ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
Research Article; 2024; 15(3);193-208

The Effect of Presenting Geometry Items with and without Shapes on the Psychometric Properties of the Test and Students' Test Scores*

İslim ATÇI ** Mustafa İLHAN ***

Abstract

This research aimed to examine the effects of presenting geometry items with and without shapes on the psychometric properties of the test and students' test scores. The study was conducted on 480 eighth grade students. Within the scope of the study, two geometry tests were crafted, one with shapes and the other without shapes. Both tests consisted of 15 multiple-choice items. In the data collection process, a counterbalanced design was followed, and the two tests were administered to the students three weeks apart. Analyses were carried out on 405 students who participated in both applications and whose test forms could be matched. The factor analysis results revealed that the factor loadings of the items and extracted variance were higher for the test with shapes compared to the test without shapes. The Cronbach's alpha coefficient of the test containing shapes was found to be significantly higher than that calculated for the test without shapes. According to item difficulties, the questions with shapes were easier for the students than the shape-free questions. In terms of discrimination indices, a difference in favor of the shape-containing test was observed in almost all items. Ferguson's delta statistic, which is a measure of discrimination for the overall test, was higher in the shape-containing test. Correlation analysis denoted a strong positive relationship between students' scores on the two tests. The paired samples *t*-test proved that there was a statistically significant difference between students' scores on the tests with and without shapes. These results indicate that the geometry tests with and without shapes differ in terms of both psychometric properties and students' test scores.

Keywords: Shape-containing geometry items, geometry items without shapes, psychometric properties

Introduction

Achievement is an abstract construct that cannot be directly observed but can be indirectly measured through tests. Therefore, reaching accurate estimations about individuals' achievement depends first and foremost on the quality of the test. There are two basic questions to be addressed when developing tests: (1) What are we going to measure, and (2) How can we measure this targeted characteristic (Lindquist, 1936)? In order to clarify what is to be measured, a test plan is usually prepared using a specification table. On the other hand, when it comes to the question of how to measure the targeted trait, various dilemmas arise (Rodriguez, 2002). These dilemmas may be related to item type, preparation of answer choices, or the structure of the item stem.

In item type dilemmas, the most appropriate item type (multiple-choice, true-false, open-ended, etc.) is decided by taking into account the construct being measured, the cognitive level of the learning objective being tested, and the number of examinees. In order to help the test developers/researchers in this decision process, many studies have been conducted to reveal how item type affects validity, reliability, item difficulty and discrimination, item response time, and examinees' ability scores (Bacon, 2003;

*This study has been produced from Master's Thesis that was conducted under the supervision of the Assoc. Prof. Dr. Mustafa İLHAN and prepared by İslim ATÇI.

** Teacher, 19 Mayıs Middle School, Ministry of National Education of the Republic of Türkiye, Mardin-Türkiye, islimatc@gmail.com, ORCID: 0009-0008-1729-4945
***Assoc. Prof. Dr., Dicle University, Ziya Gökalp Faculty of Education, Diyarbakır-Türkiye, mustafailhan21@gmail.com, ORCID: 0000-0003-1804-002X

To cite this article:
Atçı, İ. & İlhan, M. (2024). The effect of presenting geometry items with and without shapes on the psychometric properties of the test and students' test scores. *Journal of Measurement and Evaluation in Education and Psychology, 15*(3), 193-208. https://doi.org/10.21031/epod.1483567

Received: 14.05.2024
Accepted: 24.09.2024

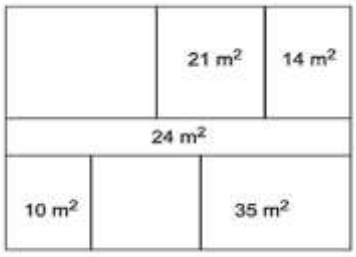
Cheng, 2004; Demir, 2010; Gültekin & Çıkrıkçı Demirtaşlı, 2012; İlhan et al., 2020; Yılmaz Koğar & Soysal, 2023; Öksüz & Güven Demir, 2019; Özer Özkan & Özaslan, 2018; Temizkan & Sallabaş, 2015; Zulaiha et al., 2021). Dilemmas about the answer choices focused on the effects of the following issues on measurements: optimal number of options (Atalmış, 2018; Baghaei & Amrahi, 2011; Haladyna & Downing, 1993; Nwadinigwe & Naibi, 2013; Raymond et al., 2019; Rodriguez, 2005; Vegada et al., 2016), the order in which options are presented (Cizek, 1994; Hohensinn & Baghaei, 2017; Karanfil & Neufeld, 2020; Lions et al, 2021; Lions et al., 2023; Shin et al., 2019), options' homogeneity (Ascalon et al., 2007; Atalmış & Kingston, 2018), and the options such as "all of the above" and "none of the above" (Atalmış & Kingston, 2017; Bishara & Lanzo, 2014; Crehan et al., 1993; Jonsdottir et al., 2021). When it comes to the item stem, in addition to issues that concern all disciplines such as the effects of item stem length (Abd El-Mohsen, 2008; Koepf, 2018), its completeness (in the form of a full or incomplete sentence) (Ascalon et al., 2007; Schaefer, 2009; Violato, 1991; Violato & Harasym, 1987; Violato & Marini, 1989) and orientation (negatively or positively) (Harasym et al., 1992; Harasym et al., 1993; Terranova, 1969) on psychometric qualities, field-specific dilemmas also arise. Mathematics is one of the disciplines where different dilemmas occur when writing item stem.

One of the basic dilemmas in item stem writing in mathematics tests is whether it would be more appropriate to compose the computational items with mathematical expressions or word problems, and whether such a change in the item stem would make a difference in the measurements (Kan et al., 2019). Another important dilemma appears in geometry, a sub-branch of mathematics. Just as items in the algebra and arithmetic areas of the mathematics tests can be written with word problems or mathematical expressions, geometry items can also be created with or without shapes. The two items in Figure 1, which the Ministry of National Education of Türkiye Republic included in the numerical ability test of the 2018 High School Entrance Examination, exemplify this.

Figure 1

The samples for geometry items with and without shapes

(a)



Above, the areas of some sections are given on the rectangular floor plan, where each section is rectangular.

If the side lengths of each of these rectangles are natural numbers in meters, the sum of the areas of the parts whose areas are not given is at least how many square meters?

A) 36 B) 54 C) 64 D) 76

(b)

A square-shaped garden with a side length of 10 m has an irrigation system only at the corners. Each irrigation system can irrigate up to a section up to 4 m away from its location. In the part of this garden that cannot be irrigated, there is a pergola with a square base. The diagonal of the base of this pergola coincides with the diagonal of the garden.

What is the maximum floor area of this pergola, whose base diagonal length is a natural number in meters?

