




<http://www.tayjournal.com>

<https://dergipark.org.tr/en/pub/tayjournal>

## **The Discrete Option Multiple Choice Items as A Measurement Instrument for Mathematics Achievement\***

 Atilla Özdemir, Asst. Prof., Corresponding Author  
Süleyman Demirel University, Türkiye  
[atillaozdemir@sdu.edu.tr](mailto:atillaozdemir@sdu.edu.tr)  
Orcid ID: 0000-0003-4775-4435

 Selahattin Gelbal, Prof. Dr.  
Hacettepe University, Türkiye  
[sgelbal@gmail.com](mailto:sgelbal@gmail.com)  
Orcid ID: 0000-0001-5181-7262

**Article Type:** Research Article

**Received Date:** 17.01.2024

**Accepted Date:** 24.06.2024

**Published Date:** 31.07.2024

**Plagiarism:** This article has been reviewed by at least two referees and scanned via a plagiarism software

**Doi:** 10.29329/tayjournal.2024.653.06

**Citation:** Özdemir, A., & Gelbal, S. (2024). The discrete option multiple choice items as a measurement instrument for mathematics achievement. *Türk Akademik Yayınlar Dergisi (TAY Journal)*, 8(2), 317-348.

\*This article was produced from Atilla Özdemir's doctoral thesis supervised by Prof. Dr. Selahattin Gelbal.

## Abstract

This study examines the applicability of Discrete Option Multiple Choice [DOMC] items in secondary school mathematics. The test included 25 questions, with 10 being traditional multiple-choice and 15 being DOMC items. Data were collected from 725 secondary school students during the second term of the 2020-2021 academic year. Among these students, 491 (68%) were in 7<sup>th</sup> grade and 234 (32%) were in 8<sup>th</sup> grade; 391 (54%) were female, and 334 (46%) were male. The findings revealed significant differences between the two item types, especially in high scores, using Classical Test Theory [CTT]. However, Item Response Theory [IRT] analysis showed that the question type did not affect estimations of students' ability levels, thus reducing errors in extreme values. This suggests that DOMC items do not significantly impact students' total scores when parameter estimations are performed using IRT instead of CTT. Additionally, some Traditional Multiple Choice [TMC] items were adapted into the DOMC format to test the applicability of various question types in this format.

**Keywords:** Classical test theory, discrete option multiple choice, item response theory, mathematics achievement, traditional multiple choice.

## Introduction

Student years are essential for observing the impact of assessments and evaluations of our lives. When national (Evaluation of High School Entrance Exam), (Examination for Transition to Higher Education), etc.) and international (Scholastic Aptitude Test [SAT]), (Graduate Record Examination [GRE]), etc.) standardized tests are considered, the measurement and evaluation processes become an influential agenda for all stakeholders in education (Cohen & Swerdlik, 2018; Erdoğan, 2003; Janda, 1997; Popham, 1999). Student success was evaluated based on the results of achievement tests. However, high-stakes tests, which are used to transition between levels and continue higher education, direct the educational processes with exams (Vidal Rodeiro & Macinska, 2022). Mathematics tests determine the decisive role of these examinations. The attainment of success by students in the field of mathematics holds paramount importance, not solely for academic achievements but also in sculpting the trajectories of their future professional careers (Forsblom et al., 2022; Wainer et al., 2015). Although only multiple choice questions were used in these exams, the answers did not show the details the students's mathematical thinking process. Therefore, students' abilities and real success cannot be precisely measured (Burt, 1911; Burt, 1972; Davis et al., 1993; Gooddenough, 1926; Lowell, 1919; Porteus, 1915; Woodworth, 1910).

Multiple choice tests are extensively utilized because of their ability to objectively evaluate, which is regarded as their most important feature (Baker, 2001). However, there are other sorts of multiple choice items, with Traditional Multiple Choice [TMC] items being the most common. Furthermore, multiple choice items can take various forms, such as matched multiple choice, best-answer, broad-matched, true-false, multiple true-false, content-dependent item sets, and Discrete Option Multiple Choice [DOMC] items (Foster & Miller, 2009).

Despite the widespread use of TMC, it has some limitations, the most notable being cheating and test wiseness. These limitations negatively affect tests's psychometric properties when there are measures other than the information a test wants to measure. In TMC tests, the

respondent chooses the alternative he thinks is most likely correct rather than directly revealing his knowledge. This selection was based on a comparison of all options at the same time. As a result, another critical disadvantage of the TMC item-containing test type is that it allows for the extraction or creation of indications pointing to the correct answer by comparing alternative answer options (Holmes, 2002). In this case, the person taking the test can answer questions using clues from the TMC item. The ability to answer questions using such clues in TMC items is called test wiseness (Gibb, 1964). Thus, individuals can find the right solution and increase their test scores by comparing the options without knowing the question using all available answer options (Bailey et al., 2022; Rost & Sparfeldt, 2007). The mere use of unwanted cues cannot decipher TMC items in carefully constructed and evaluated tests if practical item writing guidelines are followed (Adediwura et al., 2021; Haladyna & Downing, 2004). Even an experienced item writer, however, needs considerable effort to develop a genuine TMC item. Many TMC items are generated under time constraints by writers who need more excellent test development experience (Fagley, 1987). Thanks very much for your comment. The sentence has been corrected and written again (Foster & Miller, 2009). Many of these items can be answered without training or appropriate knowledge due to the wisdom cues they frequently contain (Alnasraween et al., 2022; Lions et al., 2022; Rotthoff et al., 2008).

Differences among individuals with test wiseness skills allow candidates with high test wiseness to be rewarded while punishing individuals who do not possess these skills (Baker & Baker, 2022; Taylor & Gardner, 1999). In psychometric terms, situations such as test wiseness and cheating appear as Construct-Irrelevant Variance [CIV] elements because they affect the results of test scoring containing TMC items (Guo et al., 2022). Therefore, each point obtained represents both knowledge and skill on the subject and CIV elements (test wiseness, cheating, etc.). Increasing CIV jeopardizes the test's construct validity (Haladyna & Downing, 2004; Zhai et al., 2021). One with high test wiseness or cheating can acquire dramatically different scores from people with the same degree of knowledge or skill. Given the scores, grades, certificates, or admissions obtained in this way, people with skills unrelated to the structure may come to the fore. Various TMC item formats have been proposed to reduce anxiety and to create validity and its effect on TMC test scores (Rodriguez, 2005).

As a result, tests with TMC items frequently contain solution hints and are thus sensitive to test wiseness. Similar problems are likely to be encountered, as these suggestions involve simultaneously presenting the test-taker with a choice of answers. The DOMC tests, as posited by Foster and Miller (2009), represent a prospective avenue for mitigating apprehensions related to test wiseness and cheating, presenting a viable substitute to the conventional TMC tests. Analogous to TMC items, DOMC items consist of a stem, the correct answer, and alternatives. However, they diverge fundamentally in two key aspects. Primarily, in DOMC items, response options are presented sequentially rather than simultaneously, each accompanied by a true or false designation. Participants are tasked with individually deciding on each option as it is presented sequentially, with the order randomized and no opportunity for revisiting or altering previous responses, a procedural characteristic exclusively applicable prospectively (Foster & Miller, 2009; Kingston et al., 2012). Secondly, the DOMC test employs three distinct conditions to conclude the processing of an item. Termination occurs

without further presentation of response options when any of these conditions are met: (a) the correct solution is completed (rendering additional options unnecessary), (b) the correct answer is rejected, or (c) the distractor is accepted as correct. In conditions (b) or (c) apply, additional answer options are deemed redundant, as the item has already been considered incorrect. Consequently, the presentation of options concludes upon the correct or incorrect resolution of DOMC item, in contrast to multiple-choice items that provide all answer options regardless of correctness. Foster and Miller (2009) further advocate for incorporating an additional option with a probability of .50 after the initial scoring of the item. This approach aims to diminish participants' ability to confidently discern the accuracy of their responses, thus enhancing the evaluative challenge. There are few studies on this subject (Bolt et al., 2012, 2018, 2020; Eckerly, 2017, 2018; Foster and Miller, 2009; Funk et al., 2010; Gorney and Wollack, 2022; Kingston et al., 2012; Papenberg et al., 2017, 2019; Papenberg, 2018; Willing et al., 2015), and there were no studies on DOMC item use in determining secondary school mathematics achievement.

This study examined the usability of DOMC items in measuring mathematics achievement. In this sense, comparing the psychometric properties of the DOMC item format, which is believed to provide an alternative solution to some of the limitations experienced in applying TMC items, will make an essential contribution to the literature. This study on the use of DOMC items, recently introduced in the literature, will contribute to the field. When the literature is examined, it is observed that there is no study at the middle school level, and at the high school level, only one study (Kingston et al., 2012) has been conducted. Furthermore, limited investigations on DOMC items based on actual data on importance and test criteria indicate that this study will substantially contribute to the literature.

To achieve this aim, the research problem is: How are DOMC and TMC test features compared to item response and classical test theories?

The subproblems of the problem statement are as follows.

1. How are the item and test characteristics of the DOMC and TMC tests compared to the CTT? Is there a statistically significant difference between the item difficulty indices?
2. According to the IRT what are the item and test parameters of DOMC and TMC tests?
3. How do the DOMC and TMC tests affect candidates' success? Is there a statistically significant difference between the students' test achievement scores?

## **Method**

This section provides information on the type of research, the study group, the data collection process, the data collection tool used, and the data analysis.

### **Research Model**

This study examined student mathematics achievement variations using TMC and DOMC test items. Success scores were compared to CTT and IRT, revealing similarities and differences. In this respect, a descriptive quantitative research model was used. The descriptive quantitative research model is a research method aimed at describing the current state of a specific group, situation, or event as it is. This model is used to describe, analyze, and draw

various conclusions about the current state of the research subject through observations related to this situation (Fraenkel et al., 2012).

### **Participants**

The study group consisted of 853 students in the 7<sup>th</sup> and 8<sup>th</sup> grades studying in five different secondary schools in Ankara. The study did not include one hundred twenty-eight data from these students with missing data problems. It was observed that all students with missing data issues had incomplete answers to the exam questions. Therefore, the studies of these students were not included in the research. The research groups are presented in Table 1.

**Table 1.**

*Study Group of the Research*

Gender/Grade level	7 <sup>th</sup>	8 <sup>th</sup>	Total
Female	261	130	391
Male	230	104	334
Total	491	234	725

Table 1 shows the 725 students in the study group. While 491 (68%) of the students were in the 7<sup>th</sup> grade, 234 (32%) were in the 8<sup>th</sup> grade. The study group included 391 (46%) female students and 334 (54%) male students.

