ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
Research Article; 2024; 15(1); 65-78

Investigating the Effect of Testlets Consisting of Open-Ended and Multiple-Choice Items on Reliability via Generalizability Theory*

Serpil KOCAOĞLU ** Melek Gülşah ŞAHİN ***

Abstract

This study aimed to reveal the effect on reliability of testlets consisting of open-ended and multiple-choice items with similar content. For this purpose, two different mathematics achievement tests, one with multiple-choice items and the other with open-ended items, were administered to 128 8th-grade students. Reliability estimations on the obtained data were conducted in the Edu-G program based on the Generalizability Theory. A decision study was also performed. In the achievement test with testlets consisting of open-ended items, p×i×r (p: person, i: item, r: rater) fully crossed design was used when testlet effect was not considered; p×(i:t)×r (t: testlet) nested design was used when testlet effect was considered. According to the results, the reliability coefficient was estimated higher when the testlet effect was not considered. Similarly, in the achievement test with testlets consisting of multiple-choice items, the p×i crossed design was used when the testlet effect was not considered, and the p×(i:t) nested design was used when the testlet effect was considered. According to the results, the reliability coefficient was estimated higher when the testlet effect was not considered. According to the data obtained within the scope of the study, the reliability coefficient was estimated higher in the test with open-ended items than in the test with multiple-choice items. When the testlet effect was included, the change in the reliability coefficient in the test with open-ended items was higher than the change in the test with multiple-choice items. In the decision studies, it was observed that the increase in the number of items and testlets positively affected reliability, but the increase in testlets contributed to reliability more. In the tests consisting of open-ended items, it was observed that the increase in the number of raters contributed to reliability less than items and testlets.

Keywords: Open-ended items, multiple-choice items, testlet, generalizability theory, reliability

Introduction

Assessments are carried out for different purposes at every stage of the education processes. Recognizing students, identifying and eliminating students' learning deficiencies, organizing learning experiences, determining students' learning levels, organizing the learning environment, etc., can be stated within these purposes. In the assessment process, firstly, the appropriate measurement tool should be selected, and the measurement process should be planned. When measurement tools are examined in general, it can be stated that they have different characteristics. Multiple-choice items, the most effective and useful way of measuring knowledge (Haladyna, 2004), are frequently preferred in national and international examinations. Multiple-choice tests are objective tests in terms of scoring. The data obtained by applying multiple-choice tests to a large number of groups can be evaluated in a short time. Scoring is also easy and takes a short time. It can be prepared at all levels of the cognitive taxonomy (Downing, 2006). However, measuring high-level cognitive skills in assessments conducted with multiple-choice tests is difficult (Ko, 2010). Open-ended items can especially be used in the measurement of high-level skills. Open-ended items have three important advantages over multiple-choice items (Bridgeman, 1992). With open-ended items, there is no possibility of finding the correct

*This study has been produced from Master's Thesis that was conducted under the supervision of the Assoc. Prof. Dr. Melek Gülşah ŞAHİN and prepared by Serpil KOCAOĞLU.

** Teacher, Republic of Türkiye Ministry of National Education, Ankara-Türkiye, zncl10@gmail.com, ORCID ID: 0009-0008-0566-4371

***Assoc. Prof. Dr., Gazi University, Gazi Faculty of Education, Ankara-Türkiye, mgulsahsahin@gazi.edu.tr, ORCID ID: 0000-0001-5139-9777

To cite this article:
Kocaoğlu, S. & Şahin, M.G. (2024). Investigating the effect of testlets consisting of open-ended and multiple-choice items on reliability via generalizability theory. *Journal of Measurement and Evaluation in Education and Psychology, 15*(1), 65-78. https://doi.org/10.21031/epod.1429423

Received: 1.01.2024
Accepted: 25.03.2024

answer by guessing. In addition, although it is not intended as feedback, individuals may realize that they have made a mistake when they cannot find what they think as the answer in the options of multiple-choice items; however, such feedback is not possible in open-ended items. Thirdly, it is impossible to reach the correct answer in open-ended items by eliminating the options as in multiple-choice items.