A) 18 B) 48 C) 52 D) 72

As can be seen in the Figure 1.a, a geometric shape was presented in the item and the basic information related to the question was explained on this shape. In the question in the Figure 1.b, on the other hand, all information was given verbally and no geometric shape was provided. Such differences in the item stem can affect the cognitive processes that need to be employed to answer the item correctly (Kan et al., 2019). For example, the visuospatial skills needed to solve the geometry items with and without shapes may differ. In a similar vein, answering a geometry question that does not contain shapes and consists only of verbal expressions correctly may require more intensive verbal skills. In order to be able to develop more purposeful geometry tests and to read the measurement results more accurately, it is necessary to know the effect of such differences in the item stem on the measurements.

The Purpose and Importance of the Research

When writing geometry items, the test developer may encounter the following quandaries: *(a)* Should the shapes be presented compatible or incompatible with their actual values (Çetin & Türkan, 2013)?, *(b)* Should prototype drawings corresponding to the most familiar model of the geometric shape or non-prototypical drawings be employed? Certainly, the problem that is as important as these, perhaps even before these, is how the presentation of the item with shapes and only verbal expressions without shapes will affect the measurements. To put it more clearly, one of the research problems that needs to be answered is whether presenting geometry questions with or without shapes will make a difference in students' test scores and the psychometric properties of the test. However, when the literature is examined, it is seen that the number of studies on this subject is quite limited. One of these studies was conducted by Aydın et al. (2006) with 12th grade students, and students in the same class were randomly divided into two groups. Students in the first group answered the geometry test in which verbal expressions and shapes were presented together. The other group was administered a test consisting only of verbal expressions without shapes. As a result of the research, they determined that the averages in the group to which the shape-containing was applied were higher in all items. Karpuz et al. (2014), on the other hand, carried out a qualitative study to analyze students' responses to shape-containing and shape-free geometry questions comparatively. In the literature, no study was found to examine the impact of presenting geometry items with and without shapes on the psychometric properties of the test and whether it created a significant difference in students' test scores.

This empirical research attempts to determine the effect of presenting geometry items with or without shapes on the psychometric properties of the test and students' test scores. For this purpose, answers to the following problems were sought in the study.

1. Do the test forms in which geometry questions are presented with or without shapes differ in terms of *(a)* factor structures, *(b)* item difficulty and discrimination indices, and *(c)* internal consistency coefficients?
2. *(a)* What is the relationship between students' scores on geometry tests with and without shapes? *(b)* Is there a statistically significant difference between their scores of these two tests?

Since the past studies comparing the geometry tests with and without shapes have not dealt with these research problems, it is thought that this study has original value and will contribute to both mathematics education and measurement and evaluation literature. It is hoped that the study results will benefit teachers, mathematics education, and measurement and evaluation experts in the preparation of geometry tests and also indirectly shed light on the points to be considered in geometry teaching processes.

Methods

Research Design

The present study employed a descriptive-comparative design to contrast geometry tests with and without shapes. This research approach focuses on two variables, compares these variables by following a well-planned but not manipulated formal process, and aims to reveal which of the two variables/situations is better as a result of the comparison (Paler-Calmorin & Calmorin, 2007).

Participants

The study was carried out with eighth grade students because secondary school students were a more accessible group for the researcher who carried out the data collection process and because the number of objectives learned by eighth grade students in the field of geometry learning area was higher compared to secondary school students in lower grades. Accordingly, 480 eighth grade students from three different schools in Mardin province constituted the participants of the study. Nevertheless, 46 students who participated in one of the with and without shapes test administrations but did not

participate in the other, and 29 students whose forms could not be matched because they did not use the same nickname on the two tests they answered, were excluded from the analysis. Therefore, the analyses were carried out on a total of 405 students, 218 (53.83%) of whom were female and 187 (46.17%) of whom were male, who participated in both tests and whose answered test forms could be matched.

Instruments

Research data was collected via two tests prepared to cover the objectives of the 7th grade and the previous terms in the geometry learning field of the mathematics curriculum. Both tests had 15 multiple choice items. While the first test consisted of geometry questions with shape, in the second one the items were presented only with verbal expressions. The two test forms were parallel except that one was prepared with shapes and the other consisted of only verbal expressions without any shapes. In both tests, the items had four options. The order of the items and options, and the option corresponding to the correct answer were identical for the two tests.

After the draft form for the tests was created, opinions were received from two field experts. The first of the experts was a faculty member whose field of study includes geometry teaching, who gives courses on geometry teaching at the undergraduate level, and who has a doctorate in mathematics education. The second expert was a mathematics teacher with 10 years of professional experience. These two experts reviewed the items in terms of their suitability for the research purpose and scientific accuracy. Experts expressed their opinion that formal corrections were needed in the items. For example, they pointed out that there are differences in terms of font and font size from one item to another in the tests and that there should be a standard in this regard. Besides, they emphasized that the "x" symbol used to indicate angle or length was written with a capital letter in some questions and with a lowercase letter in some questions, and recommended the use of lowercase letters in all questions. Furthermore, the wording of some questions was changed based on the experts' opinions. Additionally, one of the field experts proposed that the first item may be removed from the tests, saying, "*This item will mostly be solved correctly, whether it is presented with or without a shape.*" However, considering that the tests were crafted according to the objectives of the 7th grade and previous years and that the students may have forgotten the topics, it was thought that the ease of the first item would increase students' motivation for the test. Hence, it was decided to keep the relevant item in the test.

Following the changes that were made based on the opinions of field experts, the opinion of an expert with the title of associate professor in the field of measurement and evaluation in education was consulted. This expert whose undergraduate was in the field of secondary school mathematics teaching reviewed both tests and he specified that the items were in accordance with measurement principles such as *(a)* emphasizing negative judgements, *(b)* listing the options from the largest to smallest or smallest to largest, and *(c)* ensuring a balanced distribution of the correct answer among the options. Finally, a Turkish teacher had a look at the tests in terms of spelling and punctuation rules, and necessary corrections were made to the items in line with her comments. Then, a preliminary trial was conducted on 12 students who were heterogeneous in terms of mathematics achievement. After answering the test, the students were interviewed, and there were no statements that the students had difficulty understanding in the items or in the test instructions. Thus, it was concluded that the tests were ready for application. An example of geometry items with and without shapes was presented in the Appendix.