### **Data Collection Process**

Research data was collected during the second semester of the 2020-2021 academic year. The data were collected through the “Scorpion™” platform prepared by Caveon for online testing (Caveon, 2020). This platform allows the test content to reach the students online. The application, planned to be carried out in schools due to the covid-19 pandemic, was conducted online through this platform. Questions were created, and data were collected using the same TMC and DOMC items platform.

Before the data collection process, a video introducing the platform interface was prepared for the students. A ten-question trial exam was prepared for students to recognize the system and have information about the multiple choice item format with discrete options. Along with the introductory video, the trial exam has also been activated, so preparations before the final application have been completed.

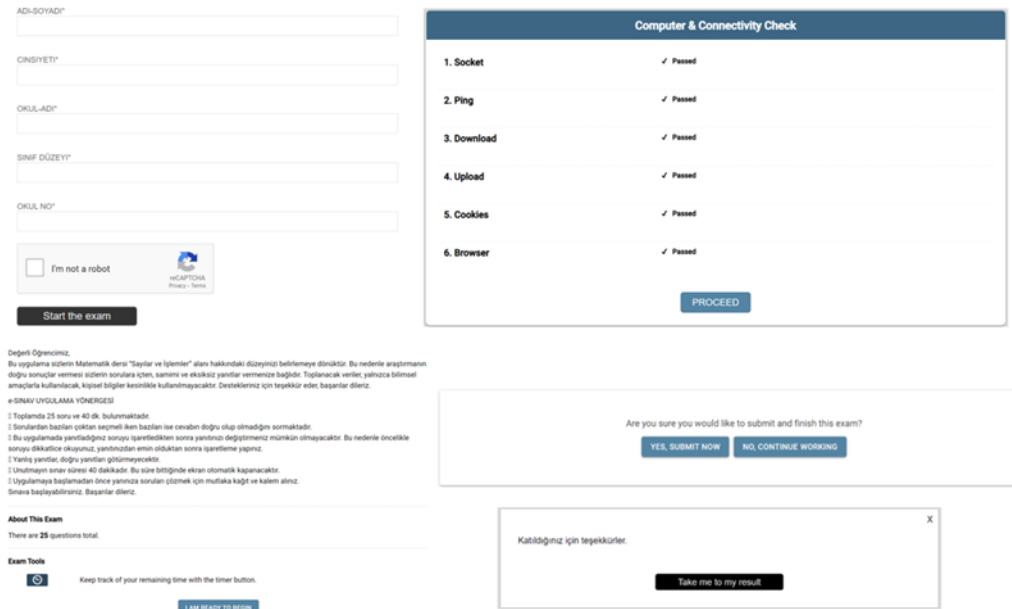
A ten-question trial exam was prepared to ensure students were familiar with the system and the DOMC item format. Alongside an introductory video, the trial exam was activated, thus completing the preparatory phase before the final implementation. The link to the online test prepared for the final application was shared with the school administrations. Some students accessed the system and completed the exam during math class hours in certain schools, while others used the application outside of regular class hours. Upon entering the system, students encountered a screen that progressed step by step (Figure 1). During the data collection phase, in the event of technical issues, the researcher intervened in the system to enable students to resume the exam.

On the screen with exam questions, the exam duration and number of questions are on the upper left screen. Accordingly, students can see which question they are in from the

question numbers on the upper-left screen. Although there were 25 questions in total in the exam, 15 consisted of DOMC, and 10 were TMC items. The number of items in DOMC is greater than the number of items in TMC. This is because there is a requirement in DOMC to write multiple items measuring the same question for some of the items in TMC. In this respect, the number of DOMC and TMC items in the system is the same. Another critical issue in the system is the difference in the order of questions for each student at this stage. The system automatically and randomly assigns questions to each student. Accordingly, each student's order of true and false statements on the DOMC items and distractors on the TMC items varied. The system completely automated this process. After seeing all the questions, the students who completed the 25<sup>th</sup> question received an exam completion warning on the next screen. Students who completed the exam were directed to the screen to learn about the results. At this stage, students can see the scores obtained from the exam.

**Figure 1.**

### Exam Screenshots



Students see their scores and exam times on the exam results screen. The data collection process was completed between 1-31 May 2021 for all schools participating in the research. The purpose of determining these dates was for all students to complete their lectures on the question items used. All students voluntarily participated in the study, and each student who completed the study was given a pencil.

### Data Collection Tool

As a data collection tool in the research, within the scope of TÜBİTAK's 1003 Priority Areas Research and Development Projects Support Program, "Investigation of Some Variables Regarding the Level of School and Students Affecting Turkish, Mathematics and Science Course Success and Development of Policy Suggestions for 7<sup>th</sup> and 8<sup>th</sup> Grade Students." (Project No. 117K851), a standardized mathematics test developed within the project's scope, was used as a data collection tool.



In the research, it was decided to utilize 20 selected items from the final multiple-choice test comprising 25 questions developed for the 7<sup>th</sup>-grade level within the scope of the project. However, due to the prolonged closure of schools resulting from the global covid-19 pandemic, the possibility of face-to-face implementation was eliminated. Consequently, it was decided to prepare the application in a single session, incorporating both discrete multiple-choice and multiple-choice items. To achieve this, it was agreed to use 10 items from the 20-item form. DOMC items were created for each selected TMC item using the same stem. For some TMC items, 2 or more questions were written as DOMC items. Table 2 shows the item specifications of the chosen TMC items according to CTT.

When examining Table 2, it can be inferred that the test items exhibit medium difficulty and discrimination according to CTT. Discrimination indices range from .28 to .53, while item difficulties vary between .32 and .74, and the KR-20 reliability coefficient was found to be .76.

**Table 2.**

*Data Collection Tool CTT Item Parameters*

Items	Discrimination	Item difficulty	Item standard deviation
I16	.38	.74	.44
I17	.51	.60	.49
I18	.33	.52	.50
I19	.32	.38	.49
I20	.44	.42	.49
I21	.28	.32	.47
I22	.53	.52	.50
I23	.53	.59	.49
I24	.45	.65	.48
I25	.48	.53	.50

When the data were examined according to IRT, a three-parameter model was observed according to the model data fit. The obtained data are listed in Table 3.

**Table 3.**

*Data Collection Tool IRT Item Parameters*

Items	a parameter (Discrimination)	b parameter (Item difficulty)	c parameter (Guessing)
I16	1.512	-.828	.097
I17	1.962	-.328	.000
I18	.893	-.117	.001
I19	2.326	1.021	.230
I20	2.889	.671	.191
I21	3.395	1.215	.214
I22	4.359	.282	.210
I23	3.173	-.008	.167
I24	2.652	-.054	.268
I25	3.198	.338	.227

Table 3 shows that parameter a varied between .893 and 4.359, parameter b varied between -.828 and 1.215, and parameter c varied between 0 and .268. Theoretically, the discrimination parameter (a) can range from 0 to positive infinity. Higher values indicate that the item is more effective at discriminating between individuals with different latent trait levels. The difficulty parameter (b) typically ranges from negative to positive infinity. Negative values indicate easier items and positive values indicate more difficult items. The guessing parameter (c) has a range from 0 to 1. A higher value of c indicates a higher chance of guessing

correctly (Baker & Kim, 2004). When Table 3 is examined, the questions are of medium difficulty and discrimination, and guessing is low.

The item characteristic curves of the data according to the IRT are shown in Figure 2.

**Figure 2.**

*Data Collection Tool Item Characteristic Curve*

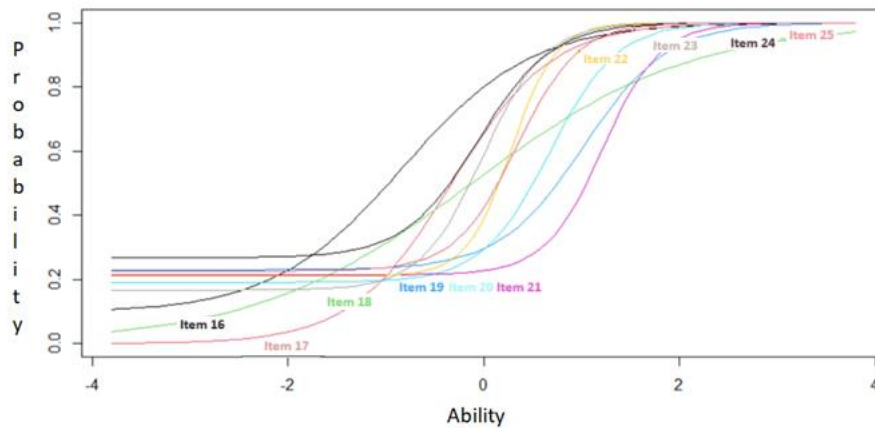


Figure 2 shows that the probability of solving the other items by chance is high, except for three items (items 16, 17 and 18), according to the item characteristic curves obtained according to the three PLM. In addition, it was found that the b parameters generally showed a distribution of approximately 0 (zero). Figure 1 shows that the slope parameters, indicators of the parameters, are also close to each other. The IRT reliability coefficient was calculated as .62.

### Preparing Discrete Option Multiple Choice Items

While preparing discrete options for multiple choice items, a plan was made to prepare at least one question for each TMC item. Owing to the nature of the DOMC, more than one DOMC was ready for some TMC items. While preparing the items, multiple choice item roots were preserved, and similar questions were created by making changes only in numerical information. Table 4 shows the sequence of the items in the applied test as well as their distribution according to multiple choice item groups.

Table 4 shows that the first 15 items in the test are multiple choice with discrete choice, while the latter ten are regular multiple choice, for a total of 25 items. Four DOMC items (1, 2, 3, and 4) were written to provide the sixteenth question from the TMC items. Similarly, two DOMC (5, 6) items were written in response to the seventeenth TMC question. Two DOMC (8, 9) items were written in response to the nineteenth TMC question. One DOMC item was written for the other TMC items on the test. The main reason for writing more than one question for some TMC items was to ensure that the questions could be measured similarly while being converted to DOMC items.



**Table 4.***Ranking and Distribution of the Items in the Test*

DOMC items	TMC items
I1	I16
I2	
I3	
I4	
I5	I17
I6	
I7	I18
I8	I19
I9	
I10	I20
I11	I21
I12	I22
I13	I23
I14	I24
I15	I25

When Table 4 is examined, for example, the 17<sup>th</sup> question prepared as a TMC question is given in Figure 3.

**Figure 3.***TMC Question 17*

**17.** Some equations are given below.

$$(14.8) + (8.\triangle) = 20.8$$

$$(\square.5) + (5.3) = 10.5$$

**Which of the following numbers should replace the symbols  $\triangle$  and  $\square$  in these equations?**

- |    |             |           |
|----|-------------|-----------|
|    | $\triangle$ | $\square$ |
| A) | 6           | 9         |
| B) | 5           | 7         |
| C) | 5           | 9         |
| D) | 6           | 7         |

For the TMC question 17 in Figure 3, questions 5 and 6 in the DOMC item type were given in Figure 4.

