BOĞAZİÇİ UNIVERSITY JOURNAL OF EDUCATION
2025, Vol. 42-4, No. 1
https://doi.org/10.52597/buje.1521294

The Comparison of Methods in Multiple Response Item Type Scoring[a]

Ömer Doğan[b] and İsmail Karakaya[c]

Abstract

The aim of this study is to examine the scoring methods of the multiple-answer item type, where a question can have more than one correct answer. Concordantly, the differences regarding test difficulty, reliability, item difficulty and discrimination among eight scoring methods to score achievement tests comprising of multiple response item types were examined and compared. According to the data obtained, it is understood that the multiple-answer item type is applicable for both numerical and verbal lessons. In addition, it has been found that standardized scoring methods are in a way that allows the use of different scoring methods by offering alternatives according to the purpose of the measurement and evaluation, without being bound by a single rule. Thus, scoring methods can be determined according to the purpose of using the multiple-answer item type, and the use of the item type can be made widespread with similar studies.

Keywords: multiple response item type, partial scoring, scoring methods

Article info

Received: 23.07.2024
Revised: 25.12.2024
Accepted: 17.03.2025
Published online: 30.04.2025

Introduction

In parallel with the abundance and variety of decisions on the students in education, several information gathering ways have been benefited. Based on a certain course, the behaviours aimed to make the students acquire and their levels of acquirement is measured through achievement tests (Koç, 2009; Tan, 2010). At the present time, multiple choice item type springs to mind, in the first place, among the applications of achievement tests. Multiple choice items have been commonly utilized as a measurement instrument within in- class achievement measurements and large-scale studies. This is mainly due to the fact that multiple choice items have numerous advantages. These items may be used for diagnosis and for formative objectives. Besides, it is possible to measure them easily, quickly, detachedly and economically through the people or measurement tools. Such advantages make those items suitable for a wide range of objectives from in- class achievement tests to large- scale standard tests and enable them to be a prepotent test technique (Auer & Tarasowa, 2013; Baghaei & Amrahi, 2011; Ben-Simon et al. 1997; Frisbie & Sweeney, 1982; Jodoin, 2003; Ma, 2004; Wan & Henly, 2012). The structure of multiple-choice item type has certain characteristics causing it both to be prepotent and to be criticized. Tan (2010) suggested that multiple choice item types were suitable for the behaviours concerned with knowledge, comprehension and application levels although it was not appropriate to evaluate the creativity levels of respondents. However, Umay (1997) asserts that the biggest problem with multiple choice tests is they do not allow monitoring the examinees' thinking processes and answering behaviours. On the contrary, she adds that the dearth of this situation provides objectivity during scoring. This is considered as the dilemma of multiple-choice tests. Multiple-choice questions do not offer students the opportunity to explain their answers, potentially limiting the depth and breadth of knowledge gathered from them. They also struggle to assess certain aspects of inquiry-based science, such as complex reasoning or coherent understanding (Liu et al., 2011). According to Siddiqui et al. (2016), an individual may be seeking the correct item choice by revising the options while solving multiple choice tests. First, s/ he may eliminate the incorrect ones, then, focus upon less number of item choices in order to decide upon the correct one. This refers to the luck in success and indicates that chance success may be increased. Baykul (2015) stated that the chance success makes the items easier and reduces item difficulty index. This decline varies based on different numbers of item choices. Chance success minimizes covariance among items in addition to declining test reliability (in split- half reliability, parallel forms and KR (Kuder- Richardson)

[a] *This article is derived from Ömer Doğan's (2020) master's thesis.*

[b] *Corresponding author, Ministry of National Education, 64omerdogan64@gmail.com, ORCID: 0000-0001-5169-520X*
[c] *Gazi University, Department of Educational Sciences, ikarakaya2002@gmail.com, ORCID: 0000-0003-4308-6919*

To cite: Dogan, O. & Karakaya, I. (2025). The comparison of methods in multiple response item type scoring. *Boğaziçi University Journal of Education*, *42-4*(1), 23-45. https://doi.org/10.52597/buje.1521294

reliability). Moreover, chance success decreases test validity based on the levels of response of mean and variances belonging to the predictor and criterion.

Traditionally, multiple choice tests have been constructed in a way that correct answers are scored with a value of one whereas incorrect responses (included blank and omitted items) with a value of zero. A major concern about this scoring method is that it is unable to allocate partial point to the respondents. Therefore, the respondents are only able to get a one point on the condition that s/he gives the correct answer; in such method, it is not possible to put forward a scoring system that reveals and rewards his/ her partial knowledge about a certain subject. There are several other types of multiple-choice items in terms of structure. As a result, multiple choice items are divided into three groups as correct answer, stem and a set of responses. The items are divided into five groups depending on the quality and number of correct answer (Doğan, 2009; Turgut & Baykul, 2013). In this study, multiple response item type, which is included in abovementioned groups, has been examined and certain comparisons concerning scoring methods in multiple response item type have been made. Multiple response item types, in which more than one of the provided options may be correct, allow the students to choose more than one answer. It is essential to select correct options and to leave all the incorrect options unmarked in this item type, which enables partial credit, to award the examinee full credit. Multiple response item type may be treated as a set of true- false options consisting of a stem, an instruction, and a few true or false options. The examinees respond correctly on the condition that they select correct options and leave the incorrect ones unmarked (Verbic, 2012). In case that there is more than one correct option in an item, the examinee's approach would be different from the item with single- correct alternative. Here, the examinee may want to eliminate the alternatives. Yet, s/he would investigate each option's probability of accuracy separately. The accuracy of an alternative would not ensure the examinee that s/he does not require to consider the following option (Siddiqui et al., 2016). The broader solution field in multiple response item type minimizes the examinees' guessing behaviours and, by this way, the students are able to more easily differentiate what they have learnt. In addition, the tests having multiple response items may be designed in such a way that they may allow an efficient feedback application about a broad field (Peterson et al., 2016). Two types of multiple response items are generally used in test implementation. The examinees are either informed about the number of true alternatives or s/he is requested to select all the alternatives considered to be correct without reporting the number of true alternatives. Multiple response items provide examinees to respond in various levels of their versatility and allow the stages of cognitive process to be systematically meaningfully monitored as well as enabling a valid measurement of examinees' levels of knowledge (Ma, 2004).

Resnick (1991) articulated that higher- order thinking processes mostly led to multiple solutions and entailed control over the thinking process and higher- level of endeavour (as cited in Doğanay, 2007). Thanks to flexible thinking, included in the skills of higher- order thinking, a multifaceted perspective may be adopted instead of dealing with the situations or problems from one point of view. Individuals with flexible thinking are able to create alternative ideas, maintain open- mindedness and consider alternatives with the awareness of having different options (Duman, 2018). As a consequence, in the age of rapid changes, tendency towards certain implementation fostering flexible thinking skills is needed in order to bring up individuals who are capable of selecting information consciously, having flexible thinking skill and proposing alternative solutions. The fact that multiple response item type could be used for this aim has been reported by the researchers in the literature (Hsu et al., 1984; Ma, 2004; Verbic, 2012). The tests with multiple response items are suitable to use apart from both paper- based and computer- based summative tests. For example, in formative tests, like standard setting or diagnosis tests, monitor student progress, focusing on what they are or are not able to do rather than to what extent they know. In recent years, multiple response items are worth highlighting due to their flexibility in answering and item presentation that could not be achieved through conventional type of multiple-choice items. Moreover, it has been acknowledged that multiple response items are easier to measure complicated talents, knowledge and skills rather than multiple choice items. Besides, multiple response item type demonstrates superiority compared to performance assessment tasks in terms of the accuracy and economy of scoring (Ma, 2004).

The use of multiple response item type is not as prevalent as that of other types in spite of its sound characteristics. According to Siddiqui et al. (2016), scoring system was a significant factor at this point. The dearth of practical and applicable method is responsible factor for the fact that multiple response items are not generally used. There is no agreement and, correspondingly, common scoring model concerning the scoring of multiple response items.