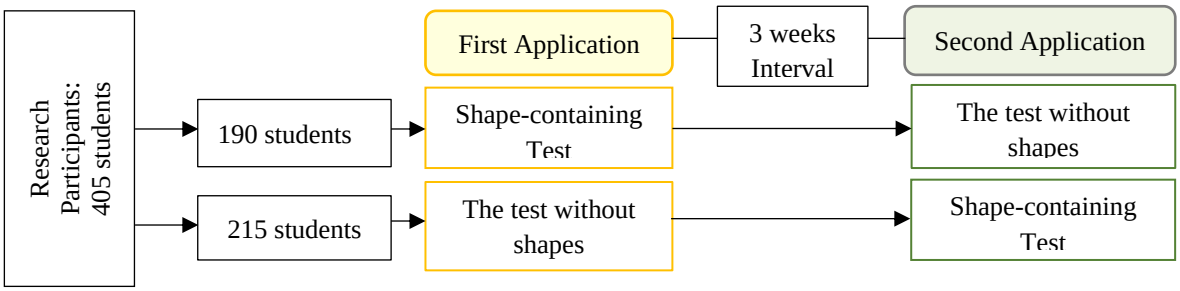
Data Collection Process

The data collection process was completed in two stages. In studies where two different instruments are administered to the same group at a certain interval, there may be effects arising from the order in which the instruments are applied, and this may threaten the internal validity of the research (Corriero, 2017). To prevent this, that is, to eliminate possible effects that may arise from the sequence of application of the tests, it is recommended to use a counterbalanced design (Graveter & Forzano, 2018). In this context, research data was collected according to the pattern summarized in Figure 2. Half of the group was first administered the shape-containing test and then the test without shapes, while the other half followed the reverse order.

Another important issue in studies involving repeated measures is the time elapsed between two applications. This period should be long enough that students do not remember their answers and short enough that participants do not experience changes due to maturation/learning (Crocker & Algina, 1986; Goldfarb, 2021). Since the time required to achieve this will vary depending on the developmental characteristics of the participant group and the nature of the measured construct (Mitchell et al., 2000), there is no clear opinion on how much time should be between two applications. However, a period of 2-3 weeks is generally considered ideal for achievement tests. Therefore, in the current study, two test forms were administered to the students three weeks apart.

Figure 2

The path followed in the collection of the research data



Before the data collection process, approval was received from Dicle University, Social and Human Sciences Ethics Committee regarding the compliance of the study with current scientific ethical principles. In the next step, a preliminary application was made for Research, Competition and Social Event permission on the Ministry of National Education website. After the application, the leave petition was submitted for approval by the Mardin/Artuklu District Governorship. Once all necessary permissions obtained, data collection started in secondary schools in Artuklu district of Mardin province, and the applications were carried out in the classroom environment, in paper-pencil form and on a voluntary basis, between December 2023 and January 2024. Prior to application, students were informed about the aim of the study and it was emphasized that the data would be used only for scientific purposes and would not be shared with any other person or institution. In addition, students were reminded that they did not need to write their actual names on the test forms, but it was stated that they should write a nickname that they would not forget in the space provided at the beginning of the tests in order to match the two test forms they would answer. There was no student who refused to participate in the study in any classroom where the application was carried out.

Data Analyses

Procedures for data analysis were presented under three headings: preliminary analyses, validity and reliability analyses, and analysis for comparing the students' scores in two tests.

Preliminary Analyses

This title includes the processes carried out to prepare the data sets for analysis and the results of the analyses applied to check the distribution of the data. While performing the analyses in question, the JASP 0.18.1.0 program (JASP Team, 2022) and the web tool running R software in its background developed by Aybek (2021) were utilized. Since the multiple-choice tests were employed as the instruments in the study, correct answers were scored as 1, and incorrect answers and blank items were scored as 0. Therefore, there were no missing values in the data file. Moreover, univariate and multivariate outliers were not found in the data set. After this determination, skewness and kurtosis coefficients were examined for univariate normality and Henze–Zirkler statistics for multivariate normality. Table 1 depicts the results of the normality test.

Table 1

The results for univariate and multivariate normality tests

Test	Skewness		Kurtosis		Henze–Zirkler
	Statistic	Std. Error	Statistic	Std. Error	
With shapes	.14	.12	-1.15	.24	4.45*
Without shapes	.76	.12	.02	.24	2.22*

*p<.01

The fact that the skewness and kurtosis coefficients are within ±1.5 is judged as the evidence of univariate normality (Tabachnick & Fidell, 2013). Accordingly, it is understood that the research data meet the assumption of univariate normality. The statistical significance of the Henze Zirkler test, on the other hand, indicates that multivariate normality is violated.

Validity and Reliability Analyses

In the research, Exploratory Factor Analysis (EFA) was performed to ascertain the factor structure of the tests. In EFA, the Kaiser–Meyer–Olkin (KMO) values were found to be .781 and .541 for the shape-containing test and shape-free test, respectively. Besides, Bartlett's sphericity test results were significant for both forms [$\chi^2_{shape-containing\ test}$=3634.221, $\chi^2_{shape-free\ test}$= 2386.016; *df* = 105; p < .001]. The calculated KMO values over .50 and the statistically significant Bartlett's tests reflect that the sample is adequate and the correlation matrices are suitable for acquiring reliable factors (Field, 2013). Thereby, the analysis continued and since the multivariate normality assumption was violated, the principal axis factoring technique (Şahin, 2022), which does not require any prerequisites about the distribution of the data, was operated as the estimator in EFA. Parallel analysis method was used to decide the number of factors, and analyses were carried out based on the tetrachoric correlation matrix as the data had a dichotomous (1-0) structure.

Subsequently, the two test forms were compared in terms of item difficulty and discrimination indices, and reliability. For the items' discrimination, the discrimination index (r_{jx}) based on 27% lower-upper group comparisons and the point biserial correlations (r_{pb}) were calculated. Also, in order to attain a discrimination index for the entire test, Ferguson's delta (δ) statistic (Ferguson, 1949) was calculated using Equation 1, where *k* is the number of items, *n* is the number of test takers (i.e., sample size) and *f* is the frequency value of each score. Ferguson's delta provides information about how heterogeneous the examinees' test scores are (Zhang & Lidbury, 2013) and can take values ranging from 0 to 1 (Hernandez and Zalava, 2017). The value of .90 is recommended as the threshold for this statistic (Kline, 1993).

$$\delta = \frac{(k+1)(n^2 - \sum f^2)}{kn^2} \qquad (1)$$

Within the scope of reliability analysis, Cronbach's alpha internal consistency coefficients were calculated and the significance of the difference between the Cronbach's alpha values of the two forms was tested using the method recommended by Feldt et al. (1987). In psychometric analyses, JASP 0.18.1.0 program (JASP Team, 2022) was utilized for EFA. While reliability coefficients and item statistics were calculated in TAP (Test Analysis Program) software (Brooks & Johanson, 2003), Ferguson's delta statistics were computed in Microsoft Excel. To compare Cronbach's alpha coefficients statistically, on the other hand, the interface running the cocron package in R programming language, developed by Diedenhofen and Musch (2016) was used.