**Figure 4.***DOMC Questions 5 and 6*

**5.** The number to replace the symbol  $\square$  in the equation  $(12.6) + (6.\square) = 21.6$ ;

7

8

9

10

11

**6.** The number to replace the symbol  $\circ$  in the equation  $(\circ.7) + (7.8) = 12.7$ ;

2

3

4

5

6

## **Analysis of Data**

CTT and IRT were used to analyze the data to address research problems. Basic information about these theories is presented in detail on the theoretical basis of this study. Item and test statistics were calculated according to CTT. As a result, the item discrimination and difficulty of the items were determined, as were the average difficulty and discrimination of the test and Cronbach's alpha reliability coefficient. The model data fit for TMC items was analyzed using IRT, and item statistics (a parameter (discrimination), b parameter (item difficulty), c parameter (guessing), and test reliability coefficient) were provided. DOMC's model data fit was investigated using Testlet Response Theory [TRT], and the test's reliability coefficient was computed using item statistics.

## **Limitations and Assumptions**

We need to acknowledge the readers for certain limitations and assumptions before discussing the results. Although the TMC item test paper-pencil form of the research and the DOMC test was planned to be applied face-to-face at school with a computer program, because of the covid-19 pandemic conditions, it could be used by the students online. The data collection for this study was facilitated through the utilization of the "Scorpion™" platform developed by Caveon. It is imperative to acknowledge that a limitation inherent in the study arises from the automated structuring of the order of questions and options by the program. With these limitations, we made some assumptions in our study. It was assumed that each candidate performing the test was adequately performed. Each candidate was assumed to use computer communication technologies (Information and Communications Technology or Technologies [ICT]). It is assumed that each candidate answered questions independently. All participants are assumed to know basic expressions such as "YES" – "NO."

Although Turkish expressions have been added to computer programs, the interface is still in English.

## **Ethical Permits of Research:**

In this study, all the rules specified to be followed within the scope of "Higher Education Institutions Scientific Research and Publication Ethics Directive" were complied with. None of the actions specified under the heading "Actions Contrary to Scientific Research and Publication Ethics", which is the second part of the directive, have been taken.

## **Ethics Committee Permission Information:**

Name of the committee that made the ethical evaluation = Hacettepe University Ethics Committee

Date of ethical review decision= 25.02.2020

Ethics assessment document issue number= 51944218-300/00000987002

## Findings

This section analyzes the findings to answer the research questions, and the answers to each problem are presented. The findings obtained from the data analysis were converted into tables and graphics.

### Research Problem 1: Comparison of Item and Test Properties of DOMC and TMC Tests with CTT, and Examination of the Statistical Significance Between Item Difficulty Indices

The discrimination and difficulty indices of the DOMC and TMC test items and test statistics were calculated based on the CTT to answer the first research question above. The calculated items and test statistics are presented in Tables 5 and 6.

In Table 5, item discrimination and item difficulty indices for the DOMC items are presented. When the findings were examined, the lowest discrimination item was 13 (.05), and the highest was 5 (.58). When item difficulties were discussed, it was found that the easiest item was 3 (.74), and the most difficult item was 10 (.12). The mean discrimination in the DOMC items was .39, and the mean item difficulty was .43. For 15 DOMC items, Cronbach's alpha reliability coefficient was .78.

**Table 5.**

*Item and Test Statistics for Discrete Option Multiple Choice Items*

Items	Discrimination ( $r_{jx}$ )	Item difficulty ( $p_j$ )
1	.44	.58
2	.45	.57
3	.47	.74
4	.37	.49
5	.58	.57
6	.58	.57
7	.05	.24
8	.41	.33
9	.38	.33
10	.11	.12
11	.33	.15
12	.56	.34
13	.05	.22
14	.46	.59
15	.56	.55
<i>M</i>	.39	.43
Reliability (Cronbach alpha)	.78	

CTT presents the data obtained from TMC test items in Table 6. In Table 6, item discrimination and item difficulty indices for the TMC items are presented. When the findings were examined, the lowest discrimination item was 19 (.35), and the highest was 22 (.53). When the item difficulties were concerned, it was seen that the easiest item was 24 (.72), and the most difficult item was 21 (.32). The item difficulty was .53, while the mean item discrimination of multiple choice items was .44. Cronbach's alpha reliability coefficient for ten classic multiple choice items was .78.

**Table 6.**

*Item and Test Statistics for Traditional Multiple Choice Items*

Items	Discrimination (rjx)	Item Difficulty (pj)
16	.42	.64
17	.48	.62
18	.36	.54
19	.35	.39
20	.52	.33
21	.36	.32
22	.53	.54
23	.47	.65
24	.41	.72
25	.51	.52
<i>M</i>	.44	.53
Reliability (Cronbach alpha)	.78	

The z-ratio test examined whether a statistically significant difference existed between item difficulty indices. When the findings were examined, it was determined that the item difficulty indices calculated for both item types of all items except item 15 showed statistically significant differences. Accordingly, all item difficulties, except item 15, obtained from the TMC item type were found to be statistically significance, that is, easier than the DOMC. Table 7 lists the item difficulty indexes of the DOMC and TMC items.

**Table 7.**

*Comparison of Item Difficulty and Discrimination of DOMA and TMC Items*

Items TMC♦	Items DOMC♦♦	pj♦	pj♦♦	p	Cohen's h	rjx♦	rjx♦♦	z	Cohen's q
	1								
16	2	.64	.59	.03*	.09	.42	.43	-47	.01
	3								
	4								
17	5	.62	.57	.01*	.11	.48	.58	-5.34*	.14
	6								
18	7	.54	.24	0*	.61	.36	.05	16.21*	.32
	8								
19	9	.39	.33	0*	.14	.36	.39	-2.04*	-.05
	10								
20	10	.33	.12	0*	.51	.52	.11	21.29*	.46
	11								
21	11	.32	.15	0*	.40	.36	.33	1.35*	.03
	12								
22	12	.54	.34	0*	.41	.52	.56	-1.84	-.05
	13								
23	13	.65	.22	0*	.90	.47	.05	23.08*	.47
	14								
24	14	.72	.59	0*	.28	.41	.46	2.46*	.06
	15								
25	15	.52	.55	.23	-.06	.51	.56	-2.29*	-.06

♦TMC, ♦♦DOMC, \*p<.05

**Research Problem 2: Determination of Item and Test Parameters of DOMC and TMC Tests According to IRT**

The item and test parameters of the DOMC and TMC test items were calculated based on IRT to answer the second research question above. The data from the DOMC, according to IRT, were analyzed using the TRT model. Preliminary to the analyses, the assessment involved comparisons with diverse fit criteria to adjudicate the most appropriate model. Based on the results garnered, the 2PL-TRT model emerged as the optimal choice, as evidenced by its minimal AIC, BIC, and DIC values coupled with the highest log-likelihood value. Further scrutiny included an examination of S-X2 statistics to evaluate the conformity of item model-data fit. Upon review, it was ascertained that all items demonstrated compliance within the

framework of the 2PL-TRT model. When the model data fit was examined, it was observed that the data were compatible with the two-parameter TRT model (2PL-TRT). The obtained data are presented in Table 8.

**Table 8.**

*a and b Parameters according to TRT for DOMC Items*

Items	a parameter (Discrimination)	b parameter (Item difficulty)
I1	.814	.287
I2	.848	.255
I3	1.190	1.063
I4	.588	-.001
I5	1.721	.436
I6	1.689	.426
I7	.047	-.696
I8	.807	-.633
I9	.756	-.621
I10	.131	-1.166
I11	.583	-1.188
I12	1.334	-.635
I13	.029	-.767
I14	.898	.311
I15	1.359	.275
<i>M</i>	.852	-.176
Reliability coefficient	.810	

When Table 8 was examined, it was seen that the highest parameter was in item 5 (1.721), and the lowest was in item 13 (.029). These results are consistent with the CTT results. As shown in Table 8, the lowest b parameter was Item 11 (-1.188), and Item 3 (1.063) was the highest. These results differed from the CTT results. This is because the items were grouped while calculating the TRT parameters of items 1-4, 5-6, and 8-9. However, in the CTT, the parameters were calculated, assuming that each item is independent. The TRT reliability coefficient for the DOMC item score was .81.

When the model data fit of the data obtained from the TMC test items according to IRT was examined, it was observed that the data were compatible with the three-parameter IRT model (3PLM), and the item parameters were calculated within this framework. The obtained data are presented in Table 9.

**Table 9.**

*IRT Results on Traditional Multiple Choice Items*

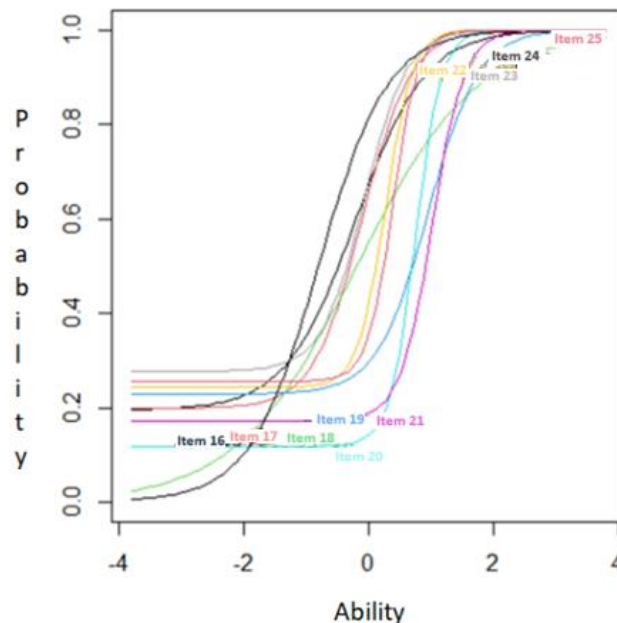
Items	a parameter (Discrimination)	b parameter (Item difficulty)	c parameter (Guessing)
I16	1.788	-.208	.193
I17	2.559	-.100	.199
I18	1.039	-.177	.004
I19	2.464	.958	.229
I20	4.865	.754	.119
I21	3.886	1.023	.173
I22	4.535	.272	.244
I23	2.934	-.071	.277
I24	1.825	-.799	.001
I25	4.927	.390	.257
<i>M</i>	3.082	.204	.169
Reliability coefficient		.610	

When Table 9 is examined, it is seen that the highest parameter is in item 25 (4.927), and the lowest parameter is in item 18 (1.039). In Table 9, it was found that the item with the highest b parameter was Item 21 (1.023), and the lowest was Item 24 (-.799). According to the results of the IRT, the easiest item was 24, while the most difficult item was 21. Upon scrutiny of the guessing parameter, it is observed that the minimum chance of guessing for an item is 24 (.001), whereas the maximum is 25 (.257). The computed KR-20 reliability coefficient for the TMC item scores is .61.