In addition, in multiple-choice items, the individual can find the answer with less effort compared to open-ended items (Attali et al, 2016). While students are sometimes expected to give a short and strictly defined answer to the items, sometimes the student can be left free in terms of the quality and length of the answer. When considered in this context, open-ended items are categorized under two headings: restricted and unrestricted response (Berberoğlu, 2006). In restricted response questions, the answers are mostly short since some limitations are imposed on the quality or length of the answer. On the other hand, in free-response questions, since the respondent is given a certain amount of freedom regarding the quality or length of the answer, these questions are long-answer questions (Doğan, 2009). In addition to the multiple-choice and open-ended items mentioned here, item types such as short-answer, fill-in-the-blank, true-false, and matching are frequently used in the literature (Doğan, 2019a; 2019b; Karatoprak Erşen & Gündüz, 2023; Nitko & Brookhart, 2014; Popham, 2014; Russell & Airasian, 2008).

In the selection of the item type, the purpose of the test, the feature to be measured, the group to be measured, the application conditions, etc., should be taken into consideration. Another issue that should also be considered is how the items will be presented. When the use of items in both national and international exams is examined, it is seen that items are presented independently or in testlets. The concept of testlets was first introduced by Wainer and Kiely (1987) to refer to a group of items with a common stimulus. In their study, they used testlets in Computerized Adaptive Testing (CAT) to balance the content and eliminate the importance of context effect and item order. Testlets have been widely used especially in recent years. Item writers prefer to prepare items based on a common material because it saves time and energy (Wainer et al. 2000). Moreover, it has been observed that following consecutive items based on a common root is more successful (Lee et al, 2000). The fact that it makes it possible to measure high-level cognitive skills is effective in making testlets popular. In a testlet, a common material such as a graph, table, figure, or map is used to answer two or more items. Some rules, such as the comprehensibility of the material, the ability to respond to the items correctly based only on the material, and careful determination of the number of items, should be considered in developing testlets (Tekin, 2009). Although there is no definite rule in determining the number of items that should be included in testlets (Kaya Uyanık & Ertuna, 2022), there may be between 2-12 items (Yaman, 2016) depending on the characteristics of the common material. In determining the number of items in testlets, the characteristics of the structure shared by the items and content validity can also be taken into consideration.

The limitation is that the items in testlets have local dependency on each other. Therefore, the testlet effect should be considered in estimating the test reliability in which testlets are included. When the items in the testlets are considered independently, the reliability value can be estimated to be higher than when the testlet effect is taken into account. (Lee et al, 2000; Sireci et al, 1991; Taşdelen Teker, 2014; Wainer & Thissen, 1996).

This study aimed to estimate the reliability of testlets consisting of open-ended items and multiple-choice items prepared in similar content within the framework of generalizability theory in cases where the testlet effect was taken and not considered. In recent years, it is seen that the frequency of testlets has increased both in Türkiye’s national exams such as the Academic Personnel and Postgraduate Education Entrance Exam (ALES), Foreign Language Exam (YDS), Higher Education Institutions Foreign Language Exam (YÖKDİL) and in international exams such as Test of English as a Foreign Language (TOEFL) and Program for International Student Assessment (PISA). Although using testlets consisting of multiple-choice items is preferred chiefly due to the scoring advantage, the use of testlets consisting of open-ended items cannot be ignored. In this context, the reliability estimations of testlets with similar content prepared with different item types will shed light on the researchers who would like to work on this subject. The study used the Generalizability (G) theory to determine the reliability estimation. According to Shavelson and Webb (1991), G theory is a statistical theory that gives an idea about the reliability of behavioral measurements. As an extension of both Classical Test Theory and analysis of variance, G theory is a mathematical model in which multiple sources of error can be addressed. The advantage of G-theory is that different error sources can be estimated simultaneously with a single analysis. In other words, unlike Classical Test Theory, it considers the results of different error sources separately and in a single interaction. G theory also allows us to estimate the reliability of scores for different interpretations. While in Classical Test Theory, only relative assessments are made in which individuals are compared with each other, in G theory, it is also possible to make absolute assessments in which only the performance of individuals is evaluated independently of each other. In other words, while Classical Test Theory provides researchers with information to make only relative decisions, G theory offers sufficient information for both relative and absolute decisions at the same time (Brennan, 2001; Güler et al, 2012; Shavelson & Webb, 1991). Within the scope of the study, reliability estimation was performed using the Generalizability approach, and a decision study was conducted.