The multiple response items both performing objective, rapid and economic scoring and offering partial credit information have a number of scoring methods (Domnich et al., 2015; Verbic, 2012). The reasons why scoring methods show differences are related to various scenarios as how to distribute partial credit, how to make

corrections if necessary and under what circumstances zero point would be given. Although certain scoring methods do not make corrections, some are able to make plenty of corrections; one single false selection results in no mark in certain methods whereas it does not in some scoring methods. Therefore, it may be alleged that various scoring methods may be employed according to the intended purpose of the test. The eight scoring methods used in this study are as follows:

1. Scoring Model- Dichotomous: To get full points, all correct options must be marked and incorrect options mustn't be marked.
2. Scoring Model – Polytomous Trapdoor: Each correct alternative marked is valued one point; however, any incorrect marking is given zero point.
3. Scoring Model – Negative Scoring: Each correct alternative marked is valued one point although any incorrect marking is given -1 point.
4. Scoring Model – Multiple True/False: While one point is given for each correct option marked, one point is given for each incorrect option not marked.
5. Scoring Model – Positive Count: Each correct alternative marked is valued one point. Yet, no scoring is performed in cases of the selection of incorrect option or leaving the options unmarked (Muckle, Becker & Wu, 2011).
6. Scoring Model – Morgan Algorithm One point is awarded for each correct option marked, -1 point is awarded for each incorrect option marked, and no action is taken for unmarked options. (Morgan, 1979).
7. Scoring Model – Ripkey Algorithm: Points are awarded as much as the ratio of the number of correct options selected to the total number of correct options. Zero points are given for markings made more than the total number of correct options.
8. Scoring Model – Balanced Scoring: Points are awarded as much as the ratio of the number of correct options selected to the total number of correct options. Penalty points are applied for markings made more than the total correct number of options. The penalty score differs depending on the number of options marked.

The ratio of the sum of correct alternatives selected and incorrect alternatives left unmarked by the examinee in Multiple True/ False scoring method to the number of alternatives is between 0-1. Thus, the comparison of scoring methods is facilitated. Auer and Tarasowa (2013) and Siddiqui et al. (2016) made transformation in this way while comparing the scoring methods in their studies. Likewise, when Positive Count method is adapted in a way that it is distributed between 0-1, it is quite enough to rate the number of correct alternatives selected by the examinee to the number of correct alternatives. Negative Scoring method, as conducted by Domnich et al. (2015), can be transformed into the standard score interval, between 0-1. For this aim, as stated by Muckle et al. (2011), the number of incorrect alternatives marked is subtracted from the number of correct alternatives in a way that minimum score is zero and the score is divided into the number of alternatives. On the other hand, in Polytomous Trapdoor, the number of correct alternatives selected by the examinee is compared to the number of correct alternatives. However, zero point is given if the examinee selects an incorrect alternative. Through these transformations, eight scoring methods are standardized.

It is crucial to determine which method is more reliable, valid and useful compared to the others in the scoring of an achievement test consisting of multiple response items. The awareness regarding alternative scoring methods of multiple response items and the comparison and examination of their impact on test and item's statistics may be guiding for those who would utilize this test type. The present study will be undertaken to standardize the scoring methods in order to eliminate the inconsistency in scoring, which is most criticized, and the difficulties of statistical analyses, and, based on this standard, to make comparisons and to examine item and test characteristics of eight scoring methods that could be standardized.

The purpose of the study is to compare the scoring methods used in the scoring process of multiple response item type which is generally used in achievement tests and in- class assessment and evaluation practices and does not have important lacks and limitations of multiple-choice items. In addition, the study attempts to introduce this item type, to define its characteristics and to extend its use in assessment and evaluation practices of multiple response items. In the present study, the answer to the question which is 'Is there a significant

difference among the methods used in the scoring of multiple response item type in terms of test and item's statistics?' has been sought. In this regard, the answers to the following questions have been sought:

1. Is there a statistically significant difference among difficulties of test scores, obtained through various scoring methods, of a test prepared by using multiple response item type?
2. Is there a significant difference among reliability coefficients of test scores, obtained through various scoring methods, of a test prepared by using multiple response item type?
3. Is there a statistically significant difference among item difficulty indices calculated through the methods used in the scoring of multiple response item type?
4. Is there statistically significant different among item discrimination indices calculated through the methods used in the scoring of multiple response item type?

Method

Research Design

The purpose of the present study is to compare and examine the methods in multiple response item type scoring, administered to achievement tests of eighth grade students, in the context of test and item statistics. To achieve this objective, basic research model was administered in the study. Basic research produces knowledge for theory development. The aim of this research is to add to existing knowledge. The understanding of "knowledge for knowledge's sake" is fundamental (Karasar, 2012).

Study Group

The study group of the current research consists of the eighth-grade students who were going to take the high school entrance exam. The participants aim to enrol in a school whose admission is determined by the exam. The study group involves voluntary students. Therefore, the study group has been generated by using purposive sampling method, one of the non- probability sampling methods.

Data Collection Instruments

The achievement tests developed by the researcher on the basis of the acquisitions in Mathematics and Turkish courses have been employed as data collection instruments. The tests consist of multiple response item types which are in accordance with eighth grade. In this study, item pool regarding Turkish and Mathematics was composed to constitute the tests. The items have been presented to a group of 11 experts' point of views. Following the analyses on the first implementation, 3 items in Turkish and Mathematics tests have been changed and then, the second implementation has been realized. As a result of item analyses at the end of the second implementation, the reliability of achievement tests and item statistics have been found to be at appropriate levels and, thus, the tests were implemented as they were.

Data Collection

The current study was conducted through using achievement tests consisting of multiple response items in accordance with the acquisitions of Mathematics and Turkish courses in eighth grades in 2019- 2020 academic year. The frequencies regarding pilot study and main study are presented in Table 1.

Table 1

The Frequencies Concerning Pilot Study and Main Study

Study	The Number of Students Performing Tests	
	Mathematics	Turkish
Study I	158	134
Study II	88	92
Main Study	263	277

The Turkish and mathematics achievement tests, which were finalized after the applications, were administered to eighth grade students at one-week intervals. The application was carried out by the researcher him/herself and with the experience obtained from the trial applications, the Turkish test was applied first and the mathematics test was applied the week after the application. In addition, in order to ensure students' motivation for the test, an introductory presentation including the areas and exams where multiple response item type is used abroad was prepared and presented before the application.

Data Analysis

In the first stage of the study, the data were evaluated using descriptive survey and inferential statistics methods. The data in the second stage of the study consisted of the responses of eighth grade students to achievement tests prepared in accordance with the achievements of mathematics and Turkish lessons. The achievement tests were scanned with the Zipgrade program and transferred to the computer environment.

For the statistical procedures related to the first question of the research, it is necessary to determine the average difficulty and reliability indices of the tests. The average difficulty of the test is obtained by dividing the sum of the difficulty indices of the items in the test by the number of items in the test. The relationship between the average difficulties of the test scores obtained with different scoring methods was determined by analysis of variance. The reliability of the test shows whether the items in the test are consistent with each other. The reliability of the research data was determined by selecting the appropriate KR-20 or Cronbach Alpha internal consistency method according to the structure of the scoring method. To determine the relationship between the reliability of the test scores obtained with different scoring methods, the reliability coefficients were first converted into Fisher's Z scores as follows (Tan, 2016):

$$Z_r = \frac{1}{2} \ln \frac{1+r_{xy}}{1-r_{xy}}$$

The significance was then calculated in pairs starting from the values with the largest differences between the coefficients obtained from the following equation (Akhun, 1995):

$$Z = \frac{Z_{r1} - Z_{r2}}{\sqrt{\frac{1}{n_1-3} + \frac{1}{n_2-3}}}$$

For the statistical procedures related to the second question of the study, it is necessary to determine the difficulty indices of the test items and the discrimination indices of the items. The item difficulty index is calculated as the ratio of the average of the scores obtained from an item to the item score range. With this calculation, the index to be used in both binary and multiple scoring can be calculated. The item difficulty index takes values between “0” and “1” and as the value approaches “0” the item becomes more difficult, while as the value approaches “1” the item becomes easier. The significance of the difference between the item difficulties obtained using different scoring methods was examined using the analysis of variance method. Item discrimination index is the degree to which the item has the property that it is expected to measure. For this reason, item discrimination is also called item validity. Item discrimination has different formulas according to different scoring types. These are Pearson product-moment correlation, point bi-serial correlation, bi-serial correlation, Phi and Tetrachoric correlations and the technique based on group differences. Although different percentages are taken according to group differences, the most common application is made by selecting 27% groups. While Pearson correlation and upper-lower group method are recommended when calculating the item discrimination indices of partially scored items, some authors (Kilmen, 2012; Turgut & Baykul, 2013) recommend point bi-serial correlation and some (Erkuş, 2016; Özçelik, 2013; Tan, 2016) recommend bi-serial correlation (on the grounds that double scoring is artificially discontinuous). In addition, Erkuş (2016) stated that the upper-lower group method is not useful for items scored as 0-1. Based on all this information, while it was decided to use Pearson correlation to calculate discrimination in multiple scoring, in cases where there is a possibility that the distribution is not normal for dichotomously scored items, it was decided to use the point bi-serial correlation since Tan (2016) stated that it would be more appropriate to prefer the point bi-serial correlation between the point bi-serial and bi-serial correlation coefficients. In order to determine the significance of the difference between the item discrimination indices obtained using different scoring methods, the discrimination scores were first converted into Fisher's Z scores as follows (Tan, 2016):

$$Z_r = \frac{1}{2} \ln \frac{1+r_{xy}}{1-r_{xy}}$$

The significance was then calculated in pairs starting from the values with the largest differences between the coefficients obtained from the following equation (Akhun, 1995):

$$Z = \frac{Z_{r1} - Z_{r2}}{\sqrt{\frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}}}$$

Excel and statistical package programs were used to analyze the data.