The item characteristic curves for TMC items are given in Figure 5. When Figure 5 was studied, it was noted that the likelihood of solving the other items by chance was relatively high, except for two items (items 18 and, 24), according to the item characteristic curves derived according to 3PLM. In addition, the b parameters generally exhibit a distribution of approximately 0 (zero). It can be asserted that the slope parameters, serving as indicators of the parameters of the items, exhibit a noteworthy proximity to one another.

**Figure 5.**

*Item Characteristic Curve of TMC Items*



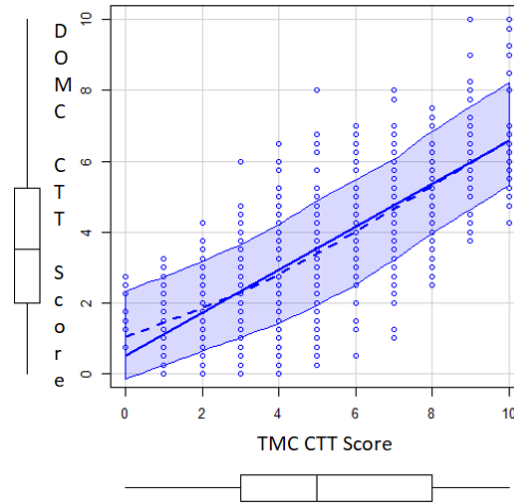
The critical finding obtained from the second subproblem of the study, when IRT calculated the scores obtained from the DOMC and TMC tests, gave close results compared to the raw scores obtained from CTT analyses. While there were significant differences between the two item types in the CTT analyses (Table 5, 6 and 7), particularly in high scores, the estimations of the student's ability levels in the IRT (Table 8 and 9) analyses were not affected by the question type, reducing the errors that may occur in extreme values.

**Research Problem 3: Investigating the Effect of DOMC and TMC Tests on Candidates' Achievement and Determining Statistically Significant Differences between Students' Test Achievement Scores**

To address this research problem, four graphs are examined. In the first graph, Figure 6 shows the raw scores obtained from the TMC test on the x-axis and the raw scores from the DOMC test on the y-axis.



**Figure 6.**  
CTT Score Chart



According to Figure 6, there is a linear relationship between the scores obtained from both test types. Observed scores are indicated by dashed blue lines in the middle of the figure, while blue lines in the middle indicate expected scores. The fact that these two lines almost overlap indicates that the expected and observed scores were very close. This shows a linear relationship between the TMC and DOMC scores. Most of the total scores are at the 95% lower and upper limits of the observed score line. Another remarkable situation in this graph is that the expected score linear line cuts the x-axis at an angle of less than  $45^\circ$ . This indicates that the total points students receive from the DOMC items are lower than those from the TMC items.

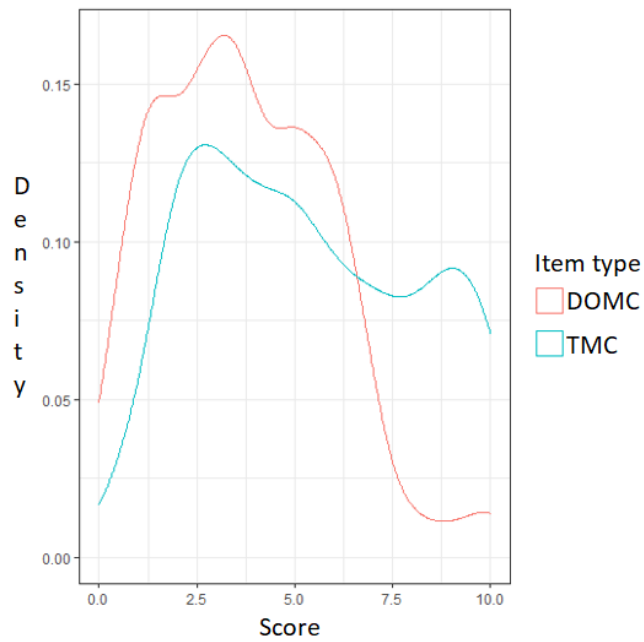
Figure 7 shows students' raw scores on the x-axis and student density on the y-axis. The red and blue lines represent the DOMC and TMC scores, respectively. Although the distributions of the scores obtained from both test types for low and medium scores were similar, student densities were higher in the DOMC-type items. This shows that students have more difficulties with the DOMC items. With a total score of 6.5, it was discovered that the number of students in the total scores of the two question types differed significantly. This shows that separation can be performed better in the DOMC test, especially with high scores.

Figure 8, on the x-axis, shows theta ability levels estimated according to the IRT obtained from the multiple choice test. On the y-axis are the theta ability levels estimated according to the IRT obtained from the discrete multiple choice test.

In Figure 9, on the x-axis, the students' IRT theta ability level values from the test are located, and on the y-axis, the student density.

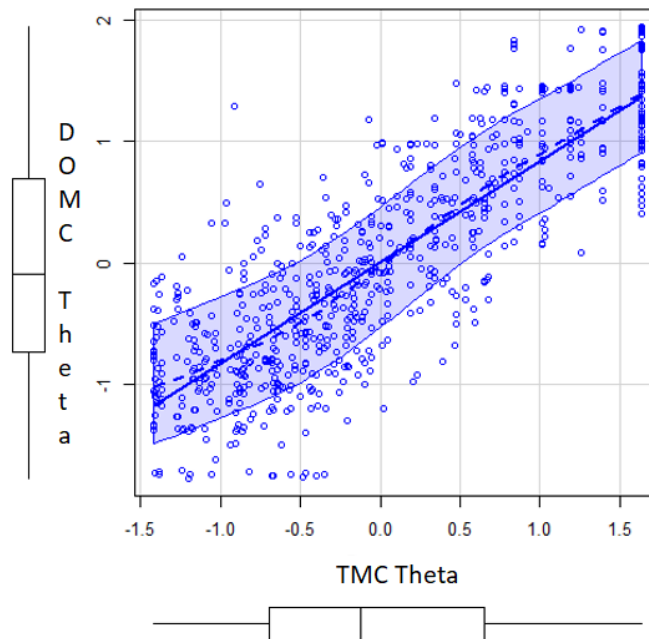
In Figure 9, like Figure 7, the red line indicates DOMC theta ability levels, and the blue line indicates TMC theta ability levels. The distribution of scores obtained from the two test types was consistent. The estimations of students' ability levels with the IRT were not affected by the type of question as with the CTT.

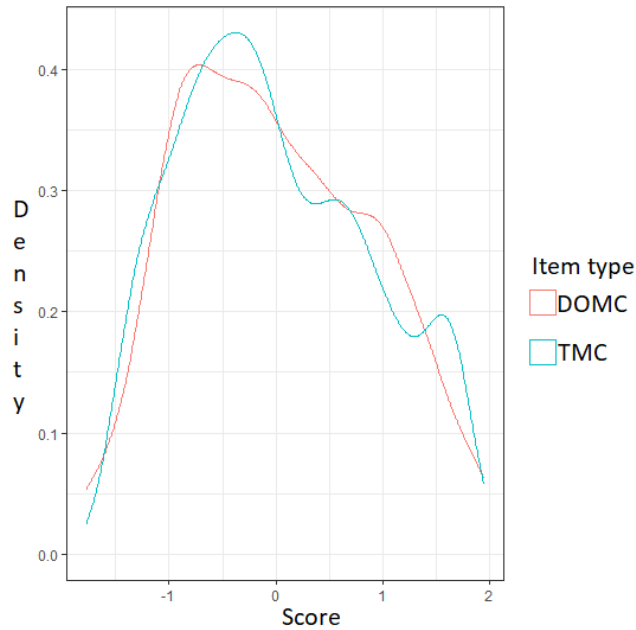
**Figure 7.**  
CTT Score Density



There was no statistically significant difference between IRT-calculated student theta levels ( $t(724) = -.34, p = .735, d = -.01$ ). According to the CTT, the total scores obtained from TMC items were statistically higher than the total scores obtained from the DOMC items ( $t(724) = 23.89, p < .001, d = .89$ ).

**Figure 8.**  
IRT Score Graph



**Figure 9.***IRT Theta Ability Level Density Graph*

## Discussion and Conclusion

This study examined the variation in students' mathematics achievement according to TMC and DOMC items, and answers were sought from research questions created within this framework. In this section, the results obtained based on the analysis of the research and the findings are presented by comparing them with other studies in literature. In addition, various suggestions were made for practitioners and researchers based on the information gained during this research.

In this study, the mathematics achievement of 725 secondary school students was compared using the DOMC and TMC item formats. The results obtained because of this study are as follows.

1. When scrutinizing the questions employed in the research, ten items were prepared for TMC, and parallel to these items, fifteen items were developed for DOMC (refer to Table 4). Notably, in the case of certain TMC items, it became imperative to compose multiple items adhering to the DOMC format rather than a singular item (refer to Figures 3 and 4). This distinction assumes significance, indicating a variance in the nature of DOMC items and the format employed in question construction compared to TMC optional items.

2. A vital feature of DOMC items over TMC items is that the correct and distracting numbers can be changed in their options. This feature of DOMC items has been used in most studies (Eckerly et al., 2017; Papenberg et al., 2017, 2019). In this study, as in previous studies, in writing the options for DOMC items, one correct and three incorrect options are presented for five items, one correct and four incorrect options for three items, one correct and five incorrect options for four items, one correct and six incorrect options for one item, two correct and four incorrect options are presented for one item, and three correct and four incorrect options are presented for one item. Since each participant has different options, DOMC items

provide a significant advantage over TMC items. It has been observed that more than one correct option can be written for some questions depending on the nature of the items so that, unlike TMC items, questions that do not have only one correct answer for an item can be produced.

3. DOMC items can be applied on a computer-based basis. For this, the software is required to write DOMC item-type questions. When the literature is examined, it is stated that different software (Webassessor™, Unipark, Macro-supported PowerPoint) are used, but the software used in many studies is not specified. This study used the Caveon Scorpion, one of the few software programs suitable for the DOMC item format. The researcher obtained a one-year free usage permit for this study. The software used is ideal for the simultaneous use of DOMC and TMC item formats. However, it has been observed that there is no user interface for organizing the data in the reporting process. Another situation is that the DOMC item type is patented, and using these items in a test or exam requires a license. This creates the DOMC item type at a disadvantage. Although there is no charge for the use of the item type for research and trial purposes, the limited number of test distribution software that supports DOMC item types, and the fact that they charge a fee in this regard, can be expressed as another critical disadvantage.