In the literature, there are studies in which testlets are handled with Item Response Theory (Sireci et al, 1991; Wainer & Thissen, 1996) and G theory (Lee & Frisbie, 1999; Lee et al, 2000; Taşdelen Teker, 2014; Kaya Uyanık & Gelbal, 2018; Kaya Uyanık & Ertuna, 2022). In addition, while Gessaroli and Folske (2002) addressed testlets with factor analysis, Kaya Uyanık and Gelbal (2018) studied two-facet patterns with the Generalizability approach in Item Response modeling using testlet data generated in simulation in their study and compared the results obtained with the results from G theory. Although it was seen that the studies conducted in the international arena occupied a larger space, it was seen that the studies conducted in the national arena were limited. In addition, there is no study in the literature examining the reliability of testlets consisting of both open-ended items and multiple-choice items with similar content within the framework of G theory. From this point of view, it is thought that this study will also contribute to the literature. In the study, a decision (D) study was conducted on the effect of the number of items and the number of testlets on reliability estimation in testlets consisting of open-ended and multiple-choice items. In addition, a D study was also conducted for the change in the number of raters for the test composed of open-ended items where more than one rater was involved. For this reason, the study is considered to be vital as it will provide a different suggestion to the users in exams where testlets are frequently used. The research problems formed in line with the purpose of the study were determined as follows:

1. In achievement tests with testlets consisting of open-ended items;
    a. What are the variance components and G and Phi coefficients for the p×i×r fully crossed design in which person (p), item (i), and raters (r) are crossed with each other when the testlet effect is not considered?
    b. What are the G and Phi coefficients for the decision studies on increasing or decreasing the number of items and raters in the p×i×r design in which the testlet effect is not taken into account?
    c. What are the variance components and G and Phi coefficients of the p×(i:t)×r design in which items (i) are nested in testlets (t) and individuals (p) and raters (r) are crossed with them?
    d. What are the G and Phi coefficients of the p×(i:t)×r design in which the testlet effect is taken into account in the decision studies for increasing or decreasing the number of testlets, items in the testlets, and the raters?
2. In achievement tests with testlets consisting of multiple-choice items;
    a. What are the variance components and G and Phi coefficients of the p×i design in which persons (p) are crossed with items (i) when the testlet effect is not considered?
    b. What are the G and Phi coefficients in the decision study for increasing or decreasing the number of items in the p×i design where the testlet effect is not considered?
    c. What are the variance components and G and Phi coefficients of the p×(i:t) partial nested design in which items (i) are nested in testlets (t) and persons (p) are crossed with them?
    d. What are the G and Phi coefficients for decision studies with increasing and decreasing the number of testlets and the number of items within a testlet in the p×(i:t) design where the testlet effect is taken into account?

Method

This study aimed to obtain and compare G and Phi coefficients within the framework of Generalizability theory in achievement tests with testlets consisting of multiple-choice and open-ended items in cases where the testlet effect was and was not taken into account. The research is basic research since it aims to obtain new information by testing the existing theory in different situations (Karasar, 1994).