Findings

The first two sub-problems of the study, in which the methods used in scoring multiple response item types were compared, analyzed the statistics obtained for the tests in general and the reliability values of the tests. In the third and fourth sub-problems, item statistics were examined and the difficulty and discrimination index values of the items were found and compared.

Findings Related to the Comparison of Test Score Difficulties Across Different Scoring Methods

In the first sub-question in which the impact of the methods used in multiple response item type on test statistics was examined, variance analysis was performed in order to determine whether there was a statistically significant difference among difficulties of test scores obtained through various scoring methods of a certain test generated by using multiple response item type. The difficulties of the tests were shown in Table 2, the results of variance analysis regarding mathematics test in Table 3 and those of Turkish test in Table 4.

Table 2

Difficulty Values of the Test Scores Obtained through Various Scoring Methods

Scoring Methods	Difficulty Values of the Test Scores	
	Mathematics	Turkish
Balanced Scoring (1)	0.51	0.61
Dichotomous Scoring (2)	0.33	0.37
Morgan Algorithm (3)	0.42	0.49
Multiple True/False (4)	0.64	0.73
Negative Scoring (5)	0.45	0.49
Positive Count (6)	0.56	0.68
Polytomous Trapdoor (7)	0.41	0.44
Ripkey Algorithm (8)	0.45	0.51

When looking at the difficulty values of test scores obtained through eight different scoring methods, it has been seen that Mathematics and Turkish tests are, in general, at moderate difficulty level.

Table 3

The Results of Variance Analysis concerning Mathematics Test Scores Obtained through Various Scoring Methods

Source of Variation	Sum of Squares	df	Mean Scores	F	*p*	Difference
Between Groups	6613.35	7	944.76	57.70	.00	(1-2), (1-3), (1-5), (1-7), (1-8), (3-2), (4-1), (4-2), (4-3), (4-5), (4-6), (4-7), (4-8), (5-2), (6-2), (6-3), (6-5), (6-7), (6-8), (7-2), (8-2)
Within Groups	34314.31	2096	16.37			
Total	40927.67	2103				

Note. Balanced Scoring (1), Dichotomous Scoring (2), Morgan Algorithm (3), Multiple True/False (4), Negative Scoring (5), Positive Count (6), Polytomous Trapdoor (7), Ripkey Algorithm (8)

The findings of the research have indicated that there are significant differences among difficulties of tests obtained from various scoring methods. The difficulty obtained through Multiple True/ False scoring method, in particular, demonstrated significant difference from difficulties obtained through other scoring methods.

Table 4

The Results of Variance Analysis concerning Turkish Test Scores Obtained through Various Scoring Methods

Source of Variation	Sum of Squares	df	Mean Squares	F	*p*	Difference
Between Groups	11451.47	7	1635.4	175.56	.00	(1-2), (1-3), (1-5), (1-7), (1-8), (3-2),
Within Groups	20574.75	2208	9.31			(3-7), (4-1), (4-2), (4-3), (4-5), (4-6),
Total	32026.22	2215				(4-7), (4-8), (5-2), (6-1), (6-2), (6-3),
						(6-5), (6-7), (6-8), (7-2), (8-2), (8-7)

Note. Balanced Scoring (1), Dichotomous Scoring (2), Morgan Algorithm (3), Multiple True/False (4), Negative Scoring (5), Positive Count (6), Polytomous Trapdoor (7), Ripkey Algorithm (8)

The results of the analysis regarding Turkish test have revealed that there are significant differences among difficulties of tests obtained through various scoring methods. Furthermore, difficulties gained from Multiple True/ False and Positive Count scoring methods have shown significant differences from those of other scoring methods.

Findings Related to the Comparison of Reliability Coefficients Across Different Scoring Methods

In the second sub- question in which the impact of the methods used in multiple response item type on test statistics was examined, whether there was a significant difference among reliability coefficients of test scores obtained through various scoring methods concerning a test prepared by using multiple response item type was investigated. For this aim, Cronbach Alpha values obtained through seven methods and KR- 20 values obtained through Dichotomous Scoring Method were found and the significance level between values obtained was analysed. The reliability values of test scores obtained through various scoring methods are presented in Table 5 and the results of the analysis in Table 6.

Table 5

The Reliability Coefficients of Test Scores Obtained through Various Scoring Methods

Scoring Methods	Type of Reliability	Reliability Coefficients of Tests	
		Mathematics	Turkish
Balanced Scoring	Cronbach Alpha	0.83	0.76
Dichotomous Scoring	KR-20	0.8	0.7
Morgan Algorithm	Cronbach Alpha	0.81	0.73
Negative Scoring	Cronbach Alpha	0.83	0.78
Multiple True/False	Cronbach Alpha	0.84	0.74
Positive Count	Cronbach Alpha	0.82	0.75
Polytomous Trapdoor	Cronbach Alpha	0.83	0.73
Ripkey Algorithm	Cronbach Alpha	0.81	0.72

As for the reliability coefficients of test scores obtained through eight scoring methods, it has been revealed that reliability coefficients of mathematics test are relatively high. However, in Turkish tests, reliability coefficients have been found to be lower than those of mathematics test; yet, it may still be alleged that they are at the moderate level. Considering the fact that reliability coefficients were affected from the structure of group and the students prepared for the exam, homogeneity and restriction of range were caused. As a result, it may be asserted that the values obtained are high. In Table 6, the comparison of scoring methods having the highest and lowest reliability coefficients for Mathematics and Turkish tests was presented.

Table 6

Z-test Statistics concerning the Highest and Lowest Reliability Coefficients of Test Scores Obtained through Various Scoring Methods

Test	Method	r	z_r	z
Mathematics	MTF	0.84	1.2211	1.3974
	DS	0.80	1.0986	
Turkish	NS	0.78	1.0445	1.6179
	DS	0.70	0.9076	

Note. Dichotomous Scoring (DS), Negative Scoring (NS), Multiple True/ False (MTF)

Fisher's Z- statistics was employed in the comparison of reliability coefficients of Mathematics and Turkish tests and no significant difference was found among test reliabilities.

Findings Related to the Comparison of Item Difficulty Indices Across Different Scoring Methods

In the third sub-question in which the impact of the methods used in the scoring of multiple response item type on item statistics was examined, variance analysis was conducted to reveal whether there was a statistically significant difference among item difficulty indices calculated from the methods used in the scoring of multiple response item type. Tables A-D in the Appendix belong to this sub-question. Based on item difficulty indices of Mathematics test obtained through various scoring methods, items with moderate difficulty have been observed to be included and the item difficulty indices obtained through DS and multiple scoring methods making corrections have been seen to be lower compared to others methods. When looking at the results of variance analysis for item difficulty indices of Mathematics test, it was found that there was a significant difference among certain scoring methods in all items. It may be alleged that the items in Turkish test mostly consist of moderate and easy items according to item difficulty indices obtained from eight scoring methods. In Turkish test, the item difficulty indices obtained through DS and multiple scoring methods making corrections were observed to be lower compared to other methods. As a result of variance analysis conducted for item difficulty indices of Turkish test, a significant difference among certain scoring methods in all items in the test was revealed.

Findings Related to the Comparison of Item Discrimination Indices Across Different Scoring Methods

In the fourth sub- question in which the impact of the methods used in the scoring of multiple response item type on item statistics was examined, in order to determine whether there is a statistically significant difference among item discrimination indices calculated from the methods used in multiple response item type, discrimination indices were firstly transformed into Fisher's Z- statistics and, then, their significance levels were investigated. Tables E-H in the Appendix belong to this sub-question. It may be inferred that almost all items in the Mathematics test have fairly sufficient discrimination levels according to item discrimination indices results calculated through eight scoring methods. The results of anaysis in Mathematics test indicated significant differences among item discrimination indices obtained through various scoring methods in 5- 7- 11 and 16^{th} items. It can be said that the items in Turkish test, in general, except two of them, have rather sufficient discrimination levels based on item discrimination indices calculated through various scoring methods. The results of analysis in Turkish test showed significant differences among item discrimination indices obtained through 3- 7- 14- 15- 17- 18 and 20^{th} items.