4. Throughout the course of the research, several challenges were encountered concerning the implementation of the DOMC item type. Primarily, participants accustomed to the TMC item format faced potential difficulties, prompting the initiation of a preliminary trial application to mitigate the impact of this circumstance. Secondly, challenges arose in the domain of item construction, where comprehensive parallelism between DOMC items and certain questions amenable to the TMC format proved unattainable. Furthermore, it is noteworthy that DOMC, as a computer-based item type, introduces potential complexities, particularly when implemented through specific software. The foremost challenge in this study pertains to the sequential presentation of options and, notably, the placement of the correct answer. Given the potential impact on participants and the consequential influence on the psychometric quality of tests incorporating DOMC items, addressing this aspect is paramount.

The results obtained from the first problem of the study are like the results in the literature, and it was observed that students had more difficulty in DOMC item types than TMC items (Foster & Miller, 2009; Funk et al., 2010; Kingston et al., 2012; Samuel & Hinson, 2012; Willing, 2013). This situation may depend on factors such as the ordered presentation of options in multiple-choice items, the possibility of multiple correct answers for certain questions, and the use of different options for each participant in DOMC items. Upon examination of item difficulty index comparisons, disregarding the effect size signs, it is observed that these comparisons span a range from .06 to .90. Consequently, it is discerned that 4 comparisons exhibit a small impact, 5 manifest a medium impact, and 1 reflects a high impact.

In Table 7, the item discrimination indexes for TMC test range from .36 to .52, while for DOMC test, the range is observed to be between .05 and .58. Upon comparing the item discrimination indexes of analogous items, a statistically significant difference is evident across all items, except for two. Despite variations in the item discrimination index for items

measuring the same construct, there exists a general proximity between the item discrimination indexes of the two tests. Upon scrutinizing the effect sizes irrespective of their directional signs, the observed range spans from .01 to .47. Hence, it is inferred that 7 comparisons yield a small effect, while 3 exhibit a medium effect.

Based on the findings obtained, the test scores were reliable. The reliability coefficients of the scores obtained for both item types were the same. Accordingly, while Cronbach's alpha reliability coefficient for the DOMC first 15 items and the last 10 TMC items of the test scores was .78, Cronbach's alpha reliability coefficient was .87 for the scores obtained from all questions in the test. When we look at the literature, a similar result was obtained in the study of Willing (2013). Considering the ascertained results, one may posit that the test scores are reliable.

The results obtained from the first problem of the study are like the results in the literature, and it was observed that students had more difficulty in DOMC item types than TMC items (Foster & Miller, 2009; Funk et al., 2010; Kingston et al., 2012; Samuel & Hinson, 2012; Willing, 2013). This situation may depend on factors such as the ordered presentation of options in multiple-choice items, the possibility of multiple correct answers for certain questions, and the use of different options for each participant in DOMC items. Upon examination of item difficulty index comparisons, disregarding the effect size signs, it is observed that these comparisons span a range from .06 to .90. Consequently, it is discerned that 4 comparisons exhibit a small impact, 5 manifest a medium impact, and 1 reflects a high impact.

In Table 7, the item discrimination indexes for TMC test range from .36 to .52, while for DOMC test, the range is observed to be between .05 and .58. Upon comparing the item discrimination indexes of analogous items, a statistically significant difference is evident across all items, except for two. Despite variations in the item discrimination index for items measuring the same construct, there exists a general proximity between the item discrimination indexes of the two tests. Upon scrutinizing the effect sizes irrespective of their directional signs, the observed range spans from .01 to .47. Hence, it is inferred that 7 comparisons yield a small effect, while 3 exhibit a medium effect.

Based on the findings obtained, the test scores were reliable. The reliability coefficients of the scores obtained for both item types were the same. Accordingly, while Cronbach's alpha reliability coefficient for the DOMC first 15 items and the last 10 TMC items of the test scores was .78, Cronbach's alpha reliability coefficient was .87 for the scores obtained from all questions in the test. When we look at the literature, a similar result was obtained in the study of Willing (2013). Considering the ascertained results, one may posit that the test scores are reliable.

When the literature on the findings obtained from the second problem of the study is examined, it is seen that ITC-based studies on DOMC focus on models related to the ordering of answer choices (Bolt et al., 2012; Bolt et al., 2018; Bolt et al., 2020). There is no comparative study on test and item parameters. Therefore, it is thought that the information obtained will contribute to the literature.

When the visualizations obtained from the third problem of the study are analyzed in general, using IRT instead of CTT to estimate the parameters does not lead to significant changes in students' total scores for the DOMC items. Using IRT decreases the possibility of errors, especially at extreme values. Estimates made using the CTT approach revealed significant disparities between the two item types, particularly for high scores. This indicates that students have more difficulty with DOMC items, which are a different type of question, than with TMC items. In a similar study with undergraduate students, Funk et al. (2010) stated that they preferred to use the TMC item format because the DOMC items were new to students and made it difficult to predict. However, Samuel and Hinson (2012) found in their study that the DOMC item format supported students' self-efficacy and intrinsic value.

Similar results were obtained in many studies when the data obtained were evaluated in general. Kingston et al. (2012) stated in their research that DOMC and TMC items measured similar structures, and TMC items were consistently easier than DOMC items. This is demonstrated in the present study. In their experimental studies, Willing et al. (2015) stated that test wiseness clues are less useful in the DOMC item format than in TMC items; therefore, DOMC items are more difficult than TMC items. In the findings obtained in the first subproblem of the study, when TMC and DOMC item difficulties were compared, it was found that DOMC items were difficult in a statistically significant way.

Another important finding of this study was related to the TMC items used. It tested whether different question types applied to the DOMC item format based on some of the TMC items used in the study. In the second subproblem, the reliability results differed for both theories. When TMC and DOMC item forms were presented in equal numbers in the literature, the findings could not be compared to those of a different study.

To date, all studies in literature have been conducted at undergraduate and higher levels. The DOMC item types could be used at the secondary school level, and no problems were encountered during the application. Thus, the usability of tests containing DOMC items for different question types and educational levels was demonstrated.

## **Recommendations**

The DOMC item format offers an essential alternative to the TMC item format, which has been used for nearly a century. However, there need to be more studies on issues such as question writing, software to be used, analysis methods, and the order of options (Bolt et al., 2018; Bolt et al., 2020). As a result, in this study, in contrast to the literature on DOMC items, various findings were revealed by making subject area, question contents, software used, grade level, and comparative analyses. The following suggestions are presented to practitioners and researchers for studies on this subject.

Recommendations for practitioners are as follows:

When studies related to DOMC items are examined, it is observed that applications are made in psychology, medicine, information technologies, the German language, and mathematics. Conducting studies on DOMC items in different fields and the content of the questions will increase our knowledge of the use of these items.



Another critical issue in the studies is that DOMC items must be delivered to the individuals who will take the test via computer-based software, so developing software on this subject is essential. A limited number of software is used in the studies carried out so far.

Considering that the groups in which DOMC items are applied are undergraduate, graduate, and adult groups in the current studies, it is evident that there is a need for studies at the K-12 level in which multiple choice tests are frequently used.

When comparing DOMC items with TMC items, preparing and applying DOMC parallel forms for different TMC items is essential. This will provide more insight into the nature of the DOMC item format.

Recommendations for researchers are as follows:

Considering the sample sizes studied on DOMC items, repeating studies in larger sample groups and testing the existing findings will positively contribute to the literature.

Considering that IRT-based studies on DOMC items are very limited in the literature, it is essential to conduct comparative studies to analyze the data obtained.

Another crucial future study area is to conduct studies to determine the effect of DOMC items on affective characteristics other than academic achievement and contribute to the limited literature on this subject.

One of the most significant criticisms of DOMC items is the order effect of options. The studies on the order effect of the options and the software to be developed based on these studies can popularize the use of DOMC items. However, studies on this subject could be more extensive in the literature.

In the context of DOMC applications, instead of scoring the entire item 0-1, the exploration of alternative scoring methodologies, such as nominal response or partial credit model for each option or combination can provide additional information.

## References

- Adediwura, A. A., Ajayi, O. S., & Ibikunle, Y. A. (2021). Students and test variables as predictors of undergraduates' self-compassion. *Journal of Research & Method in Education*, 11(3), 42-48.
- Alnasraween, M. S., Alsmadi, M. S., Al-zboon, H. S., & Alkurshe, T. O. (2022). The level of universities students' test wiseness in Jordan during distance learning in light of some variables. *Education Research International*, 1-10.
- Anastasi, A. (1988). *Psychological testing* (6<sup>th</sup> ed.). Macmillan.
- Bailey, C. D., Briggs, J. W., & Irving, J. H. (2022). Test-wiseness and multiple-choice accounting questions: Implications for instructors. *Issues in Accounting Education*, 37(2), 1-14.
- Baker, D. L., & Baker, R. L. (2022). Knowledge and wisdom: High stakes testing and learning outcomes. In *Neuroethical policy design. Studies in brain and mind*. (pp.101-118). Springer.
- Baker, F. B. (2001). *The basics of item response theory* (ED458219). ERIC. <https://eric.ed.gov/?id=ED458219>
- Baker, F. B., & Kim, S. H. (Eds.). (2004). *Item response theory: Parameter estimation techniques* (2<sup>nd</sup> ed.). CRC.
- Bolt, D. M., Kim, N., Wollack, J., Pan, Y., Eckerly, C., & Sowles, J. (2020). A psychometric model for discrete-option multiple-choice items. *Applied Psychological Measurement*, 44(1), 33-48.
- Bolt, D. M., Lee, S., Wollack, J., Eckerly, C., & Sowles, J. (2018). Application of asymmetric IRT modeling to discrete-option Multiple-Choice Test items. *Frontiers in Psychology*, 9, 1-7.
- Bolt, D. M., Wollack, J. A., & Suh, Y. (2012). Application of a multidimensional nested logit model to multiple-choice test items. *Psychometrika*, 77, 263-287.
- Bowman, M. L. (1989). Testing individual differences in ancient China. *American Psychologist*, 44(3), 576-578.
- Burt, C. (1911). Experimental tests of higher mental processes and their relation to general intelligence. *Journal of Experimental Pedagogy*, 1, 93-112.
- Burt, C. (1972). Inheritance of general intelligence. *American Psychologist*, 27(3), 175-190.
- Caveon. (2020). Technology and internet-based services subscriber agreement. Retrieved October 10, 2023, from <https://caveon.com/caveon-privacy-policy-for-students-of-education-subscribers/>
- Cohen, R. J., & Swerdlik, M. E. (2018). *Psychological testing and assessment* (9<sup>th</sup> ed.). McGraw-Hill Education.
- Cronbach, L. (1990). *Essentials of psychological testing* (5<sup>th</sup> ed.). Harper & Row.
- Davis, N. T. (1996). Transition from objectivism to constructivism in science education. *International Journal of Science Education*, 15(6), 627-636.
- Eckerly, C., Smith, R. W., & Sowles, J. (2017, September 7). *Analysis of the discrete option multiple choice item: Examples from its certification*. Conference on Test Security, Madison, WI.
- Eckerly, C., Smith, R., & Sowles, J. (2018). Fairness concerns of discrete option multiple-choice items. *Practical Assessment, Research & Evaluation*, 23(16), 1-10.
- Erdoğan, İ. (2003). *Çağdaş eğitim sistemleri* [Contemporary education systems] (5<sup>th</sup> ed.). Sistem.
- Fagley, N. S. (1987). Positional response bias in multiple-choice tests of learning: Its relation to test wiseness and guessing strategy. *Journal of Educational Psychology*, 79(1), 95-97.
- Forsblom, L., Pekrun, R., Loderer, K., & Peixoto, F. (2022). Cognitive appraisals, achievement emotions, and students' math achievement: A longitudinal analysis. *Journal of Educational Psychology*, 114(2), 346-367.
- Foster, D. F., & Miller, H. L. (2009). A new format for multiple-choice testing: Discrete option multiple-choice. Results from early studies. *Psychology Science Quarterly*, 51, 355-369.