Analyses to Compare Students' Scores in Two Tests

Since the research data held the univariate normality assumption, the relationship between student scores in the shape-containing and shape-free tests was examined by means of Pearson product-moment correlation. When interpreting the size of correlation coefficient, the following ranges offered by Salkind (2010) were taken as reference: between .00 and .20, very weak; between .20 and .40, weak, between .40 and .60, moderate; between .60 and .80, strong; between .80 and 1.00, very strong. Paired samples *t*-test was implemented to test the significance of the difference between the scores of the students from the two tests. In order to evaluate the magnitude of the significant difference observed as a result of the *t*-test, Cohen's *d* statistic was inspected. Cohen (1977) defined the cut-off points for small, medium, and large effects as .20, .50, and .80, respectively. Relying on this guideline, Cohen's *d* was interpreted as follows in the current research: if $d < .20$ the difference is negligible, if $.20 < d < .50$ the difference is small, if $.50 < d < .80$ the difference is moderate and if $d > .80$ the difference is large. Analyses to compare students' scores in the two tests were conducted in the JAPS 0.18.1.0 program.

Results

Firstly, EFA was applied for the tests with and without shapes. Table 2 shows the outputs reported in parallel analysis for the number of factors in EFA.

Table 2
Results from parallel analysis for number of factors in the tests with and without shapes

	Test with shapes		Test without shapes	
	Eigenvalues for Real Data	Eigenvalues for Simulated Data	Eigenvalues for Real Data	Eigenvalues for Simulated Data
Factor1	7.076*	1.348	4.720*	1.348
Factor2	1.201	1.264	1.672*	1.264
Factor3	0.967	1.204	1.418*	1.204
Factor4	0.873	1.158	1.208*	1.158
Factor5	0.853	1.113	0.973	1.113

Table 2 illustrates that the number of factors where the eigenvalue of the real data is greater than the eigenvalue of the simulated data was 1 in the test with shapes and 4 in the test without shapes. In other words, there was a unidimensional structure in the shape-containing test and a four-factor structure in the test without shapes. However, no interpretable structure was observed when the distribution of the items to the factors in the shape-free test was examined. More explicitly, the four factors that emerged could neither be associated with the theoretical framework such as the objectives measured by the items, nor with statistical features such as the items' difficulty indices. For this reason, considering that all items were written in a way to test the objectives belonging to the same learning domain, the number of factors was limited to 1 for the shape-free test and EFA was redone. Table 3 displays the factor analysis results obtained for the tests with and without shapes after the repeated EFA.

As can be seen from Table 3, the factor loadings of all items except Item 8 in the test without shapes are above the threshold value of .30 (Büyüköztürk, 2010). It is also noteworthy that the factor loadings of the items are generally higher in the shape-containing test compared to the test without shapes. In line with the factor loadings, the extracted variance ratio was also higher in the shape-containing test than in the shape-free one.

Table 3

Factor solutions reported in EFA for tests with and without shapes

Items	Factor Loadings	
	Test with shapes	Test without shapes
Item1	.676	.397
Item2	.595	.474
Item3	.527	.589
Item4	.495	.549
Item5	.767	.494
Item6	.624	.432
Item7	.499	.382
Item8	.781	.262
Item9	.599	.511
Item10	.715	.605
Item11	.686	.687
Item12	.730	.592
Item13	.796	.566
Item14	.595	.358
Item15	.715	.681
Extracted Variance	43.60%	26.90%

Following the factor analysis, item difficulty and discrimination indices were examined. Table 4 provides the results regarding item statistics.

Table 4

The results of the item analysis for the tests with and without shapes

Items	Test with shapes			Test without shapes		
	p	r_{jx}	r_{pb}	p	r_{jx}	r_{pb}
Item1	.85	.37	.48	.85	.26	.30
Item2	.54	.58	.53	.34	.46	.45
Item3	.33	.52	.47	.38	.48	.51
Item4	.64	.56	.46	.43	.53	.49
Item5	.53	.78	.66	.32	.46	.48
Item6	.48	.67	.56	.51	.52	.43
Item7	.64	.53	.46	.54	.45	.39
Item8	.66	.74	.64	.32	.23	.31
Item9	.64	.63	.53	.46	.61	.50
Item10	.59	.73	.62	.57	.64	.53
Item11	.37	.69	.58	.30	.58	.56
Item12	.47	.79	.63	.33	.52	.50
Item13	.51	.81	.67	.28	.54	.49
Item14	.44	.69	.54	.28	.37	.38
Item15	.51	.75	.62	.38	.66	.56
Mean	.547	.563	.655	.413	.490	.458

It can be seen from the statistics in Table 4 that compared to the test without shapes, the difficulty indices in the shape-containing test are closer to 1 in most items. Accordingly, it is understood that among the geometry questions that aim to test the same learning objectives, the ones without shapes are more difficult for the students than the ones with shapes. This can be seen more clearly in the mean difficulties calculated for the tests. Table 4 shows that the discrimination indices based on upper-lower group comparisons exceeded the .30 lower limit (Ebel & Frisbie, 1991; Erkuş, 2012) in all items except Item 1 and Item 8 in the test without shapes. The discrimination values based on point biserial correlation

meet the .30 criterion for all items in both test forms. These results reflect that both shape-containing and shape-free tests have acceptable discrimination. However, both the item-based discrimination indices and the mean discrimination values of the tests indicate that the form with shapes can distinguish students at different achievement levels better than the shape-free one.

In the present study, Ferguson's delta (δ) statistic was also explored to obtain additional evidence about the difference between the discrimination powers of the tests and it was found .986 and .961 for the tests with and without shapes, respectively. That's to say, Ferguson's delta statistic exceeds the cut-off value of .90 for both tests. These results hint that the scores of both shape-containing and shape-free tests are heterogeneous enough to assert that the instruments are distinctive. Nevertheless, the greater Ferguson's delta statistic regarding the test with shapes implies that this test is more discriminatory than the test without shapes. Following the analysis of the distinctiveness of the tests, the internal consistency of the measurements obtained from the two tests was scrutinized. Table 5 shows the Cronbach's alpha coefficients calculated for the tests with and without shapes, along with the chi-square value for the significance of the difference between these coefficients.

Table 5

Internal consistency coefficients calculated for the tests with and without shapes

Tests	Cronbach's alpha	n	df	χ^2
With Shapes	.847 (95% CI [.825, .867])	405	1	12.459*
Without Shapes	.734 (95% CI [.696, .770])			

*$p < .001$

Table 5 denotes that Cronbach's alpha coefficients are above .70 (Pallant, 2005), which is the most commonly accepted lower limit for reliability for both tests. The Cronbach's alpha value for the test with shapes was higher than the test without shapes, and the difference between the internal consistency coefficients was statistically significant. Finally, the relationship between the scores the students received from the two tests was examined and whether the difference between the scores was significant or not was explored. Table 6 contains the results of the correlation analysis and paired samples *t*-test applied for this purpose.