Discussion

The present study aimed to compare various scoring methods used for multiple response items, a format frequently utilized in achievement tests and classroom assessments, while also highlighting its advantages over traditional multiple-choice items. In this section, the results are discussed in light of the study's objectives and relevant literature.

In addressing the first sub-question of the study, the discussion focuses on the differences in test score difficulties arising from the use of various scoring methods for multiple response items. The findings revealed that the difficulties of test scores obtained through the Multiple True/False and Positive Count methods were significantly higher compared to those obtained through other scoring approaches. These results suggest that such methods can be effectively incorporated into classroom practices to monitor students' achievement of course objectives and to identify areas of learning deficiencies. On the other hand, the methods of Polytomous Trapdoor,

Dichotomous Scoring, and the Morgan Algorithm yielded notably lower difficulty levels, indicating their potential suitability for large-scale assessments where election-based or maximum score-based evaluations are preferred. These findings are consistent with the results reported by Muckle et al. (2011) and align with the conclusions drawn by Domnich et al. (2015) and Ripkey et al. (1996).

In examining the second sub-question of the study, the discussion focuses on the comparison of reliability coefficients obtained through different scoring methods for multiple response items. The fact that there is no significant difference among test reliability coefficients obtained from the methods used in the scoring of multiple response items indicates that it is not possible to classify the scoring methods of this items type as good or bad. Besides, it has been observed that the reliability coefficients obtained through multiple scoring method are higher compared to those obtained through Dichotomous Scoring Method. Hsu et al. (1984) compared the scoring methods of multiple response items and multiple-choice items by using History, Chinese and San Min Chu- I (Political Theory course by Dr. Sun Yat Sen) sub- tests included in the Joint College Entrance Examination. The results showed no significant difference among the reliabilities of the scoring methods used in History and San Min Chu- I courses. On the contrary, in Chinese course, the reliability values of Dichotomous Scoring and the method named as S1 were found to be significantly lower compared to other methods.

In addressing the third sub-question, the discussion centers on the differences in item difficulty indices calculated through various scoring methods applied to multiple response items. When looking at the item difficulty indices obtained through eight scoring methods used in the scoring of multiple response item type, it was revealed that the lowest values in both Turkish and Mathematics tests were obtained in Dichotomous Scoring, Morgan Algorithm applying penalty, Ripkey Algorithm and Polytomous Trapdoor respectively. However, Positive Count scoring method in which the practice of correction was not conducted and Multiple True/ False scoring method in which any incorrect option left unmarked was given one point are those having the highest level of item difficulty. As in mean difficulties of tests, the two items having higher level of item difficulty in all test items may be used in such tests which are to carry out formative assessment, to provide efficient feedback and to monitor learning. Nevertheless, Morgan and Ripkey Algorithms as well as Polytomous Trapdoor scoring method in which the practices of corrections are performed are recommended to be used in summative assessments the goal of which is mostly the evaluation of student learning and assessment.

In discussing the fourth sub-question, the focus is placed on the differences in item discrimination indices calculated using various scoring methods for multiple response items. Based on the examination on item discrimination indices obtained through eight standardized scoring methods, significant difference in terms of item discrimination indices of four items in Mathematics test and seven items in Turkish test. As for the investigation on four items in Mathematics test, Positive Count method was found to be significantly different whereas, in five out of seven items in Turkish test, Multiple True/ False scoring method was found to be significantly different. In addition, Dichotomous Scoring Method, in particular, was observed to be significantly low in Turkish test. In context of the highest item discrimination indices based on item by item among scoring methods, Multiple True/ False and Positive Count methods were of the greatest number of items with higher level of item discrimination indices. Yet, in Mathematics test, the methods having the greatest number of items with higher level of item discrimination indices were Polytomous Trapdoor, Multiple True/ False and Dichotomous Scoring. Verbic (2012) found that cluster scoring was more robust than item scoring and that multiple scoring methods had better results in terms of discrimination than binary scoring. Hsu et al. (1984) found that multiple scoring methods were more discriminative and reliable than binary scoring methods. When the items in the Turkish and mathematics tests are classified as easy, medium difficulty and difficult with the item difficulty index results obtained according to the scoring methods and compared with the discrimination of the items, a consistent result emerges. According to this inference, it was seen that the discrimination indices obtained with the Binary Scoring Method for easy items were slightly higher than those obtained with other scoring methods. It was noteworthy that the discrimination indices obtained with the Multiple True/False Scoring Method for medium difficulty items were slightly higher than those obtained with other scoring methods. For items with low item difficulty indices and labeled as difficult, the discrimination indices obtained with the Trap Multiple Response Scoring Method were slightly higher than those obtained with other scoring methods. Muckle et al. (2011) articulated that 'Trapdoor' scoring method were able to discriminate among the examinees with various levels of skills to the greatest extent. Domnich et al. (2015) concluded that there was no significant difference among discrimination levels of scoring methods of multiple response items.

Conclusion

According to the data gained from the tests consisting of multiple response item types, the fact that this item type is appropriate to use for quantitative and verbal courses may be inferred from the reliability values of test and item statistics. Turgut and Baykul (2013), who were opposed to the applicability and scoring methods of this item type, stated that it was easy to write multiple response items; yet, added that there would be important difficulties during item analysis and recommended not using this item type if not necessary. Karakaya and Doğan (2020) interviewed with the staff in the Measurement and Evaluation Centers, graduate students at the department of Measurement and Evaluation in Education and academicians from the same department. The study unearthed the participants' confusion regarding the scoring methods of multiple response items and, therefore, they were monitored not to be in favour of those methods. The participants' statements regarding the negative impact of the variety of scoring methods on reliability is worth highlighting. Contrary to this, the scoring methods of multiple response item type can be standardized, and item analysis be conducted based on the current research. Furthermore, it was found that standardized scoring methods could allow various scoring methods, without conforming to a single rule, to be used by introducing alternatives depending on the objectives of measurement and evaluation. Thus, Multiple True/ False and Positive Count scoring methods are recommended to determine learning deficiencies in in- class assessments and to be used in measurement and evaluation practices aiming at effective feedback. However, Polytomous Tradoor, Morgan Algorithm or Dichotomous Scoring was found to be used in large- scale exams aiming to select and distinguish among individuals. The fact that there was no significant difference among reliability values of scoring methods and those methods are sufficiently reliable would eliminate the uncertainties concerning the reliability of this item type. It has been thought that item type would become prevalent through this study and other studies alike regarding multiple response item type and scoring methods. In addition, the use of item type is considered to be extended in computer- based assessment and evaluation practices as in paper-pencil exams and large- scale examinations conducted abroad.

Suggestions

1. The selection of appropriate scoring methods based on assessment and evaluation objectives of tests constituted through multiple response items is considered to make contributions to test- makers. Therefore, test and item's statistics of the methods used to score multiple response items via the tests with multiple response items except Turkish and Mathematics courses may be compared.
2. The implementations in which the number of correct alternatives has not been known in such tests generated by using multiple response items have been realized. The impact of the situations when the number of correct alternatives is reported in test instructions on the results of the tests may be studied.
3. The measurement instruments consisting of multiple response items have been administered in educational institutions. With the help of these measurement tools, the validity and reliability of measurements made in different application areas such as medicine and engineering can be examined.
4. Multiple True/ False and Positive Count scoring methods are recommended in assessment and evaluation practices aiming to identify learning deficiencies and provide effective feedback. However, Polytomous Trapdoor, Morgan Algorithm or Dichotomous Scoring methods may be suggested to be used in large-scale examinations conducted to select and distinguish among individuals.

References

Akhun, İ. (1995). *İstatistiklerin manidarlığı ve örneklem [Significance of statistics and sampling]*. Özel Basım.

Auer, S. & Tarasowa, D. (2013). Balanced scoring method for multiple-mark questions. *Proceedings of the 5th International Conference on Computer Supported Education (CSEDU-2013),* 411–416. https://doi.org/10.5220/0004384304110416

Baghaei, P. & Amrahi, N. (2011). The effects of the number of options on the psychometric characteristics of multiple choice items. *Psychological Test and Assessment Modeling, 53*(2), 192–211.

Baykul, Y. (2015). *Eğitimde ve psikolojide ölçme: Klasik test teorisi ve uygulaması.[Measurement in education and psychology: Classical test theory and practice]*. Pegem Akademi.

Ben-Simon, A., Budescu, D. V. & Nevo, B. (1997). A comparative study of measures of partial knowledge in multiple-choice tests. *Applied Psychological Measurement, 21*(1), 65–88.