- Fraenkel, J. R., Wallen, N. E., & Hyun, H. H. (2012). *How to design and evaluate research in education* (8<sup>th</sup> ed.). McGraw Hill.
- Funk, R., Hooper, T., Hadlock, E., Whicker, J., Estes, D., & Miller, H. L. (2010). *Differential effects of the discrete-option multiple-choice format on test takers' assessment preparation and scores*. Poster presented at the Mary Lou Fulton Undergraduate Research Conference, Provo, UT.
- Gibb, B. G. (1964). *Testwiseness as a secondary cue response* (Publication No. 6407643) [Doctoral dissertation, Stanford University]. ProQuest Thesis Center.
- Goodenough, F. L. (1926). *Measurement of intelligence by drawings*. World Book Company.
- Gorney, K., & Wollack, J. A. (2022). Does item format affect test security? *Practical Assessment, Research, & Evaluation*, 27(15), 1-13.
- Guo, H., Rios, J. A., Ling, G., Wang, Z., Gu, L., Yang, Z., & Liu, L. O. (2022). Influence of selected-response format variants on test characteristics and test-taking effort: An empirical study. *ETS Research Report Series*, 2022(1), 1–20.
- Güler, N. (2011). *Eğitimde ölçme ve değerlendirme* [Measurement and evaluation in education]. Pegem.
- Haladyna, T. M., & Downing, S. M. (2004). Construct-irrelevant variance in high-stakes testing. *Educational Measurement: Issues and Practice*, 23(1), 17–27.
- Holmes, P. (2002). *Multiple evaluation versus multiple choice as testing paradigm* [Doctoral dissertation, Twente University]. University of Twente Research Information.
- Janda, L. H. (1997). *Psychological testing: Theory and applications* (1<sup>st</sup> ed.). Pearson.
- Kingston, N. M., Tiemann, G. C., Miller, H. L., & Foster, D. (2012). An analysis of the discrete-option multiple-choice item type. *Psychological Test and Assessment Modeling*, 54(1), 3–19.
- Kumandaş, H., & Kutlu, Ö. (2010). High stakes testing: Does secondary education examination involve any risks? *Procedia- Social and Behavioral Sciences*, 9, 758-764.
- Lions, S., Monsalve, C., Dartnell, P., Blanco, M. P., Ortega, G., & Lemarié, J. (2022). Does the response options placement provide clues to the correct answers in multiple-choice tests? A systematic review. *Applied Measurement in Education*, 35(2), 133–152.
- Lowell, F. (1919). A preliminary report of some group tests of general intelligence. *Journal of Educational Psychology*, 10(7), 323–344.
- Papenberg, M. (2018). *On how test wiseness and acceptance reluctance influence the validity of sequential knowledge tests* [Unpublished doctoral dissertation]. Heinrich-Heine-University Düsseldorf.
- Papenberg, M., Diedenhofen, B., & Musch, J. (2019). Experimental validation of sequential multiple-choice tests. *Journal of Experimental Education*, 89(2), 402–421.
- Papenberg, M., Willing, S., & Musch, J. (2017). Sequentially presented response options prevent the use of testwiseness cues in multiple-choice testing. *Psychological Test and Assessment Modeling*, 59(2), 245-266.
- Popham, W. J. (1999). *Modern educational measurement: Practical guidelines for educational leaders* (3<sup>rd</sup> ed.). Pearson.
- Porteus, S. D. (1915). Mental tests for the feebleminded: A new series. *Journal of Psycho-Asthenics*, 19, 200-213.
- Rodriguez, M. C. (2005). Three options are optimal for multiple-choice items: A meta-analysis of 80 years of research. *Educational Measurement Issues and Practice*. 24(2), 3–13.
- Rost, D. H., & Sparfeldt, J. R. (2007). Reading comprehension without reading? On the construct validity of multiple-choice reading comprehension test items. *Zeitschrift für Pädagogische Psychologie*, 21, 305-314.

- Rotthoff, T., Fahron, U., Baehring, T., & Scherbaum, W. A. (2008). The quality of CME questions as a component part of continuing medical education--an empirical study. *Zeitschrift für Ärztliche Fortbildung und Qualität im Gesundheitswesen*, *101*, 667-674.
- Samuel, J., & Hinson, J. (2012, March 5). *Promoting motivation through technology-based testing*. In P. Resta (Ed.), *Proceedings of the Society for Information Technology & Teacher Education International Conference* Chesapeake, VA: AACE.
- Taylor, C., & Gardner, P. L. (1999). An alternative method of answering and scoring multiple-choice tests. *Research in Science Education*, *29*, 353-363.
- Vidal Rodeiro, C., & Macinska, S. (2022). Equal opportunity or unfair advantage? The impact of test accommodations on performance in high-stakes assessments. *Assessment in Education: Principles, Policy & Practice*, *29*(4), 462-481.
- Wainer, H., Dorans, N. J., Eignor, D. R., Flaugher, R. L., Green, B. F., Mislevy, R. J., Steinberg, L., & Thissen, D. (2015). *Computerized adaptive testing a primer* (2<sup>nd</sup> ed.). Routledge.
- Willing, S. (2013). *Discrete-option multiple-choice: Evaluating the psychometric properties of a new method of knowledge assessment* [Unpublished doctoral dissertation]. Heinrich-Heine-University Düsseldorf.
- Willing, S., Ostapczuk, M., & Musch, J. (2015). Do sequentially presented answer options prevent the use of test wiseness cues on continuing medical education tests? *Advances in Health Sciences Education: Theory and Practice*, *20*(1), 247-263.
- Woodworth, R. S. (1910). Race differences in mental traits. *Science*, *31*(788), 171-86.
- Zhai, X., Haudek, K. C., Wilson, C., & Stuhlsatz, M. (2021). A framework of construct-irrelevant variance for contextualized constructed response assessment. *Frontiers in Education*, *6*, 1-13.

## **BIOGRAPHICAL NOTES**

### **Contribution Rate of Researchers**

Author 1: 50%

Author 2: 50%

### **Conflict Statement**

There is no conflict of interest in the research.

# Matematik Başarısı için Bir Ölçme Aracı Olarak Ayrık Seçenekli Çoktan Seçmeli Maddeler



## Özet

Bu çalışmanın amacı Ayrık Seçenekli Çoktan Seçmeli [ASÇS] maddelerin uygulanabilirliğinin incelenmesidir. Bu amaca ulaşmak için ortaokul matematik dersi kapsamında toplam 25 sorudan oluşan bir test kullanılmıştır. Testi oluşturan maddelerden 10 tanesi çoktan seçmeli maddelerden oluşurken 15 tanesi ASÇS maddelerden oluşmaktadır. Araştırmanın verileri 2020-2021 eğitim-öğretim yılı 2. döneminde ortaokulda öğrenim görmekte olan 725 öğrenciden elde edilmiştir. Öğrencilerin 491 (%68) tanesi 7. sınıf düzeyinde iken 234 (%32) tanesi 8. sınıf düzeyindedir. Çalışma grubundaki öğrencilerin 391 (%54)'ini kız öğrenciler, 334 (%46)'sını ise erkek öğrenciler oluşturmaktadır. Araştırmadan elde edilen bulgular incelendiğinde Klasik Test Kuramı [KTK] ile yapılan analizlerde özellikle yüksek puanlarda iki madde türü arasında büyük farklılıklar gözlenirken, Madde Tepki Kuramı [MTK] ile yapılan analizlerde öğrencilerin yetenek düzeylerinin kestirimlerinin soru tipinden etkilenmediği böylece uç değerlerde oluşabilecek hataları düşürdüğü gözlenmiştir. ASÇS maddelerinin, KTK yerine MTK ile parametre kestirimlerinin gerçekleştirilmesiyle öğrencilerin toplam puanlarında çok büyük bir farklılığa yol açmayacağı söylenebilir. Çalışmada kullanılan bazı Geleneksel Çoktan Seçmeli [GÇS] maddeler, ASÇS madde formatında iki veya daha fazla soru olacak şekilde seçilmiş ve farklı soru türlerinin ASÇS madde formatında uygulanabilirliği test edilmiştir.

**Anahtar Kelimeler:** Klasik test kuramı, ayrık seçenekli çoktan seçmeli, madde tepki kuramı, matematik başarısı, geleneksel çoktan seçmeli.

## Giriş

Okul yıllarımıza kadar fark etmesek de hayatımızın her alanında karşımıza çıkacak olan ölçme ve değerlendirme süreçleri yaşamımızın doğal bir parçası haline gelmiştir. Günlük hayatımızın olağan akışında sıklıkla ölçme-değerlendirmeler yapar ve kararlar alırız. Örneğin seçeceğimiz ayakkabı için ayak ölçümüze uygun olan ayakkabıyı inceleriz, bir futbol sahası yapmamız için en ve boy ölçülerine ihtiyaç duyarız, trafik düzenini sağlamak için kullandığımız lambalar için zaman ölçülerinden yararlanarak verimli bir model geliştirmeye çalışırız. Bunlar gibi birçok farklı alanda ölçme-değerlendirme işleminden faydalanırız. Farklı alanlarda yapılan tüm bu ölçmeler kendine özgü bir ölçme aracına ve ölçme birimine sahiptir. Bu araç ve birimler bizlere ölçmelerimizi doğrudan, dolaylı ya da türetilmiş şekilde yapmaya olanak tanır. Doğrudan ölçmelerde ölçülmek istenilen özellik doğrudan gözlenebilirken, dolaylı ölçmelerde ölçülmek istenen özelliğin ölçülüp gözlenmesi bir başka özelliğin yardımıyla ölçülebilir. Türetilmiş ölçmelerde ise ölçülmek istenilen özellik kendisinden farklı iki ya da daha fazla özelliğin arasındaki matematiksel bir bağlantıyla ölçülebilir (Güler, 2011).