Table 6

The results of correlation analysis and paired samples t-test for the scores of the tests with and without shapes

Tests	Mean	SD	Pearson r	Paired samples *t*-test		
				t	df	Cohen d
With Shapes	8.202	4.079	.648* (%95 CI [.588, .701])	12.78*	404	.635
Without Shapes	6.195	3.236				

* $p < .001$

As seen in Table 6, students' scores in the test with shapes were higher than the test without shapes. The calculated correlation coefficient reflects that there was a strong positive relationship between the students' scores in the two tests. The outputs of paired samples *t*-test indicated that there was a statistically significant difference between students' scores in the tests with and without shapes. The Cohen's *d* statistic represents a medium-sized difference, thus signs that the statistically significant difference detected was also noteworthy in practice.

Conclusion and Discussion

In this study, the effects of presenting geometry items with and without shapes on the psychometric properties of the test and students' test scores were examined. First, two tests were compared in terms of factor structures. The results showed that the factor loadings and extracted variance of the shape-containing test were higher compared to the form without shapes. This result reflects that the test with shapes serves the purpose more, in other words, it produces more valid measurements. Since there is no

visual support for the students in the test without shapes, variables such as language skills and reading comprehension ability may have a greater impact on the students' performance in this test compared to the shape-containing one. The interference of such sources of variability in the measurement results other than geometry knowledge may have decreased the validity of the measurements. The fact that a multidimensional structure emerged in the shape-free test when there was no limitation on the number of factors supports this view. As a matter of fact, Kan et al. (2019) compared the tests in which the item stem was presented in mathematical expressions and verbal form in terms of dimensionality and found that the two item types differed in terms of the skills required to answer the question correctly.

When the two forms were compared in terms of item difficulty indices, it was disclosed that the test with shapes was easier for the students than the one without shapes. That is to say, while students showed higher success in the form with shapes, they had difficulty in solving the questions when the same items were presented with only verbal expressions. This finding is coherent with the results of the study conducted by Karpuz et al. (2014). Karpuz et al. (2014) prepared two tests, one in which the concept and the shape were presented together, and the second in which the concept was presented but the shape was not. They administered these tests, both of which contained eight open-ended questions, to 120 high school students, one month apart, first the form without shapes and then the form with shapes. As a result of the study, they reported that students were more successful in solving the shape-containing questions. Drawing wrong shapes that do not meet the generalizability condition, being mostly influenced by prototype shapes while solving, or not being able to create any shape that corresponds to the conceptual knowledge in the item were listed as the factors that caused students to have more difficulty in solving the questions without shapes. A similar result was obtained in the study by Aydın et al. (2006). In the study just mentioned, an application was made over five open-ended geometry questions and students were randomly divided into two groups in the classroom environment. The students in the first group were presented the items with both verbal expressions and shapes. The students in the second group were asked the same questions without shapes. As a result of the study, they observed that when the item was presented only with verbal expressions, students had difficulty in transferring the expression in the question to the shape and consequently had more difficulty in these types of questions. The attained results regarding the item difficulties also match with the positions of Michael–Chrysanthou et al. (2024) who stated that "*A figure is a representation of a geometrical situation easier to understand compared to a representation with linguistic elements only*." Therefore, it can be said that the item difficulties calculated in this study for the tests with and without shapes are in line with the results of the previous studies.

The fact that students have more difficulty with the geometry questions without shapes can also be explained by the fact that they rarely encounter these types of questions. Because when exams do not go beyond certain question types, namely, when students always see similar types of items, they may have difficulty with different types of questions (Yılmaz, 2007). This situation manifested itself in the opinions written by the students under the questions in the test without shapes. Although there was not a particular space on the tests for students to write their comments about the questions, some students expressed that they did not know what to do with the shape-free questions and wrote notes on their test papers such as "*I can't understand as there is no shape*", "*algebraic expressions were already all we need in the geometry questions*", "*what kind of geometry question are these"*. Based on these opinions, it can be asserted that students' unfamiliarity with shape-free geometry items caused them to have more difficulty with these questions.

Another noteworthy result regarding item difficulties was as follows: The difficulty indices calculated for the first item in the tests with and without shapes were equal to each other. Indeed, one of the field experts whose opinion was consulted about the tests remarked that this item would be solved correctly whether it was presented with or without a shape. In this sense, the difficulty index calculated for the related item confirmed the expert opinion. Accordingly, it may be useful to get the experts' opinions before the application about the necessity of the shape in geometry questions or in which questions removing the shape may make a difference in the difficulty index.

When the item discriminations calculated for the two tests were compared, a difference was observed in favor of the shape-containing form. This result reflects that when geometry questions are presented with shapes, students with different levels of achievement can be effectively discriminated from each other, whereas when the same questions are presented only with verbal expressions, it becomes difficult to distinguish students at different achievement levels. In line with this, the Ferguson's delta statistic, which provides information about how heterogeneous the examinees are in terms of their test scores, was higher for the shape-containing test. This difference between the discrimination powers of the tests was reflected in Cronbach's alpha coefficients, and a significantly higher internal consistency coefficient was estimated for the shape-containing test compared to the one without shapes. Accordingly, it is possible to conclude that there is a higher consistency between the items of the shape-containing test and that the shape-free form is more prone to random errors than the one with shapes. In the literature, there is no study directly comparing geometry tests with and without shapes in terms of discrimination and internal consistency. However, considering that there is a difference between the discrimination values and internal consistency coefficients of the tests even when geometric shapes are presented in accordance with their real values and different from their real values (Çetin & Türkan, 2013), it is thought that presenting the questions without shapes will affect these statistics more explicitly. Therefore, it can be said that the results of the study conducted by Çetin and Türkan (2013) indirectly support the findings of the study, although not directly.

In the second problem of the study, students' scores in tests with and without shapes were compared. The obtained correlation coefficient elicited that there was a strong positive relationship between the students' scores in the two tests. This result indicates that the two tests ranked the students largely similarly in terms of their geometry achievement. More clearly, there was a high relative agreement between the achievement scores obtained from the geometry tests with and without shapes. On the other hand, there was a statistically significant difference between the students' scores in the two test forms. This finding signs that there is no absolute agreement between the scores of the two tests. Likewise, Aydın et al. (2006) reported that the students' item scores were higher in the test with shapes compared to the one without shapes.