Doğan, N. (2009). Çoktan seçmeli testler [Multiple choice items]. In H. Atılgan (Ed.), *Eğitimde ölçme ve değerlendirme* (pp. 223–268). Anı.

Doğanay, A. (2007). Üst düzey düşünme becerilerinin öğretimi [Teaching higher order thinking skills]. In A. Doğanay (Ed.), *Öğretim ilke ve yöntemleri* (pp. 280–327). Pegem Akademi.

Domnich, A., Panatto, D., Arata, L. Bevilacqua, I., Apprato, L., Gasparini, R. & Amicizia, D. (2015). Impact of different scoring algorithms applied to multiple-mark survey items on outcome assessment: An in-field study on health-related knowledge. *Journal of Preventive Medicine and Hygiene*, *56*(4), E162.

Duman, E. Z. (2018). Bir düşünme türü olarak esnek düşünme [Flexible thinking as a type of thinking]. *Turkish Studies Social Sciences*, *13*(26), 547–561.

Erkuş, A. (2016). *Psikolojide ölçme ve ölçek geliştirme-I [Measurement and scale development in psychology-I]*. Pegem Akademi.

Frisbie, D. A. & Sweeney, D. C. (1982). The relative merits of multiple true-false achievement tests. *Journal of Educational Measurement*, *19*(1), 29–35.

Hsu, T.C., Moss, P.A., Khampalikit, C. (1984). The merits of multiple-answer items as evaluated by using six scoring formulas. *The Journal of Experimental Education*, *52*(3), 152–158.

Jodoin, M. G. (2003). Measurement efficiency of innovative item formats in computer-based testing. *Journal of Measurement*, *40*(1), 1–15.

Karakaya, İ. & Doğan, Ö. (2020). Çoklu cevaplı madde türünün psikometrik özellikleri hakkında ödm çalışanı, alan uzmanı ve akademisyenlerin görüşlerinin incelenmesi [Examination of the opinions of the staff of the assessment and evaluation center, field specialists and academicians about the psychometric properties of the multiple response item type] *Ankara II. Uluslararası Bilimsel Araştırmalar Kongresi Bildirileri, 1,* 271–279.

Karasar, N. (2012). *Bilimsel irade algı çerçevesi ile bilimsel araştırma yöntemi [Scientific research method with the framework of perception of scientific will]*. Nobel Yayınları.

Kilmen, S. (2012). Madde analizi, madde seçimi ve yorumlanması [Item analysis, item selection, and interpretation]. In N. Çıkrıkçı-Demirtaşlı (Ed.), *Eğitimde ölçme ve değerlendirme*, (pp. 363–385). Elhan.

Koç, N. (2009), *Eğitimimizde psikolojik testlerin (standart testlerin) kullanım durumu, sorunlar ve öneriler*. *[Use of psychological tests (standard tests) in our education, problems and suggestions]* I. Ulusal Eğitimde ve Psikolojide Ölçme ve Değerlendirme Kongresi bildirileri 1, 305–316.

Liu, O. L., Lee, H. S., & Linn, M. C. (2011). An Investigation of Explanation Multiple-Choice Items in Science Assessment. *Educational Assessment*, *16*(3), 164–184. https://doi.org/10.1080/10627197.2011.611702

Ma, X. (2004). *An investigation of alternative approaches to scoring multiple response items on a certification exam*. (Publication No. 305176174) [Doctoral dissertation, University of Massachusetts Amherst]. ProQuest Dissertations Publishing.

Morgan, M. (1979). MCQ: An interactive computer program for multiple-choice self-testing. *Biochemical Education*, *7*(3), 67–69.

Muckle, T., Becker, K. A. & Wu, B. (2011). Investigating the multiple answer multiple choice item format. Presentation and scoring consideration. *2011 Annual Meeting of the National Council on Measurement in Education*, New Orleans, LA.

Özçelik, D. A. (2013). *Test hazırlama kılavuzu [Test preparation guide]*. Pegem Akademi.

Peterson, A., Craig M. & Danny, P. (2016). Employing multiple answer multiple choice questions. *Proceedings of the 2016 ACM Conference on Innovation and Technology in Computer Science Education*, 252–253. https://doi.org/10.1145/2899415.2925503

Ripkey, D., Case, S. & Swanson, D. (1996). A "new" item format for assessing aspects of clinical competence. *Academic Medicine*, *71*(10), 34–36.

Siddiqui, N. I., Bhavsar, V. H., Bhavsar, A. V., & Bose, S. (2016). Contemplation on marking scheme for Type X multiple choice questions, and an illustration of a practically applicable scheme. *Indian journal of pharmacology*, *48*(2), 114–121.

Tan, Ş. (2010). *Öğretimde ölçme ve değerlendirme [Measurement and evaluation in teaching]*. Pegem Akademi.

Tan, Ş. (2016). *SPSS ve Excel uygulamalı temel istatistik-1 [Basic statistics-1 with SPSS and Excel]*. Pegem Akademi.

Turgut, M. F. & Baykul, Y. (2013). *Eğitimde ölçme ve değerlendirme [Measurement and evaluation in education]*. Pegem Akademi.

Umay, A. (1997). Yanıtlayıcı davranışlarının analizi yolu ile matematikte problem çözümleri için bir güvenirlik ve geçerlik araştırması [A reliability and validity study for problem solving in mathematics through the analysis of respondent behavior]. *Hacettepe University Journal of Education, 13*, 47–56.

Verbic, S. (2012). Information value of multiple response questions. *Psihologija*, *45*(4), 467–485.

Wan, L. & Henly, G. A. (2012). Measurement properties of two innovative item formats in a computer-based test. *Applied Measurement in Education, 25*, 58–78.

Çoklu Cevaplı Madde Türünün Puanlanmasında Yöntemlerin Karşılaştırılması

Öz

Bu çalışmanın amacı, bir sorunun birden fazla doğru cevabının olabildiği çoklu cevaplı madde türünün puanlama yöntemlerini incelemektir. Bu doğrultuda, çoklu cevaplı madde türlerinden oluşan başarı testlerinin puanlanmasında kullanılan sekiz puanlama yönteminin madde güçlüğü, güvenirlik ve ayırt edicilik açısından farklılıkları incelenmiş ve karşılaştırılmıştır. Elde edilen verilere göre, çoklu cevaplı madde türünün hem sayısal hem de sözel dersler için uygulanabilir olduğu anlaşılmıştır. Ayrıca standardize edilmiş puanlama yöntemlerinin tek bir kurala bağlı kalmadan, ölçme ve değerlendirmenin amacına göre alternatifler sunarak farklı puanlama yöntemlerinin kullanılmasına olanak sağlayacak şekilde olduğu görülmüştür. Böylece çoklu cevaplı madde türünün kullanım amacına göre puanlama yöntemleri belirlenebilir ve benzer çalışmalarla madde türünün kullanımı yaygınlaştırılabilir.

Anahtar kelimeler: çoklu cevaplı madde türü, kısmi puanlama, puanlama yöntemleri

Appendix

Table A

Item Difficulty Indices of Mathematics Test Calculated from Different Scoring Methods

Item Number	Difficulty Indices of Mathematics Test Items According to Scoring Methods							
	BaS	DS	MA	NP	MTF	PS	TMR	RA
1	0.83	0.66	0.82	0.8	0.92	0.86	0.77	0.78
2	0.57	0.39	0.45	0.51	0.72	0.63	0.43	0.46
3	0.53	0.37	0.41	0.46	0.66	0.54	0.45	0.51
4	0.33	0.18	0.25	0.18	0.58	0.51	0.18	0.18
5	0.48	0.23	0.34	0.34	0.62	0.56	0.28	0.32
6	0.49	0.35	0.4	0.46	0.62	0.5	0.45	0.48
7	0.52	0.26	0.42	0.42	0.66	0.56	0.39	0.47
8	0.64	0.48	0.55	0.6	0.75	0.7	0.53	0.54
9	0.68	0.57	0.59	0.66	0.71	0.68	0.65	0.68
10	0.31	0.18	0.19	0.27	0.44	0.32	0.26	0.3
11	0.59	0.39	0.51	0.54	0.71	0.62	0.52	0.56
12	0.55	0.3	0.47	0.48	0.69	0.56	0.47	0.53
13	0.71	0.43	0.65	0.61	0.81	0.79	0.5	0.56
14	0.53	0.34	0.45	0.5	0.61	0.54	0.47	0.54
15	0.29	0.27	0.23	0.27	0.45	0.34	0.27	0.27
16	0.51	0.29	0.45	0.43	0.62	0.56	0.38	0.43
17	0.34	0.11	0.18	0.26	0.42	0.34	0.24	0.34
18	0.49	0.33	0.4	0.43	0.59	0.52	0.41	0.46
19	0.47	0.25	0.33	0.38	0.61	0.52	0.33	0.4
20	0.41	0.29	0.34	0.29	0.59	0.48	0.29	0.29