Ölçme ve değerlendirmenin günlük hayattaki etkisinin fark edildiği en önemli süreç öğrencilik yıllarıdır. Bunun en önemli nedeni ders başarılarının değerlendirilmesinde öğrencilere uygulanan başarı testlerinden elde edilen ölçümlerin kullanılmasıdır. Bununla birlikte kademeler arası geçiş ve üst öğrenime devam etmek için kullanılan yüksek risk içeren (high stakes tests) (Kumandaş ve Kutlu, 2010) ulusal standart testlerde düşünüldüğünde



ölçme ve değerlendirme süreçleri eğitimin bütün paydaşlarının önemli bir gündemi haline gelmektedir. Geliştirilen başarı testlerinin tamamında amaç psikolojik yapıların ölçülmesi olduğundan tarihsel süreç içerisinde gelişimleri paralellik göstermiştir.

Testler genel anlamdan bireyleri tanımak ve bilgi sahibi olmak için kullanılırken (Cronbach, 1990), psikolojik testler bireye ait zeka, yetenek, beceri, tutum vb. davranışların standart ölçümlerini ifade etmektedir (Anastasi, 1988). Psikolojik test ve test programlarının kullanımı kaynaklarda MÖ 2200 yıllarındaki antik Çin'e kadar dayandırılır (Cohen & Swerdlik, 2018; Janda, 1997; Popham, 1999). Uygulanan testler subay ve sivil memur seçimleri için kullanılan ve imparatorluk sınavları ismi verilen bir takım zorlu süreçleri içermekteydi. Testlerin uygulama yöntemleri açısından günümüzle benzerlikleri dikkat çekicidir. Bu durumun en basit açıklaması ise Çinliler tarafından geliştirilen sınav uygulama prosedürleri ve esas alınan psikometrik nitelikler Fransa (1791), Hindistan (1833) ve Amerika Birleşik Devletleri (1883) gibi ülkelerde benzer uygulamalar için temel teşkil etmiştir. Sınav uygulamalarında ise aday isimlerinin gizli tutulması, sınav salonlarının her bir aday için benzer koşullarda, özel sınav binalarında bulunan küçük odacıklarda sınavların yapılması, sınav kağıtlarını puanlarken en az iki bağımsız değerlendiricinin olması gibi günümüzdeki uygulamalara benzer süreçler geliştirildiği görülmektedir (Bowman, 1989; Cohen & Swerdlik, 2018).

1900'lü yılların başlarında test süreçleriyle ilgili hem istatistiksel çalışmalar hem de farklı becerilerle ilgili performansları ölçebilmek için farklı türlerde testlerin ortaya çıktığı görülmektedir (Burt, 1911; Burt, 1972; Goodenough, 1926; Lowell, 1919; Porteus, 1915; Woodworth, 1910). Bu gelişmelerle birlikte testlerin bireysel olarak değil toplu olarak uygulanmaya başlamış, grup testleriyle birlikte çoktan seçmeli testlerin kullanımı yaygınlaşmıştır. Bu konuda ilk defa ABD'de 1901 yılında üniversiteye giriş sınavında uygulaması başlatılmış devam eden süreçte ise bu konuda bir komite kurularak 1926 yılında "Scholastic Aptitude Test [SAT]" isimli bir yetenek testi geliştirilerek kullanılmaya başlanmıştır. Testin kullanımı giderek yaygınlaşmış ve sadece üniversiteye girişte değil bursların verilmesinde de etkili bir rol oynamaya başlamıştır (Wainer vd., 2015). Günümüzde Amerika'daki pek çok üniversite öğrenci kabulünde ortaöğretim notları, SAT, Graduate Record Examination [GRE] ve Graduate Management Admission Test [GMAT] gibi yetenek ve başarı testlerinden alınan puanları kullanmaktadır. Bu süreçte tavsiye mektupları da göz önünde bulundurulurken, bu belgelere ilave olarak bazı üniversiteler ayrı bir seçme sınavı da uygulayabilmektedir (Erdoğan, 2003).

Çoktan seçmeli testlerin bu kadar geniş bir kullanım alanına sahip olmasının en temel nedeni objektif olarak değerlendirilmeleridir (Baker, 2001). Çoktan seçmeli maddelerin çeşitli türleri olmasına rağmen, Geleneksel Çoktan Seçmeli [GÇS] maddeler sıklıkla kullanılır. GÇS maddelerin yaygın olarak kullanılmasına rağmen bazı sınırlılıklarının olduğu gözlemlenmiştir. En bariz olanlar test bilgeliği ve kopyadır. Test bilgeliği becerisine sahip bireyler arasındaki farklılıklar, test bilgeliği yüksek olan adayları ödüllendirirken, bu beceriye sahip olmayanları cezalandırmaktadır (Baker & Baker, 2022; Taylor & Gardner, 1999). Psikometrik açıdan test bilgeliği ve kopya çekme gibi durumlar, GÇS maddeleri içeren test değerlendirmelerinin sonuçlarını etkiledikleri için yapıyla ilgisiz varyans öğeleri olarak karşımıza çıkmaktadır (Guo

vd., 2022). Bu nedenle, elde edilen her puan, ilgili bilgi, beceri ve ölçülmemiş varyansın bileşenlerini (test bilgeliği, kopya vb.) temsil eder. Yapıyla ilgisiz varyansın arttırılması, testin yapı geçerliliğini tehdit eder (Haladyna & Downing, 2004; Zhai vd., 2021). Aynı bilgi ve beceri düzeyine sahip iki kişiden, test bilgeliği yüksek olan veya kopya çeken aday, önemli ölçüde farklı puanlar alabilir. Bu şekilde elde edilen puanlar, dereceler, sertifikalar veya kabuller göz önüne alındığında, ölçülen yapıyla ilgisi olmayan becerilere sahip kişiler ön plana çıkabilir. Bu sınırlamalar, bir testin ölçmek istediği bilgilerden başka ölçümler olduğunda testin psikometrik özelliklerini olumsuz etkiler.

Ayrıca GÇS öğeleriyle yapılan testler genellikle çözüme yönelik ipuçları içerir ve bu nedenle test bilgeliği stratejilerine karşı savunmasızdır. Bu noktada literatürde son yıllarda yer bulmaya başlayan “Ayrık Seçenekli Çoktan Seçmeli [ASÇS]” madde içeren testler önemli bir alternatif olarak karşımıza çıkmaktadır. Bu araştırma, ASÇS madde içeren testleri kullanarak, GÇS madde içeren testlere yöneltilen test bilgeliği ve kopya eleştirileri için çözümler sunmaktadır. Literatür incelendiğinde ASÇS maddelerle ilgili çalışmaların sınırlı sayıda olduğu görülmektedir (Bolt vd., 2012, 2018, 2020; Eckerly vd., 2017, 2018; Foster & Miller, 2009; Funk vd., 2010; Kingston vd., 2012; Papenberg, 2018; Papenberg vd., 2019; Willing vd., 2015). Ayrıca ortaokul matematik başarısını belirlemede ASÇS madde kullanımına ilişkin herhangi bir çalışmada bulunmamaktadır. Ek olarak, ASÇS maddelerinin önemi ve test parametreleri hakkındaki gerçek verilerden elde edilen sınırlı çalışmalar, bu çalışmanın literatüre önemli ölçüde katkı sağlayacağını göstermektedir.

Bu nedenlerle bu çalışmada ASÇS maddelerin matematik başarısını ölçmede kullanılabilirliğini incelenmiştir. Bu amaca ulaşmak için nicel araştırma türlerinden betimsel model kullanılmıştır (Fraenkel vd., 2012).

Araştırmanın temel problemi şu şekildedir: ASÇS ve GÇS test özellikleri madde yanıt ve klasik test teorilerine göre nasıldır?

Problem cümlesinin alt problemleri aşağıda verilmiştir:

1. ASÇS ve GÇS testlerin madde ve test özellikleri Klasik Test Kuramına [KTK] göre nasıldır? Madde güçlük indeksleri arasında istatistiksel olarak anlamlı bir fark var mıdır?

2. ASÇS ve GÇS testlerin madde ve test parametreleri Madde Tepki Kuramına [MTK] göre nasıldır?

3. ASÇS ve GÇS testlerin adayların başarısına olan etkisi nasıldır? Öğrencilerin testlerden aldıkları başarı puanları arasında istatistiksel olarak anlamlı bir farklılık var mıdır?

## **Yöntem**

Bu çalışmada, GÇS ve ASÇS test maddeleri kullanılarak öğrencilerin matematik başarısındaki değişimler incelenmiştir. Başarı puanları KTK ve MTK ile karşılaştırılarak benzerlikler ve farklılıklar ortaya çıkarılmıştır. Bu bağlamda, betimsel bir nicel araştırma modeli kullanılmıştır (Fraenkel vd., 2012).

### **Araştırmanın Etik İzinleri:**

Bu çalışmada "Yükseköğretim Kurumları Bilimsel Araştırma ve Yayın Etiği Yönergesi" kapsamında uyulması gerektiği belirtilen tüm kurallara uyulmuştur. Yönergenin ikinci bölümü olan "Bilimsel Araştırma ve Yayın Etiğine Aykırı Eylemler" başlığı altında belirtilen eylemlerin hiçbiri gerçekleştirilmemiştir.

### **Etik Kurul İzin Bilgileri:**

Etik değerlendirmeyi yapan kurulun adı = Hacettepe Üniversitesi Etik Komisyonu

Etik Kurul Etik inceleme karar tarihi= 25.02.2020

Etik değerlendirme belgesi konu numarası= 51944218-300/00000987002

### **Bulgular**

Araştırma sonucunda elde edilen bulgular genel olarak değerlendirildiğinde, parametre kestirimi KTK yerine MTK ile yapıldığında ASÇS maddelerinin öğrencilerin toplam puanlarında anlamlı bir farklılığa neden olmayacağı söylenebilir. MTK kullanımının özellikle uç değerlerde oluşabilecek hataları azalttığı görülmüştür. KTK yöntemi ile yapılan tahminlerde iki madde türü arasında özellikle yüksek puanlarda anlamlı farklar olduğu görülmüştür. Bu durum öğrencilerin farklı bir soru türü olan ASÇS maddelerde, GÇS maddelere göre daha fazla zorlandıklarını göstermektedir. Lisans öğrencileriyle yapılan benzer bir çalışmada Funk vd. (2010), ASÇS maddelerinin öğrenciler için yeni olması ve tahmin etmeyi zorlaştırması nedeniyle GÇS madde formatını kullanmayı tercih ettiklerini belirtmişlerdir. Ancak Samuel ve Hinson (2012) çalışmalarında ASÇS madde formatının öğrencilerin öz yeterliklerini ve içsel değerlerini desteklediğini bulmuşlardır.

