ve Weiss, 2011; Şahin, 2020; Erdem Kara ve Doğan, 2022).

Modül uzunluklarına göre yönlendirme modülünün uzunluğu arttıkça yanlış yönlendirme oranının azaldığı sonucuna erişilmiştir. Fakat yönlendirme modülü uzunluğunun arttırılması, uyarlanabilir aşamalar olan diğer aşamaların uzunluğunu azalttığından ölçme kesinliğinin azalmasına neden olmaktadır. Bu noktada hem yanlış yönlendirme oranları hem de ölçme kesinliği birlikte düşünülerek en uygun modül uzunluğunun kullanılması önerilmektedir. Yönlendirme modülü uzunluğunun arttırılmasının yanlış yönlendirmeyi azaltığına dair mevcut araştırmanın bulguları, Kim ve diğerleri (2015) ve Karatoprak Erşen ve Lee'nin (2023) bulguları ile benzerlik göstermektedir.

Yönlendirme modülünün geniş yetenek aralığında tasarlanması, dar yetenek aralığına göre yanlış yönlendirme oranını azalttığı sonucuna ulaşılmıştır. Ayrıca yönlendirme modülünün geniş ve dar yetenek aralığında tasarlanmasının ölçme kesinliği açısından bir farklılık yaratmadığı sonucuna ulaşılmıştır. Dolayısıyla yönlendirme modülünün geniş aralıkta tasarlanması önerilmektedir. Mevcut çalışmanın sonuçlarına benzer şekilde, Cai ve diğerleri (2021), yönlendirme modülünün modül bilgi miktarı arttığında yanlış yönlendirmenin azaldığını belirtmektedir. Ayrıca araştırmanın bu bulgusuna benzer şekilde, Şahin (2020), b parametreleri arasındaki fark değişiminin ölçme kesinliği üzerinde önemli bir fark yaratmadığı sonucuna ulaşmıştır.

Yanlış yönlendirilenlerin yetenek düzeylerinin modül bilgi fonksiyonlarının kesiştiği noktaların etrafında yoğunlaştığı sonucuna ulaşılmıştır. Ölçme kesinliği sonuçlarına göre, yanlış yönlendirilenlerin ölçme kesinliği doğru yönlendirilenlere göre daha düşüktür, fakat ölçme kesinliği aşırı farklı olmadığından yanlış yönlendirmenin önemli bir tehdit oluşturmadığı söylenebilir. Yanlış yönlendirilenlerin ölçme kesinliğinin aşırı farklı olmamasının sebebi olarak bu katılımcıların yetenek ölçeğinin orta noktalarında yoğunlaşması olabilir. Bu araştırmanın sonuçlarına benzer şekilde, Kim ve Moses (2014), yanlış yönlendirilenlerin modüller arası bilgi fonksiyonlarının kesme noktalarında yoğunlaştığını, dolayısıyla yanlış yönlendirilenlerin ölçme kesinliğine etkisinin minimum düzeyde olduğunu belirtmektedir. Yine benzer şekilde, Karamese (2022), yanlış yönlendirilenlerin kesme puanı etrafında yoğunlaştığını ifade etmektedir.

Bu araştırmanın sonuçları BÇAT'ta yanlış yönlendirmenin ölçme kesinliğini az da olsa düşürdüğünü göstermektedir. Yanlış yönlendirme oranının azaltılması için uygulayıcılar, BÇAT tasarım ve temel bileşenleri üzerinde değişiklikler yapabilirler. Bu araştırmanın sonuçlarına göre BÇAT tasarımının aşama sayısının arttırılması, test uzunluğunun arttırılması ve geniş yönlendirme modülü tasarlanması yanlış yönlendirme oranını azaltacaktır. Modül uzunluklarına göre ise hem yanlış yönlendirme oranı hem de ölçme kesinliği açısından modül uzunluklarının eşit olması önerilebilir. Gelecek çalışmalarda yönlendirme kurallarının, yetenek kestirimi yöntemlerinin ve madde havuzu özelliklerinin yanlış yönlendirme oranı üzerindeki etkisine dair araştırmalar tasarlanması önerilebilir.

Etik Kurul Onayı: Araştırmada yer alan veriler yer alan veriler bilgisayarda simülasyon ortamında üretilmiş olup herhangi bir insan veya canlı katılımcıdan veri toplanmamıştır.

Araştırmacıların Katkı Oranı: Bu çalışmanın araştırma sürecinin tüm aşamalarında yazarlar eşit oranda katkı sağlamışlardır.

Çatışma Beyanı: Yazarlar herhangi bir potansiyel çıkar çatışması olmadığını beyan ederler.

References

Breithaupt, K. J., Mills, C. N., & Melican, G. J. (2006). Facing the opportunities of the future. In D. Bartram & R. Hambleton (Eds.), *Computer-based testing and the Internet: Issues and advances,* 219-251. John Wiley & Sons Ltd.

Cai, L., Albano, A. D., & Roussos, L. A. (2021). An investigation of item calibration methods in multistage testing. *Measurement: Interdisciplinary Research and Perspectives, 19*(3), 163–178. https://doi.org/10.1080/15366367.2021.1878778

Demir, S. (2022). The effect of item pool and selection algorithms on computerized classification testing (CCT) performance. *Journal of Educational Technology and Online Learning, 5*(3), 573-584. https://doi.org/10.31681/jetol.1099580

Erdem Kara, B. (2022). Yönlendirme yöntemlerinin çok aşamalı testler üzerindeki etkisi [Effect of routing methods on the performance of multi-stage tests]. *Uluslararası Türk Eğitim Bilimleri Dergisi 10* (19), 343-354. https://doi.org/10.46778/goputeb.1123902

Erdem Kara, B., & Doğan, N. (2022). The effect of ratio of items indicating differential item functioning on computer adaptive and multi-stage tests. *International Journal of Assessment Tools in Education*, *9*(3), 682–696. https://doi.org/10.21449/ijate.1105769

Erkuş, A. (2012). Psikolojide ölçme ve ölçek geliştirme [Measurement and scale development in psychology]. *Pegem Akademi Yayınları*.