Balanced Scoring (BaS), Dichotomous Scoring (DS), Morgan Algorithm (MA) Negative Scoring (NP), Multiple True/False Scoring (MTF), Positive Counting (PS), Trap Multiple Response (TMR), Ripkey Algorithm (RA)

Table B

Analysis of Variance Results for Item Difficulty Indices of Mathematics Test Calculated with Different Scoring Methods

Item Number	Variance Source for Item	Sum of Square	df	Mean Square	F	*p*	Differences
1	Between group	10.18	7	1.45	14.95	.00	(1-2), (3-2), (4-2), (4-3), (4-5), (4-6), (4-7), (4-8), (5-2), (6-2), (6-7), (7-2), (8-2)
	Within group	203.84	2096	.09			
	Total	214.02	2103				
2	Between group	22.98	7	3.28	15.81	.00	(1-2), (1-7), (4-1), (4-2), (4-3), (4-5), (4-7), (4-8), (6-2), (6-3), (6-7), (6-8)
	Within group	435.29	2096	.20			
	Total	458.28	2103				
3	Between group	15.04	7	2,150	10.90	.00	(1-2), (1-3), (4-1), (4-2), (4-3), (4-5), (4-7), (4-8), (6-2), (6-3), (8-2)
	Within group	413.36	2096	.19			
	Total	428.41	2103				
4	Between group	48.77	7	6.96	39.04	.00	(1-2), (4-1), (4-2), (4-3), (4-5), (4-7), (4-8), (6-1), (6-2), (6-3), (6-5), (6-7), (6-8)
	Within group	374.05	2096	.17			
	Total	422.82	2103				
5	Between group	35.81	7	5.11	27.96	.00	(1-2), (1-3), (1-7), (1-8), (4-1), (4-2), (4-3), (4-5), (4-7), (4-8), (5-2), (6-2), (6-3), (6-5), (6-7), (6-8)
	Within group	383.52	2096	.18			
	Total	419.33	2103				
6	Between group	11.84	7	1.69	8.54	.00	(1-2), (4-1), (4-2), (4-3), (4-5), (4-6), (4-7), (4-8), (6-2), (8-2)
	Within group	414.89	2096	.18			
	Total	426.74	2103				
7	Between group	26.29	7	3.75	24.03	.00	(1-2), (1-3), (1-7), (4-1), (4-2), (4-3), (4-5), (4-7), (4-8), (5-2), (6-2), (6-3), (6-5), (7-2), (8-2)
	Within group	327.61	2096	.15			
	Total	353.90	2103				
8	Between group	15.32	7	2.19	10.86	.00	(1-2), (4-2), (4-3), (4-5), (4-7), (4-8), (6-2), (6-3), (6-7), (6-8)
	Within group	422.32	2096	.20			
	Total	437.60	2103				
9	Between group	4.23	7	.60	2.98	.00	(4-2), (4-3)
	Within group	424.92	2096	.20			
	Total	429.15	2103				
10	Between group	12.28	7	1.75	11.03	.00	(1-2), (1-3), (4-1), (4-2), (4-3), (4-5), (4-6), (4-7), (4-8), (6-2), (6-3)
	Within group	333.26	2096	.15			
	Total	345.54	2103				

Balanced Scoring (1), Dichotomous Scoring (2), Morgan Algorithm (3) Negative Scoring (4), Multiple True/False Scoring (5), Positive Counting (6), Trap Multiple Response (T7), Ripkey Algorithm (8)

Table B (continues)

Item Number	Variance Source for Item	Sum of Square	df	Mean Square	F	p	Differences
11	Between group	15.94	7	2.27	13.07	.00	(1-2), (3-2), (4-1), (4-2), (4-3), (4-5), (4-7), (4-8), (5-2), (6-2), (7-2), (8-2)
	Within group	365.02	2096	.17			
	Total	380.96	2103				
12	Between group	21.37	7	3.05	19.73	.00	(1-2), (3-2), (4-1), (4-2), (4-3), (4-5), (4-6), (4-7), (4-8), (5-2), (6-2), (7-2), (8-2)
	Within group	324.32	2096	.15			
	Total	345.70	2103				
13	Between group	33.46	7	4.78	32.30	.00	(1-2), (1-7), (1-8), (3-2), (4-2), (4-3), (4-5), (4-7), (4-8), (5-2), (5-7), (6-2), (6-3), (6-5), (6-7), (6-8)
	Within group	310.13	2096	.14			
	Total	343.59	2103				
14	Between group	12.31	7	1.76	9.95	.00	(1-2), (4-2), (4-3), (4-5), (4-7), (5-2), (6-2), (7-2), (8-2)
	Within group	370.54	2096	.17			
	Total	382.86	2103				
15	Between group	8.75	7	1.25	5.90	.00	(4-1), (4-2), (4-3), (4-5), (4-7), (4-8)
	Within group	444.29	2096	.21			
	Total	453.05	2103				
16	Between group	19.89	7	2.84	16.18	.00	(1-2), (1-7), (3-2), (4-2), (4-3), (4-5), (4-7), (4-8), (5-2), (6-2), (6-5), (6-7), (6-8)
	Within group	368.03	2096	.17			
	Total	387.92	2103				
17	Between group	19.00	7	2.71	22.26	.00	(1-2), (1-3), (1,7), (4-2), (4-3), (4-5), (4-7), (5-2), (6-2), (6-3), (6-7), (7-2), (8-2), (8-3), (8-7)
	Within group	255.56	2096	.12			
	Total	274.56	2103				
18	Between group	12.07	7	1.72	8.77	.00	(1-2), (4-2), (4-3), (4-5), (4-7), (4-8), (6-2), (6-3), (8-2)
	Within group	412.03	2096	.19			
	Total	424.10	2103				
19	Between group	25,.7	7	3.59	20.53	.00	(1-2), (1-3), (1-7), (4-1), (4-2), (4-3), (4-5), (4-7), (4-8), (5-2), (6-2), (6-3), (6-5), (6-7), (6-8), (8-2)
	Within group	367.09	2096	.17			
	Total	392.27	2103				
20	Between group	23.43	7	3.34	15.39	.00	(4-1), (4-2), (4-3), (4-5), (4-7), (4-8), (6-2), (6-3), (6-5), (6-7), (6-8)
	Within group	455.76	2096	.21			
	Total	479.20	2103				

Balanced Scoring (1), Dichotomous Scoring (2), Morgan Algorithm (3) Negative Scoring (4), Multiple True/False Scoring (5), Positive Counting (6), Trap Multiple Response (T7), Ripkey Algorithm (8)

Table C

Item Difficulty Indices of Turkish Test Calculated from Different Scoring Methods

Item Number	Difficulty Indices of Turkish Test Items According to Scoring Methods							
	BaS	DS	MA	NP	MTF	PS	TMR	RA
1	0.61	0.25	0.44	0.39	0.73	0.66	0.33	0.52
2	0.53	0.31	0.48	0.31	0.73	0.69	0.31	0.31
3	0.47	0.16	0.28	0.31	0.63	0.49	0.28	0.42
4	0.71	0.56	0.64	0.65	0.80	0.73	0.64	0.68
5	0.58	0.51	0.37	0.56	0.66	0.62	0.53	0.54
6	0.56	0.39	0.49	0.39	0.75	0.69	0.39	0.39
7	0.58	0.27	0.47	0.46	0.72	0.73	0.30	0.35
8	0.62	0.32	0.52	0.52	0.77	0.66	0.48	0.56
9	0.36	0.38	0.60	0.45	0.53	0.49	0.44	0.45
10	0.85	0.72	0.81	0.81	0.89	0.89	0.76	0.78
11	0.65	0.35	0.56	0.53	0.78	0.70	0.46	0.54
12	0.78	0.68	0.72	0.75	0.84	0.81	0.71	0.73
13	0.47	0.16	0.29	0.31	0.64	0.51	0.27	0.40
14	0.74	0.44	0.69	0.64	0.82	0.86	0.47	0.51
15	0.65	0.39	0.57	0.55	0.75	0.74	0.44	0.49
16	0.74	0.43	0.72	0.43	0.81	0.87	0.43	0.43
17	0.56	0.19	0.28	0.40	0.60	0.57	0.32	0.56
18	0.48	0.27	0.41	0.29	0.69	0.64	0.27	0.27
19	0.53	0.16	0.34	0.42	0.61	0.52	0.37	0.52
20	0.77	0.51	0.66	0.70	0.80	0.77	0.66	0.77