The fact that the students' scores in the test without shapes were significantly lower means that they could not use their conceptual knowledge in questions without shapes, had difficulty in creating visual representations of verbal expressions and were unable to mobilize the knowledge in their minds when they encountered questions without shapes. As a matter of fact, Çiftçi and İşleyen (2022) stated that students comprehend geometry problems in a shape-oriented manner, cannot transfer the verbal expressions to the shapes or they transfer them incorrectly, and that some students even skip the questions presented only verbally without reading them at all. In addition, they emphasized that the fact that students proceed directly through shapes without fully learning geometric concepts comes to light by the deficiencies in visualizing verbal expressions. In a similar vein, Barut and Retnawati (2020) showed the lack of visualization ability as one of the difficulties experienced in geometry lessons in their study conducted with secondary school students. Considering all these, it can be argued that one of the main factors that led students to get lower scores in the shape-free test is the deficiency in visualization skills, which Duval (1998) defines as one of the three basic cognitive processes of geometry teaching (cited in Çiftçi & İşleyen, 2022).

Implications, and Suggestions for Future Researches

The research results demonstrated that geometry tests, in which items are presented with or without shapes, differ in terms of both their psychometric properties and the test scores of the students. From the point of these results, it is possible to offer the following suggestions for practice: First and foremost, when experts from the field of mathematics education and measurement and evaluation need to prepare parallel forms of a geometry test, they should take into account that shape-containing and shape-free questions to test the same learning objective are not equivalent. Considering that students' performance on geometry items without shape is lower, teachers should focus more on conceptual learning in the lesson and provide opportunities for students to draw the shape of a geometric term given a definition.

Supporting the interaction between concept and shape with activities can improve students' visualization and spatial thinking skills and promote their performance on geometry questions without shapes. A similar situation is also valid for the textbooks. Including shape-containing geometry questions as well as shape-free items in textbooks may support students' visualization skills and prevent them from floundering when they encounter shape-free questions.

While interpreting the study results and implications based on these results, it should be kept in mind that the research has certain limitations and further research is needed to overcome these limitations. First of all, the current study was limited to the data obtained from two 15-item tests, one consisting of shape-containing and the other consisting of shape-free questions for eighth grade level. Therefore, it may be recommended to conduct a similar study with students at different grade levels. In addition, in the present investigation, no opinion was requested from the field experts about whether the shape would be necessary or not in the items prepared. In future studies, experts can be consulted, and it can be tested whether the differences found between the item statistics of shape-containing and shape-free questions are compatible with the experts' opinions about the necessity of shape. Finally, it can be tested whether the difference between the item statistics of questions with and without shapes changes according to whether the shape in the question is prototypical or unusual.

Declarations

Author Contribution: İslim ATÇI: conceptualization, methodology, development of the instruments, data collection and analysis, writing, and visualization. Mustafa İLHAN: determining the research problem, methodology, data analysis, writing-review & editing, and supervision.

Conflict of Interest: No potential conflict of interest was reported by the authors.

Ethical Approval: Ethical rules were followed in this research. Ethical approval for the study was received from Dicle University, Social and Human Sciences Ethics Committee dated 10.11.2023 numbered 598100.

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

Consent to Publish: Written consent was sought from each author to publish the manuscript.

Competing Interests: The authors have no relevant financial or non-financial interests to disclose.

References

Abd El-Mohsen, M. M. (2008). *The effect of stem length in multiple choice questions on item difficulty in syllabus-based vocabulary test items difficulty in syllabus-based vocabulary test items* [Unpublished Master Theses, The American University in Cairo]. Retrieved from https://fount.aucegypt.edu/retro_etds/2195/

Ascalon, M. E., Meyers, L. S., Davis, B. W., & Smits, N. (2007). Distractor similarity and item-stem structure: Effects on item difficulty. *Applied Measurement in Education*, *20*(2), 153–170. https://doi.org/10.1080/08957340701301272

Atalmış, E. H. (2018). The use of three-option multiple choice items for classroom assessment. *International Journal of Assessment Tools in Education*, *5*(2), 314–324. https://doi.org/10.21449/ijate.421167

Atalmış, E. H., & Kingston, N. (2017). Three, four, and none of the above options in multiple-choice items. *Turkish Journal of Education*, *6*(4), 143–157. https://doi.org/10.19128/turje.333687

Atalmış, E. H., & Kingston, N. M. (2018). The impact of homogeneity of answer choices on item difficulty and discrimination. *Sage Open*, *8*(1). https://doi.org/10.1177/2158244018758147

Aybek, E. C. (2021). *Data preparation for factor analysis*. https://shiny.eptlab.com/dp2fa/

Aydın, E., Kertil, M., Yılmaz, K., & Topçu, T. (2006). *Examining contextual support in geometry learning in terms of student and question level* [*Geometri öğreniminde bağlamsal desteğin öğrenci ve soru seviyesi açısından incelenmesi]* [Full text oral presentation]. VII. National Science and Mathematics Education Congress, Gazi University, Gazi Education Faculty, Ankara.

Bacon, D. R. (2003). Assessing learning outcomes: A comparison of multiple-choice and short-answer questions in a marketing context. *Journal of Marketing Education*, *25*(1), 31–36. https://doi.org/10.1177/0273475302250570

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.

Barut, M. E. O., & Retnawati, H. (2020). Geometry learning in vocational high school: Investigating the students' difficulties and levels of thinking. *Journal of Physics: Conference Series 1613*(1), 012058. https://doi.org/10.1088/17426596/1613/1/012058

Bishara, A. J., & Lanzo, L. A. (2014). All of the above: When multiple correct response options enhance the testing effect. *Memory*, *23*(7), 1013–1028. https://doi.org/10.1080/09658211.2014.946425

Brooks, G. P., & Johanson, G. A. (2003). TAP: Test Analysis Program. *Applied Psychological Measurement*, *27*(4), 303–304. https://doi.org/10.1177/0146621603027004007

Büyüköztürk, Ş. (2010). *A manual of data analysis for social sciences* [*Sosyal bilimler için veri analizi el kitabı]* (11. ed). Pegem Academy.

Cheng, H. (2004). A comparison of multiple-choice and open-ended response formats for the assessment of listening proficiency in English. *Foreign Language Annals*, *37*(4), 544–553. https://doi.org/10.1111/j.1944-9720.2004.tb02421.x

Cizek, G. J. (1994). The effect of altering the position of options in a multiple-choice examination. *Educational and Psychological Measurement*, *54*(1), 8–20. https://doi.org/10.1177/0013164494054001002

Cohen, J. (1977). *Statistical power analysis for the behavioral sciences*. Academic.

Corriero, E. (2017). Counterbalancing. In *The SAGE Encyclopedia of Communication Research Methods* (Vol. 4, pp. 278–281). Sage. https://doi.org/10.4135/9781483381411

Crehan, K. D., Haladyna, T. M., & Brewer, B. W. (1993). Use of an inclusive option and the optimal number of options for multiple-choice items. *Educational and Psychological Measurement*, *53*(1), 241–247. https://doi.org/10.1177/0013164493053001027

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. HarcourtBrace Jovanovich.