Eroğlu, M. G., & Kelecioğlu, H. (2015). Bireyselleştirilmiş bilgisayarlı test uygulamalarında 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: Conducting simulation studies in psychometrics. *Educational Measurement Issues and Practice*, *35*(2), 36–49. https://doi.org/10.1111/emip.12111

Fraenkel, J. R., Wallen, N. E., & Hyun, H. H. (2012). *How to design and evaluate research in education. In*(E. Kedelapan (Ed.). McGraw-Hill Companies.

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

Karamese, H. (2022). A *comparison of final scoring methods under the multistage adaptive testing framework.* [Unpublished Doctoral Dissertation]. The University of Iowa.

Karatoprak Ersen, R., & Lee, W.-C. (2023). Pretest item calibration in computerized multistage adaptive testing. *Journal of Educational Measurement*. https://doi.org/10.1111/jedm.12361

Khorramdel, L., Pokropek, A., Joo, S. H., Kirsch, I., & Halderman, L. (2020). Examining gender DIF and gender differences in the PISA 2018 reading literacy scale: A partial invariance approach. *Psychological Test and Assessment Modeling*, 62(2), 179-231.

Kim, S., & Moses, T. (2014). An investigation of the impact of misrouting under two-stage multistage testing: A simulation study: An investigation of the impact of misrouting. ETS Research Report Series, 2014(1), 1–13. https://doi.org/10.1002/ets2.12000

Kim, S., Moses, T., & Yoo, H. H. (2015). A comparison of IRT proficiency estimation methods under adaptive multistage testing: A comparison of IRT proficiency estimation methods. *Journal of Educational Measurement, 52*(1), 70–79. https://doi.org/10.1111/jedm.12063

Kirsch, I., & Lennon, M. L. (2017). PIAAC: a new design for a new era. *Large-Scale Assessments in Education*, *5*(1), 1-22. https://doi.org/10.1186/s40536-017-0046-6

Ling, G., Attali, Y., Finn, B., & Stone, E. A. (2017). Is a computerized adaptive test more motivating than a fixed-item test?. *Applied Psychological Measurement*, *41*(7), 495–511. https://doi.org/10.1177/0146621617707556

Luo, X., & Kim, D. (2018). A top-down approach to designing the computerized adaptive multistage test: Top-down multistage. *Journal of Educational Measurement*, *55*(2), 243–263. https://doi.org/10.1111/jedm.12174

Ma, Y. C. (2020). *Investigating hybrid test designs in passage-based adaptive tests.* [Doctoral dissertation, The University of Iowa]. https://doi.org/10.17077/etd.005590

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

Martin, A. J., & Lazendic, G. (2018). Computer-adaptive testing: Implications for students' achievement, motivation, engagement, and subjective test experience. *Journal of Educational Psychology*, *110*(1), 27–45. https://doi.org/10.1037/edu0000205

Mooney, C. Z. (1997). *Monte carlo simulation*. Sage.

Morris, T. P., White, I. R., & Crowther, M. J. (2019). Using simulation studies to evaluate statistical methods: Using simulation studies to evaluate statistical methods. *Statistics in Medicine*, *38*(11), 2074–2102. https://doi.org/10.1002/sim.8086

Patsula, L. N. (1999). *A comparison of computerized adaptive testing and multistage testing*. [Doctoral dissertation, University of Massachusetts Amherst].

Rotou, O., Patsula, L., Steffen, M., & Rizavi, S. (2007). Comparison of multistage tests with computerized adaptive and paper-and-pencil tests. ETS Research Report Series, 2007(1), 1–27. https://doi.org/10.1002/j.2333-8504.2007.tb02046.x

Sari, H. İ., & Huggins-Manley, A. C. (2017). Examining content control in adaptive tests: Computerized adaptive testing vs. Computerized adaptive multistage testing. *Educational Sciences Theory & Practice*, *17*(5), 1759–1781. https://doi.org/10.12738/estp.2017.5.0484

Şahin, M. G. (2020). Analyzing different module characteristics in computer adaptive multistage testing. *International Journal of Assessment Tools in Education*, *7*(2), 191–206. https://doi.org/10.21449/ijate.676947

Şenel, S. (2021). Bilgisayar ortamında bireye uyarlanmış testler [Computerized adaptive testing]. Pegem Yayınları.

Thompson, N. A., & Weiss, D. A. (2011). A framework for the development of computerized adaptive tests. *Practical Assessment, Research, and Evaluation*, *16*(1). https://doi.org/10.7275/wqzt-9427

Wainer, H., Mislevy, R. J., Steinberg, L., & Thissen, D. (2001). Review of computerized adaptive testing: a primer. *Language Learning & Technology*, *5*(2).

Wang, K. (2017). *A fair comparison of the performance of computerized adaptive testing snd multistage adaptive testing*. [Doctoral dissertation, Michigan State University].

Yigiter, M. S., & Dogan, N. (2023). Computerized multistage testing: Principles, designs and practices with R. Measurement: Interdisciplinary Research and Perspectives, 21(4), 254–277. https://doi.org/10.1080/15366367.2022.2158017

Zenisky, A., Hambleton, R. K., & Luecht, R. M. (2010). *Multistage testing: Issues, designs, and research*. J. Van der Linden (Ed.). Elements of adaptive testing, 355-372, Springer.