Note. Balanced Scoring (BaS), Dichotomous Scoring (DS), Morgan Algorithm (MA) Negative Scoring (NP), Multiple True/False Scoring (MTF), Positive Counting (PS), Trap Multiple Response (TMR), Ripkey Algorithm (RA)

Table D

Analysis of Variance Results for Item Difficulty Indices of Turkish Test Calculated with Different Scoring Methods

Item Number	Variance Source for Item	Sum of Square	*df*	Mean Square	*F*	*p*	Differences
1	Between group	54.03	7	7.71	59.54	.00	(1-2), (1-3), (1-5), (1-7), (3-2), (3-7), (4-1), (4-2), (4-3), (4-5), (4-7), (4-8), (5-2), (6-2), (6-3), (6-5), (6-7), (6-8), (8-2)
	Within group	286.22	2208	.13			
	Total	340.257	2215				
2	Between group	61.98	7	8.85	41.60	.00	(1-2), (1-5), (1-7), (1-8), (3-2), (3-5), (3-7), (3-8), (4-1), (4-2), (4-3), (4-5), (4-7), (4-8), (6-1), (6-2), (6-3), (6-5), (6-7), (6-8)
	Within group	469.97	2208	.21			
	Total	531.95	2215				
3	Between group	43.58	7	6.22	46.68	.00	(1-2), (1-3), (1-5), (1-7), (3-2), (4-1), (4-2), (4-3), (4-5), (4-7), (4-8), (5-2), (6-2), (6-3), (6-5), (6-7), (7-2), (8-2)
	Within group	294.47	2208	.13			
	Total	338.05	2215				
4	Between group	9.60	7	1.37	8.03	.00	(1-2), (4-2), (4-3), (4-5), (4-7), (4-8), (6-2)
	Within group	377.28	2208	.17			
	Total	386.89	2215				
5	Between group	14.47	7	2.06	7.90	.00	(1-3), (2-3), (4-2), (4-3), (5-3), (6-3), (7-3), (8-3)
	Within group	577.40	2208	.26			
	Total	591.87	2215				
6	Between group	42.74	7	6.10	26.59	.00	(1-2), (1-5), (1-7), (1-8), (3-2), (3-7), (4-1), (4-2), (4-3), (4-5), (4-7), (4-8), (6-1), (6-2), (6-3), (6-5), (6-7), (6-8)
	Within group	507.03	2208	.23			
	Total	549.78	2215				
7	Between group	63.46	7	9.06	57.43	.00	(1-2), (1-3), (1-5), (1-7), (1-8), (3-2), (3-7), (3-8), (4-1), (4-2), (4-3), (4-5), (4-7), (4-8), (5-2), (5-7), (5-8), (6-1), (6-2), (6-3), (6-5), (6-7), (6-8)
	Within group	348.54	2208	.15			
	Total	412.01	2215				
8	Between group	34.31	7	4.90	35.03	.00	(1-2), (1-3), (1-5), (1-7), (3-2), (4-1), (4-2), (4-3), (4-5), (4-6), (4-7), (4-8), (5-2), (6-2), (6-3), (6-5), (6-7), (6-8), (8-2)
	Within group	308.93	2208	.14			
	Total	343.24	2215				
9	Between group	40.96	7	5.85	19.24	.00	(1-3), (2-3), (4-1), (4-2), (4-3), (5-3), (6-3), (7-3), (8-3)
	Within group	671.41	2208	.30			
	Total	712.38	2215				
10	Between group	7.53	7	1.07	8.34	.00	(1-2), (3-2), (4-2), (4-7), (4-8), (5-2), (6-2), (6-7), (6-8)
	Within group	284.57	2208	.12			
	Total	292.11	2215				

Note. Balanced Scoring (1), Dichotomous Scoring (2), Morgan Algorithm (3) Negative Scoring (4), Multiple True/False Scoring (5), Positive Counting (6), Trap Multiple Response (T7), Ripkey Algorithm (8)

Table D (continues)

Item Number	Variance Source for Item	Sum of Square	*df*	Mean Square	*F*	*p*	Differences
11	Between group	35.72	7	5.10	34.52	.00	(1-2), (1-5), (1-7), (1-8), (3-2), (4-1), (4-2), (4-3), (4-5), (4-7), (4-8), (5-2), (6-2), (6-3), (6-5), (6-7), (6-8), (7-2), (8-2)
	Within group	326.44	2208	.14			
	Total	362.17	2215				
12	Between group	5.60	7	.80	4.89	.00	(4-2), (4-3), (4-7), (4-8), (6-2)
	Within group	361.28	2208	.16			
	Total	366.88	2215				
13	Between group	45.41	7	6.48	49.41	.00	(1-2), (1-3), (1-5), (1-7), (3-2), (4-1), (4-2), (4-3), (4-5), (4-6), (4-7), (4-8), (5-2), (6-2), (6-3), (6-5), (6-7), (6-8), (7-2), (8-2), (8-3), (8-7)
	Within group	289.88	2208	.13			
	Total	335.29	2215				
14	Between group	48.38	7	6.91	45.21	.00	(1-2), (1-5), (1-7), (1-8), (3-2), (3-7), (3-8), (4-2), (4-3), (4-5), (4-7), (4-8), (5-2), (5-7), (5-8), (6-1), (6-2), (6-3), (6-5), (6-7), (6-8)
	Within group	337.49	2208	.15			
	Total	385.87	2215				
15	Between group	34.45	7	4.921	28.17	.00	(1-2), (1-7), (1-8), (3-2), (3-7), (4-2), (4-3), (4-5), (4-7), (4-8), (5-2), (6-2), (6-3), (6-5), (6-7), (6-8)
	Within group	385.65	2208	.175			
	Total	420.10	2215				
16	Between group	72.04	7	10.29	56.71	.00	(1-2), (1-5), (1-7), (1-8), (3-2), (3-5), (3-7), (3-8), (4-2), (4-5), (4-7), (4-8), (6-1), (6-2), (6-3), (6-5), (6-7), (6-8)
	Within group	400.61	2208	.18			
	Total	472.65	2215				
17	Between group	4866	7	6.95	52.32	.00	(1-2), (1-3), (1-5), (1-7), (4-2), (4-3), (4-5), (4-7), (5-2), (5-3), (6-2), (6-3), (6-5), (6-7), (7-2), (8-2), (8-3), (8-5) (8-7)
	Within group	293.36	2208	.13			
	Total	342.02	2215				
18	Between group	57.01	7	8.14	38.61	.00	(1-2), (1-5), (1-7), (1-8), (3-2), (3-5), (3-7), (3-8), (4-1), (4-2), (4-3), (4-5), (4-7), (4-8), (6-1), (6-2), (6-3), (6-5), (6-7), (6-8)
	Within group	465.72	2208	.21			
	Total	522.73	2215				
19	Between group	39.55	7	5.65	47.85	.00	(1-2), (1-3), (1-5), (1-7), (3-2), (4-2), (4-3), (4-5), (4-7), (5-2), (6-2), (6-3), (6-5), (6-7), (7-2), (8-2), (8-3), (8-5)
	Within group	260.71	2208	.11			
	Total	300.26	2215				
20	Between group	17.74	7	2.53	19.23	.00	(1-2), (1-3), (1-7), (3-2), (4-2), (4-3), (4-5), (4-7), (5-2), (6-2), (6-3), (6-7), (7-2), (8-2), (8-3), (8-7)
	Within group	291.04	2208	.13			
	Total	308.79	2215				

Table E

Item Discrimination Indices of Mathematics Test Calculated from Different Scoring Methods

Item Number	Discrimination Indices of Mathematics Test Items According to Scoring Methods							
	BaS	DS	MA	NP	MTF	PS	TMR	RA
1	0.3	0.36	0.31	0.33	0.26	0.26	0.34	0.31
2	0.52	0.55	0.51	0.53	0.48	0.47	0.54	0.53
3	0.47	0.55	0.45	0.5	0.42	0.48	0.49	0.46
4	0.41	0.36	0.39	0.36	0.44	0.46	0.38	0.4
5	0.48	0.37	0.45	0.43	0.46	0.49	0.33	0.33
6	0.4	0.38	0.38	0.42	0.43	0.39	0.43	0.4
7	0.38	0.3	0.4	0.41	0.43	0.37	0.4	0.36
8	0.43	0.46	0.44	0.46	0.41	0.41	0.45	0.44
9	0.54	0.5	0.49	0.54	0.53	0.55	0.5	0.49
10	0.52	0.49	0.47	0.53	0.59	0.51	0.54	0.49
11	0.43	0.5	0.46	0.46	0.36	0.39	0.47	0.45
12	0.54	0.5	0.58	0.59	0.56	0.53	0.59	0.55
13	0.39	0.41	0.38	0.41	0.35	0.43	0.38	0.32
14	0.47	0.44	0.41	0.45	0.48	0.46	0.43	0.47
15	0.59	0.6	0.55	0.61	0.59	0.57	0.6	0.6
16	0.55	0.42	0.53	0.55	0.61	0.56	0.5	0.49
17	0.56	0.61	0.52	0.59	0.54	0.57	0.61	0.54
18	0.59	0.59	0.59	0.6	0.59	0.59	0.6	0.58
19	0.65	0.63	0.64	0.65	0.6	0.63	0.63	0.6
20	0.54	0.53	0.5	0.57	0.58	0.51	0.58	0.57