Çetin, B., & Türkan A., (2013). The effect of the compatibility and incompatibility of the shapes with their actual values in secondary school 8th grade geometry test questions on the psychometric properties of the test [İlköğretim 8. sınıf geometri testi sorularında şekillerin gerçek değerlerine uygun çizilmesiyle, farklı çizilmesinin testin psikometrik özelliklerine etkisi]. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi*. *4*(2), 52–63. https://doi.org/10.21031/epod.77190

Çiftçi, O., & İşleyen, T. (2022). Üçgenin açıortayları ve kenarortayları konusunda öğrencilerin karşılaştıkları öğrenme güçlükleri. *Ahi Evran Üniversitesi Kırşehir Eğitim Fakültesi Dergisi*, *23*(Özel Sayı), 509–560. https://doi.org/10.29299/kefad.943663

Demir, E. (2010). *Uluslararası öğrenci değerlendirme programı (PISA) bilişsel alan testlerinde yer alan soru tiplerine göre Türkiye'de öğrenci* (Tez No. 257803), [Yüksek lisans tezi, Hacettepe Üniversitesi]. YÖK Ulusal Tez Merkezi.

Diedenhofen, B., & Musch, J. (2016). cocron: A web interface and R package for the statistical comparison of Cronbach's alpha coefficients. *International Journal of Internet Science*, *11*(1), 51–60.

Ebel, R. L., & Frisbie, D. A. (1991). *Essentials of educational measurement* (5th ed.). Prentice Hall.

Erkuş, A. (2012). *Measurement and scale development in psychology–I: Basic concepts and procedures [Psikolojide ölçme ve ölçek geliştirme-1: Temel kavramlar ve işlemler]*. Pegem Academy.

Feldt, L. S., Woodruff, D. J., & Salih, F. A. (1987). Statistical inference for coefficient alpha. *Applied Psychological Measurement*, *11*(1), 93–103. https://doi.org/10.1177/014662168701100107

Ferguson, G. A. (1949). On the theory of test discrimination. *Psychometrika*, *14*(1), 61–68. https://doi.org/10.1007/bf02290141

Field, A. (2013). Discovering statistics using SPSS (3rd ed.). Sage.

Goldfarb, R. (2021). *Consuming and producing research in communication sciences and disorders: Developing power of professor*. Plural.

Graveter, F. J., & Forzano, L. B. (2018). *Research methods for the behavioral sciences* (6th ed.). Cengage.

Gültekin, S., & Demirtaşlı, N. Ç. (2012). Comparing the test information obtained through multiple choice, open-ended and mixed item tests based on item response theory. *Elementary Education Online*, *11*(1), 251–263.

Haladyna, T. M., & Downing, S. M. (1993). How many options is enough for a multiple-choice test item? *Educational and Psychological Measurement*, *53*(4), 999–1010. https://doi.org/10.1177/0013164493053004013

Harasym, P. H., Doran, M. L., & Brant, R., & Lorscheider, F.L. (1993). Negation in stems of single-response multiple-choice items: An overestimation of student ability. *Evaluation & the Health Professions*, *16*(3), 342–357. https://doi.org/10.1177/016327879301600307

Harasym, P. H., Price, P. G., Brant, R., Violato, C., Lorscheider, F. L. (1992). Evaluation of negation in stems of multiple-choice items. *Evaluation & the Health Professions, 15*(2), 198–220. https://doi.org/10.1177/016327879201500205

Hernandez, E. & Zalava, G. (2017). Accurate items for inaccurate in undergraduate physics students. In M.S. Ramírez- Montoya (Eds.), *Handbook of research on driving STEM learning with educational technologies* (pp. 315-340). IGI Global.

Hohensinn, C., & Baghaei, P. (2017). Does the position of response options in multiple-choice tests matter? *Psicológica*, *38*(1), 93–109.

İlhan, M., Boztunç Öztürk, N., & Şahin, M. G. (2020). The effect of the item's type and cognitive level on its difficulty index: The sample of TIMSS 2015. *Participatory Educational Research, 7*(2), 47–59. https://doi.org/10.17275/per.20.19.7.2

JASP Team (2022). *JASP (Version 0.18.1.0)* [Computer software]. https://jasp-stats.org/

Jonsdottir, A. H., Jonmundsson, T., Armann, I. H., Gunnarsdottir, B. B., & Stefansson, G. (2021, 8-9 March). *The effect of the number of distractors and the "none of the above" – "all of the above" options in multiple choice questions* [Conference presentation]. 5th International Technology, Education and Development Conference. https://doi.org/10.21125/inted.2021.1540

Kan, A., Bulut, O., & Cormier, D. C. (2019). The impact of item stem format on the dimensional structure of mathematics assessments. *Educational Assessment*, *24*(1), 13–32. https://doi.org/10.1080/10627197.2018.1545569

Karanfil, T., & Neufeld, S. (2020). The role of order and sequence of options in multiple-choice questions for high-stakes tests of English language proficiency. *International Journal of Applied Linguistics and English Literature*, *9*(6), 110–129. https://doi.org/10.7575/aiac.ijalel.v.9n.6p.110

Karpuz, Y., Koparan, T., & Güven, B. (2014). Students' use of shape and concept knowledge in geometry [Geometride öğrencilerin şekil ve kavram bilgisi]. *Turkish Journal of Computer and Mathematics Education*, *5*(2), 108–118. https://dergipark.org.tr/en/pub/turkbilmat/issue/21573/231505

Kline, P. (1993). *Handbook of psychological testing* (2nd ed.). Routledge.

Koepf, T. M. (2018). *The effect of item stem and response option length on the item analysis outcomes of a career and technical education multiple choice assessment* [Unpublished Doctoral Dissertation, Western Michigan University]. Retrieved from https://scholarworks.wmich.edu/dissertations/3366/

Lindquist, E. F. (1936). The theory of test construction. In H. W. Hawkes, E. F. Linquist & C. R. Mann (Eds.), *The construction and use of achievement examinations: A manual for secondary school teachers* (pp. 17–106). Houghton Mifflin.

Lions, S., Dartnell, P., Toledo, G., Godoy, M. I., Córdova, N., Jiménez, D., & Lemarié, J. (2023). Position of correct option and distractors impacts responses to multiple-choice items: Evidence from a national test. *Educational and Psychological Measurement*, *83*(5), 861–884. https://doi.org/10.1177/00131644221132335

Lions, S., Monsalve, C., Dartnell, P., Godoy, M. I., Córdova, N., Jiménez, D., Blanco, M. P., Ortega, G., & Lemarié, J. (2021). The position of distractors in multiple-choice test items: The strongest precede the weakest. *Frontiers in Educiton*, *6*, 731763. https://doi.org/10.3389/feduc.2021.731763

Michael–Chrysanthou, P., Panaoura, A., & Gagatsis, A. (2024). Exploring secondary school students' geometrical figure apprehension: cognitive structure and levels of geometrical ability. *Educational Studies in Mathematics*. https://doi.org/10.1007/s10649-024-10317-5

Ministry of National Education of Türkiye Republic. (2018). *Central examination for secondary education institutions that will accept students by test: Numerical part.* Retrieved from https://odsgm.meb.gov.tr/meb_iys_dosyalar/2018_06/03153730_SAYISAL_BYLYM_A_kitapYY.pdf

Mitchell, J. E., Crosby, R. D., Wonderlich, S., & Adson, D. E. (2000). *Elements of clinical research in psychiatry*. American Psychiatric.