### **Tartışma ve Sonuç**

Çalışmanın sonuçları literatürdeki birçok çalışma ile benzer sonuçların elde edildiği görülmüştür. Kingston vd. (2012) araştırmalarında ASÇS ve GÇS maddelerinin benzer yapıları ölçtüğünü ve TMK maddelerinin ASÇS maddelerine göre sürekli olarak daha kolay olduğunu belirtmişlerdir. Bu çalışmada da benzer bir sonuç bulgulanmıştır. Deneysel çalışmalarında, Willing vd. (2015), test bilgeliği ipuçlarının ASÇS madde formatında GÇS maddelerine göre daha az yararlı olduğunu, bu nedenle ASÇS maddelerinin GÇS maddelerine göre daha zor olduğunu belirtmiştir. Araştırmanın birinci alt probleminde elde edilen bulgularda, GÇS ve ASÇS madde güçlükleri karşılaştırıldığında, ASÇS maddelerinin istatistiksel olarak anlamlı bir şekilde zor olduğu bulunmuştur.

Araştırmanın bir diğer önemli bulgusu da kullanılan GÇS maddeleri ile ilgilidir. Çalışmada kullanılan bazı GÇS maddeleri, ASÇS madde formatında iki veya daha fazla soru olarak seçilmiş ve farklı soru türlerinin ASÇS madde formatında uygulanabilirliği test edilmiştir. Araştırmanın ikinci alt probleminde ise her iki teoride güvenilirlik sonuçları farklılık göstermektedir. Literatürde GÇS ve ASÇS madde formları eşit sayıda sunulduğunda bulgular farklı bir çalışma ile karşılaştırılamamıştır.

Şimdiye kadar literatürde yer alan çalışmaların tamamı lisans ve üzeri düzeyde uygulanmıştır. ASÇS madde türleri ortaokul düzeyinde kullanılabilmektedir ve uygulama

sırasında herhangi bir sorunla karşılaşılmamıştır. Böylece ASÇS maddeleri içeren testlerin farklı soru türlerinde ve eğitim düzeylerinde kullanılabilirliği ortaya konmuştur.

Araştırmada, öğrencilerin matematik başarılarının GÇS ve ASÇS test maddelerine göre değişimi incelenmiş ve bu çerçevede oluşturulan araştırma sorularına yanıt aranmıştır. Araştırmanın analizlerine ve elde edilen bulgulara dayalı olarak ulaşılan sonuçlar alan yazındaki diğer çalışmalarla karşılaştırılmış ve araştırma süresince kazanılan bilgiler ışığında, uygulayıcılara ve araştırmacılara yönelik çeşitli öneriler getirilmiştir. Bu çalışmada ortaokul düzeyinde 725 öğrencinin matematik başarıları ASÇS ve GÇS madde formatları ile karşılaştırılmıştır. Çalışma sonucunda elde edilen sonuçlar:

1. Araştırmada kullanılan sorular açısından incelendiğinde GÇS seçmeli on madde için paralel on beş madde oluşturulmuştur. Literatürden farklı olarak GÇS maddelerden bazıları için birden fazla ASÇS madde yazılması gerekmiştir. Bu durum ASÇS maddelerin doğası ve soru yazım formatının GÇS maddelerden farklı olduğunu göstermesi açısından önemlidir.

2. ASÇS maddelerin GÇS maddelerden güçlü bir özelliği seçeneklerinde doğru ve çeldirici sayılarının değiştirilebilmesidir. Literatürdeki çalışmalarında birçoğunda kullanılan maddelerde ASÇS maddelerin bu özelliği kullanılmıştır (Eckerly vd., 2017; Papenberg vd., 2019; Papenberg vd., 2017). Bu çalışmada literatürdeki çalışmalara benzer şekilde ASÇS maddelerin seçeneklerinin yazılmasında beş madde için bir doğru üç yanlış seçenek, üç madde için bir doğru dört yanlış, dört madde için bir doğru beş yanlış, bir madde için bir doğru altı yanlış, bir madde için iki doğru dört yanlış, bir madde için üç doğru dört yanlış seçenek sunulmuştur. Her bir katılımcıya farklı sayıda seçenek gösterildiği düşünüldüğünde GÇS maddelere göre ASÇS maddelerin önemli bir avantaj sağladığı görülmektedir. Maddelerin doğasına bağlı olarak bazı sorular için birden fazla doğru seçeneğin yazılabildiği böylece GÇS maddelerden farklı olarak bir madde için sadece bir doğru cevabın olmadığı soruların üretilmediği görülmüştür.

3. ASÇS maddeler bilgisayar tabanlı olarak uygulanabilmektedir. Bunun için ASÇS madde türü soruların yazılabileceği bir yazılıma ihtiyaç vardır. Literatür incelendiğinde farklı yazılımlar (Webassessor™, Unipark, Makro destekli power point) kullanıldığı ifade edilmiş ancak birçok çalışmada kullanılan yazılım belirtilmemiştir. Bu çalışmada ASÇS madde formatına uygun az sayıda yazılımlardan biri olan Caveon Scorpion kullanılmıştır. Araştırmacı tarafından bu çalışma için bir yıllık ücretsiz bir kullanım izni alınmıştır. Kullanılan yazılım ASÇS madde formatı ve GÇS madde formatının aynı anda kullanımına uygundur. Bununla birlikte raporlama sürecinde verilerin düzenlenmesi için çok pratik bir ara yüzünün olmadığı görülmüştür. Bir diğer önemli konu ise ASÇS madde türünün patentli ve bu maddelerin bir test veya sınavda kullanılması için bir lisans gereğinin olmasıdır. Bu durum ASÇS madde türü için bir dezavantaj oluşturmaktadır. Her ne kadar araştırma ve deneme amaçlı olarak madde türünün kullanımından bir ücret talep edilmese de ASÇS madde türlerini destekleyen test dağıtım yazılımlarının sınırlı sayıda olması ve bu konuda ücret talep etmeleri de bir diğer önemli dezavantaj olarak ifade edilebilir.

Sonuç olarak bu çalışmada ASÇS maddelere ilişkin literatürden farklı olarak konu alanı, soru içerikleri, kullanılan yazılım, sınıf düzeyi ve karşılaştırmalı analizler yapılarak çeşitli

bulgular ortaya konulmuştur. ASÇS madde formatının yaklaşık yüzyıldır kullanılan GÇS madde formatı için önemli bir alternatif sunduğu görülmektedir. Bununla birlikte soru yazımı, kullanılması gereken yazılım, analiz yöntemleri ve sınırlı sayıda çalışma olan seçenek sıralaması (Bolt vd., 2018; Bolt vd., 2020) gibi konularda yeterince çalışmanın olmadığı görülmektedir.

## **Öneriler**

ASÇS madde formatı, yaklaşık bir asırdır kullanılan GÇS madde formatına önemli bir alternatif sunmaktadır. Ancak soru yazımı, kullanılacak yazılım, analiz yöntemleri, seçeneklerin sıralanışı gibi konularda daha fazla çalışmaya ihtiyaç duyulmaktadır (Bolt vd., 2018; Bolt vd., 2020). Sonuç olarak bu çalışmada ASÇS maddeler ile ilgili literatürden farklı olarak konu alanı, soru içerikleri, kullanılan yazılım, sınıf düzeyi ve karşılaştırmalı analizler yapılarak çeşitli bulgular ortaya konulmuştur. Bu konuda yapılacak çalışmalar için uygulayıcılara ve araştırmacılara aşağıdaki öneriler sunulmuştur.

Uygulayıcılar için öneriler şunlardır:

ASÇS maddeleri ile ilgili çalışmalar incelendiğinde psikoloji, tıp, bilişim teknolojileri, Alman dili ve matematik alanlarında uygulamalar yapıldığı görülmektedir. Farklı alanlardaki ASÇS maddeleri ve soruların içeriği üzerine çalışmalar yapılması bu maddelerin kullanımına ilişkin bilgimizi artıracaktır.

Çalışmalardaki bir diğer kritik konu ise ASÇS maddelerinin bilgisayar tabanlı yazılımlar aracılığıyla sınava girecek bireylere ulaştırılması gerekliliğidir, dolayısıyla bu konuda yazılım geliştirilmesi elzemdir. Şu ana kadar yapılan çalışmalarda sınırlı sayıda yazılım kullanılmıştır.

Mevcut çalışmalarda ASÇS maddelerinin uygulandığı grupların lisans, lisansüstü ve yetişkin grupları olduğu düşünüldüğünde, çoktan seçmeli testlerin sıklıkla kullanıldığı K-12 düzeyinde de çalışmalara ihtiyaç olduğu açıktır.

ASÇS maddeleri ile GÇS maddelerini karşılaştırırken, farklı GÇS maddeleri için ASÇS paralel formlarının hazırlanması ve uygulanması önemlidir. Bu, ASÇS madde formatının doğası hakkında daha fazla bilgi sağlayacaktır.

Araştırmacılar için öneriler şu şekildedir:

ASÇS maddeleri üzerinde çalışılan örneklem büyüklükleri dikkate alındığında, çalışmaların daha büyük örneklem gruplarında tekrarlanması ve mevcut bulguların test edilmesi literatüre olumlu katkı sağlayacaktır.

Alanyazında ASÇS maddeleri üzerinde MTK temelli çalışmaların oldukça sınırlı olduğu göz önünde bulundurulduğunda, elde edilen verilerin analiz edilmesi için karşılaştırmalı çalışmaların yapılması elzemdir.

Gelecekteki önemli çalışma alanlarından bir diğeri de ASÇS maddelerinin akademik başarı dışındaki duyuşsal özellikler üzerindeki etkisini belirlemeye yönelik çalışmaların yapılması ve bu konudaki sınırlı literatüre katkı sağlanmasıdır.

ASÇS maddelerine yönelik en önemli eleştirilerden biri seçeneklerin sıra etkisidir. Seçeneklerin sıra etkisi üzerine yapılacak çalışmalar ve bu çalışmalara dayalı olarak

geliştirilecek yazılımlar ASÇS maddelerinin kullanımını yaygınlaştırabilir. Ancak literatürde bu konudaki çalışmalar daha kapsamlı olabilir.

ASÇS uygulamaları bağlamında, maddenin tamamını 0-1 puanlamak yerine, her bir seçenek veya kombinasyon için nominal yanıt veya kısmi kredi modeli gibi alternatif puanlama yöntemlerinin araştırılması ek bilgi sağlayabilir.