Note. Balanced Scoring (BaS), Dichotomous Scoring (DS), Morgan Algorithm (MA) Negative Scoring (NP), Multiple True/False Scoring (MTF), Positive Counting (PS), Trap Multiple Response (TMR), Ripkey Algorithm (RA)

Table F

Results of Comparison of Item Discrimination Indices of Mathematics Test Calculated from Different Scoring Methods

Item Number	Method	r_{jx}	z_r	z
1	DS	0.36	0.376	1,2635
	MTF	0.26	0.266	
2	DS	0.55	0.618	1,2349
	PS	0.47	0.510	
3	DS	0.55	0.618	1,9461
	MTF	0.42	0.447	
4	PS	0.46	0.497	1,3730
	NP	0.36	0.376	
5	PS	0.49	0.536	2,2131*
	TMR	0.33	0.342	
	PS	0.49	0.536	2,2131*
	RA	0.33	0.342	
	BaS	0.48	0.522	2,0540*
	TMR	0.33	0.342	
	BaS	0.48	0.522	2,0540*
	RA	0.33	0.342	
6	MTF	0.43	0.459	0,6822
	MA	0.38	0.400	
7	RA	0.46	0.497	2,1411*
	BiS	0.30	0.309	
8	NP	0.46	0.497	0,7034
	PS	0.41	0.435	
9	PS	0.55	0.618	0,9386
	MA	0.49	0.536	
10	MTF	0.59	0.677	1,9108
	MA	0.47	0.510	
11	DS	0.50	0.549	1,9658*
	MTF	0.36	0.376	
12	TMR	0.59	0.677	1,4635
	DS	0.50	0.549	
13	PS	0.43	0.459	1,4622
	RA	0.32	0.331	
14	MTF	0.48	0.522	0,9962
	MA	0.41	0.435	
15	NP	0.61	0.708	1,0323
	MA	0.55	0.618	
16	MTF	0.61	0.708	2,9784**
	DS	0.42	0.447	
	PS	0.56	0.632	2,1109*
	DS	0.42	0.447	
	MTF	0.61	0.708	1,9709*
	RA	0.49	0.536	
17	TMR	0.61	0.708	1,5116
	MA	0.52	0.576	
18	NP	0.60	0.693	0,3498
	RA	0.58	0.662	
19	BaS	0.65	0.775	0,9366
	MTF	0.60	0.693	
20	TMR	0.58	0.662	1,2901
	MA	0.50	0.549	

Note. Balanced Scoring (BaS), Dichotomous Scoring (DS), Morgan Algorithm (MA) Negative Scoring (NP), Multiple True/False Scoring (MTF), Positive Counting (PS), Trap Multiple Response (TMR), Ripkey Algorithm (RA)

Table G

Item Discrimination Indices of Turkish Test Calculated from Different Scoring Methods

Item Number	Discrimination Indices of Turkish Test Items According to Scoring Methods							
	BaS	DS	MA	NP	MTF	PS	TMR	RA
1	0.33	0.33	0.31	0.31	0.33	0.38	0.32	0.31
2	0.37	0.36	0.35	0.37	0.32	0.37	0.35	0.36
3	0.38	0.25	0.41	0.35	0.46	0.4	0.35	0.41
4	0.51	0.46	0.51	0.51	0.47	0.5	0.5	0.49
5	0.39	0.44	0.43	0.41	0.36	0.37	0.43	0.42
6	0.4	0.28	0.38	0.29	0.37	0.38	0.29	0.32
7	0.34	0.25	0.3	0.29	0.29	0.41	0.29	0.27
8	0.36	0.37	0.39	0.39	0.44	0.41	0.4	0.36
9	0.27	0.29	0.29	0.29	0.26	0.21	0.3	0.31
10	0.33	0.33	0.35	0.34	0.35	0.32	0.34	0.33
11	0.34	0.44	0.33	0.39	0.34	0.27	0.42	0.4
12	0.5	0.46	0.46	0.5	0.5	0.52	0.45	0.45
13	0.36	0.3	0.38	0.35	0.37	0.34	0.34	0.31
14	0.45	0.32	0.4	0.4	0.52	0.51	0.32	0.3
15	0.45	0.38	0.39	0.45	0.52	0.45	0.42	0.43
16	0.39	0.39	0.34	0.41	0.4	0.4	0.41	0.41
17	0.47	0.4	0.44	0.45	0.54	0.5	0.44	0.42
18	0.44	0.29	0.45	0.36	0.39	0.45	0.36	0.36
19	0.43	0.36	0.39	0.44	0.5	0.43	0.41	0.42
20	0.44	0.42	0.39	0.4	0.42	0.46	0.35	0.37

Note. Balanced Scoring (BaS), Dichotomous Scoring (DS), Morgan Algorithm (MA) Negative Scoring (NP), Multiple True/False Scoring (MTF), Positive Counting (PS), Trap Multiple Response (TMR), Ripkey Algorithm (RA)

Table H

Results of Comparison of Item Discrimination Indices of Turkish Test Calculated from Different Scoring Methods

Item Number	Method	r_{jx}	z_r	z
1	PS	0.38	0.4	0.934
	RA	0.31	0.32	
2	NP	0.37	0.388	0.6669
	MTF	0.32	0.331	
3	MTF	0.32	0.331	2.8416**
	DS	0.36	0.376	
	MA	0.35	0.365	2.6935**
	DS	0.36	0.376	
	RA	0.41	0.435	2.1168*
	DS	0.36	0.376	
	PS	0.4	0.423	1.9763*
	DS	0.36	0.376	
4	BaS	0.51	0.562	0.3132
	RA	0.49	0.536	
5	DS	0.44	0.472	11,200
	MTF	0.36	0.376	
6	BaS	0.4	0.423	15,972
	DS	0.28	0.287	
7	PS	0.41	0.435	2.1168*
	DS	0.25	0.255	
8	MTF	0.44	0.472	11,200
	RA	0.36	0.376	
9	RA	0.31	0.32	12,613
	PS	0.21	0.213	
10	MTF	0.35	0.365	0.397
	PS	0.32	0.331	
11	DS	0.44	0.472	22,950
	PS	0.27	0.276	
	TMR	0.42	0.447	20,067
	PS	0.27	0.276	
12	PS	0.52	0.576	10,765
	RA	0.45	0.484	
13	MA	0.38	0.4	10,636
	DS	0.3	0.309	
14	MTF	0.52	0.576	3.1344**
	RA	0.3	0.309	
	MTF	0.52	0.576	2.8744**
	TMR	0.32	0.331	
	MTF	0.52	0.576	2.8744**
	DS	0.32	0.331	
	PS	0.51	0.562	2.9745**
	RA	0.3	0.309	

Item Number	Method	r_{jx}	z_r	z
	PS	0.51	0.562	2.7146**
	TMR	0.32	0.331	
	PS	0.51	0.562	2.7146**
	DS	0.32	0.331	
15	MTF	0.52	0.576	2.0708*
	DS	0.38	0.4	
16	RA	0.41	0.435	0.9576
	MA	0.34	0.354	
17	MTF	0.54	0.604	2.1204*
	TMR	0.4	0.423	
	MTF	0.54	0.604	2.1204*
	DS	0.4	0.423	
18	MA	0.45	0.484	2.1865*
	DS	0.29	0.298	
	PS	0.45	0.484	2.1865*
	DS	0.29	0.298	
	BaS	0.44	0.472	2.0400*
	DS	0.29	0.298	
19	NP	0.44	0.472	11,200
	DS	0.36	0.376	
20	MTF	0.52	0.576	2.4774**
	TMR	0.35	0.365	
	MTF	0.52	0.576	2.2075*
	RA	0.37	0.388	

Note. Balanced Scoring (BaS), Dichotomous Scoring (DS), Morgan Algorithm (MA) Negative Scoring (NP), Multiple True/False Scoring (MTF), Positive Counting (PS), Trap Multiple Response (TMR), Ripkey Algorithm (RA)