Nwadinigwe, P. I., & Naibi, L. (2013). The number of options in a multiple-choice test ıtem and the psychometric characteristics. *Journal of Education and Practice*, *4*(28), 189–196. Retrieved from https://www.iiste.org/Journals/index.php/JEP/article/view/9944

Öksüz, Y., & Güven Demir, E. (2019). Comparison of open ended questions and multiple choice tests in terms of psychometric features and student performance. *Hacettepe University Journal of Education*, *34*(1), 259–282. https://doi.org/10.16986/HUJE.2018040550

Özer Özkan, Y., & Özaslan, N. (2018). Student achievement in Turkey, according to question types used in PISA 2003-2012 mathematic literacy tests. *International Journal of Evaluation and Research in Education (IJERE)*, *7*(1), 57–64. https://pdfs.semanticscholar.org/7e84/37899e70c78fbe2dde7ab179ccca7eb6a0a0.pdf

Paler-Calmorin, L., & Calmorin, M. A. (2007). *Research methods and thesis writing* (2nd ed.). Rex Book Store.

Pallant, J. (2005). *SPSS survival manual: A step by step guide to data analysis using SPSS for Windows (Version 12).* Allen & Unwin.

Raymond, M. R., Stevens, C., & Bucak, S. D. (2019). The optimal number of options for multiple-choice questions on high-stakes tests: Application of a revised index for detecting nonfunctional distractors. *Advances in Health Science Education*, *24*, 141–150. https://doi.org/10.1007/s10459-018-9855-9

Rodriguez, M. C. (2002). Choosing an item format. In G. Tindal & T. M. Haladyna (Eds.), *Large-scale assessment programs for all students: Validity, technical adequacy, and implementation* (pp. 213–231). Lawrence Erlbaum Associates.

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. https://doi.org/10.1111/j.1745-3992.2005.00006.x

Salkind, N. J. (2010). *Statistics for people who (think they) hate statistics* (3rd ed.). Sage.

Schaefer, J. M. L. (2009). *The effects of stem completeness and stem orientation on multiple-choice item difficulty and discrimination* [Unpublished Master Theses, California State University]. Retrieved from https://hdl.handle.net/10211.9/162

Shin, J., Bulut, O., & Gierl, M. J. (2019). The effect of the most-attractive-distractor location on multiple-choice item difficulty. *The Journal of Experimental Education*, *88*(4), 643–659. https://doi.org/10.1080/00220973.2019.1629577

Şahin, M. D. (2022). Exploratory factor analysis [Açımlayıcı faktör analizi]. In S. Göçer Şahin & M. Buluş (Eds.), *Applied statistics step by step* [*Adım adım uygulamalı istatistik]* (pp. 309–342) Pegem Academy.

Tabachnick, B. G., & Fidell, L. S. (2013). *Using multivariate statistics* (6th ed.). Pearson.

Temizkan, M., & Sallabaş, M. E. (2015). Comparison of multiple choice tests and open-ended questions in the assessment of reading comprehension skills [Okuduğunu anlama becerisinin değerlendirilmesinde çoktan seçmeli testlerle açık uçlu yazılı yoklamaların karşılaştırılması]. *Dumlupınar University Journal of Socıal Sciences*, *30*, 207–220.

Terranova, C. (1969). *The effects of negative stems in multiple-choice test items*. Unpublished doctoral dissertation, State University of New York at Buffalo. (30, 2390A).

Vegada, B., Shukla, A., Khilnani, A., Charan, J., & Desai, C. (2016). Comparison between three option, four option and five option multiple choice question tests for quality parameters: A randomized study. *Indian Journal of Pharmacology*, *48*(5), 571–575. https://doi.org/10.4103/0253-7613.190757

Violato, C. (1991). Item difficulty and discrimination as a function of stem completeness. *Psychological Reports*, *69*(3), 739–743. https://doi.org/10.2466/pr0.1991.69.3.739

Violato, C., & Harasym, P. H. (1987). Effects of structural characteristics of stem format of multiple-choice items on item difficulty and discrimination. *Psychological Reports*, *60*(3_part_2), 1259–1262. https://doi.org/10.1177/0033294187060003-251.1

Violato, C., & Marini, A. E. (1989). Effects of stem orientation and completeness of multiple-choice items on item difficulty and discrimination. *Educational and Psychological Measurement*, *49*(1), 287–295. https://doi.org/10.1177/0013164489491032

Yılmaz Koğar, E., & Soysal, S. (2023). Examination of response time effort in TIMSS 2019: Comparison of Singapore and Türkiye. *International Journal of Assessment Tools in Education*, *10*(Special Issue), 174–193. https://doi.org/10.21449/ijate.1343248

Yılmaz, S. (2007). *Misconceptions of second-degree primary school's students about problem solving* (Thesis Number. 200688). [Master Thesis, Eskişehir Osmangazi University], Eskişehir.

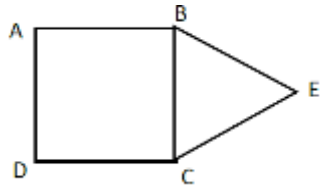
Zhang, F., & Lidbury, B. A. (2013). Evaluating a genetics concepts inventory. In F. Zhang (Eds.), *Sustainable language support practices in science education: Technologies and solutions* (pp. 116–128). Medical Information Scince Reference.

Zulaiha, R., Dian Rahdiani, F., Rahman, A., & Al Anfal, M. F. (2021). Analysis of difficulty level and discriminating power between multiple choices and essay items on math test. *Advances in Social Science, Education and Humanities Research*, *545*, 62–68. https://doi.org/10.2991/assehr.k.210423.065

Appendix

Figure 1

An example of geometry items with and without shapes

The item with shape	The item without shape
 If ABCD is a square and BEC is an equilateral triangle, find the angle of m($\widehat{DCE}$).	ABCD is a square and BEC is an equilateral triangle. If the square ABCD and equilateral triangle BEC have side [BC] in common, find the angle of m($\widehat{DCE}$).
A) 120° B) 130° C) 140° D) 150°	A) 120° B) 130° C) 140° D) 150°