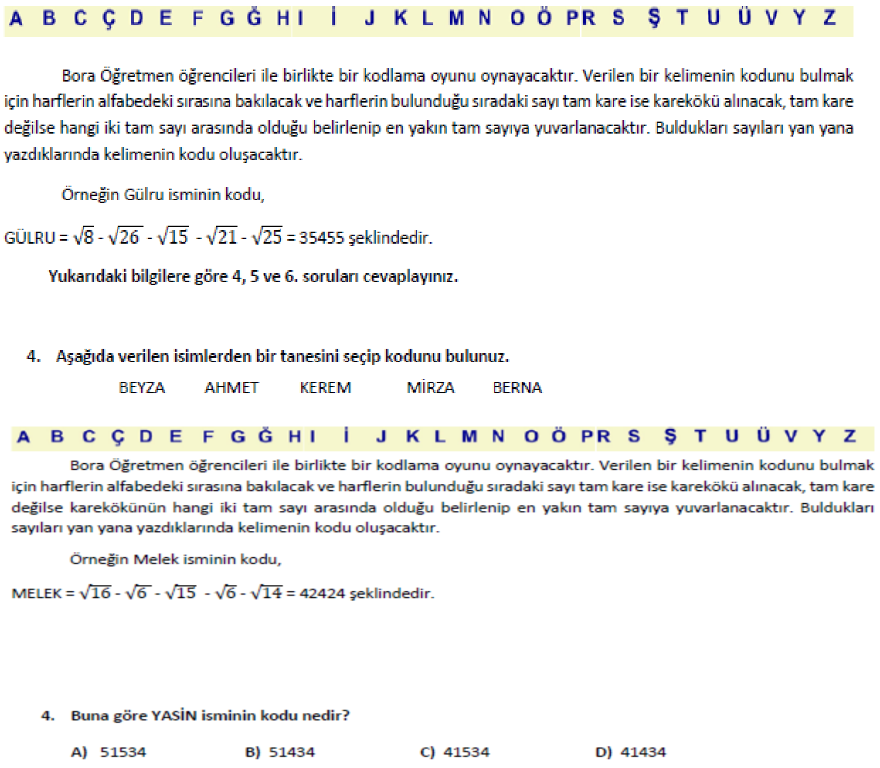
Participants

The study group of the research consists of 8th-grade students studying in public elementary schools affiliated with the Ministry of National Education in Ankara province in the 2022-2023 academic year. Ethical approval of the research was obtained from the Gazi University Ethics Commission. The pilot study was conducted in 3 elementary schools in Ankara and Beypazarı district. These schools were selected because of the large number of students they have. Since the number of students in Beypazarı district was insufficient, the final implementation was carried out in a public elementary school in the central district of Ankara. While determining this school, it was taken into consideration that it should be similar to the achievement levels of the schools in the pilot study. In the pilot study, 119 students solved the mathematics achievement test consisting of open-ended items, and 115 students solved the mathematics achievement test consisting of multiple-choice items. Of the 119 students who solved the achievement test consisting of open-ended items, 31 students were not included in the analysis because they either left all the items blank or scored zero points. As a result, 88 students' responses to open-ended items and 115 students' responses to multiple-choice items were analyzed in the pilot application. The final application was carried out with the participation of 157 students. Since it was observed that 29 of these students either did not answer the achievement test consisting of open-ended items at all or answered incorrectly and received zero points, the results of these students were not included. In the final application, the responses of 128 students to the achievement tests consisting of both open-ended and multiple-choice items were evaluated.

Data Collection Tools

The data required for this study were collected through two separate mathematics achievement tests with testlets consisting of open-ended and multiple-choice items. First, two achievement tests with similar content, one with open-ended items and the other with multiple-choice items, were prepared for pilot testing. Each test included four testlets and four items in each testlet. The items were prepared in line with the achievements of "exponential expressions, square root expressions, data analysis, and probability of simple events" from the 8th-grade mathematics curriculum of the 2022-2023 academic year. In the assessment of testlets consisting of open-ended items, rubric and evaluation form were prepared for each item. Accordingly, the grade was calculated by 3 points for the entirely correct answers and 2 points and 1 point for the partially correct answers. Blank and other answers were evaluated as 0 points. An example of open-ended and multiple-choice items from the same content is given in Figure 1.

Figure 1
Example of Open-Ended and Multiple-Choice Test Items



A B C Ç D E F G Ğ H I İ J K L M N O Ö P R S Ş T U Ü V Y Z

Bora Öğretmen öğrencileri ile birlikte bir kodlama oyunu oynayacaktır. Verilen bir kelimenin kodunu bulmak için harflerin alfabedeki sırasına bakılacak ve harflerin bulunduğu sıradaki sayı tam kare ise karekökü alınacak, tam kare değilse hangi iki tam sayı arasında olduğu belirlenip en yakın tam sayıya yuvarlanacaktır. Buldukları sayıları yan yana yazdıklarında kelimenin kodu oluşacaktır.

Örneğin Gülru isminin kodu,

GÜLRU = $\sqrt{8}$ - $\sqrt{26}$ - $\sqrt{15}$ - $\sqrt{21}$ - $\sqrt{25}$ = 35455 şeklindedir.

Yukarıdaki bilgilere göre 4, 5 ve 6. soruları cevaplayınız.

4. Aşağıda verilen isimlerden bir tanesini seçip kodunu bulunuz.

BEYZA AHMET KEREM MİRZA BERNA

A B C Ç D E F G Ğ H I İ J K L M N O Ö P R S Ş T U Ü V Y Z

Bora Öğretmen öğrencileri ile birlikte bir kodlama oyunu oynayacaktır. Verilen bir kelimenin kodunu bulmak için harflerin alfabedeki sırasına bakılacak ve harflerin bulunduğu sıradaki sayı tam kare ise karekökü alınacak, tam kare değilse karekökünün hangi iki tam sayı arasında olduğu belirlenip en yakın tam sayıya yuvarlanacaktır. Buldukları sayıları yan yana yazdıklarında kelimenin kodu oluşacaktır.

Örneğin Melek isminin kodu,

MELEK = $\sqrt{16}$ - $\sqrt{6}$ - $\sqrt{15}$ - $\sqrt{6}$ - $\sqrt{14}$ = 42424 şeklindedir.

4. Buna göre YASİN isminin kodu nedir?

A) 51534 B) 51434 C) 41534 D) 41434

Seven expert opinions were obtained for the pilot forms of the developed open-ended and multiple-choice tests. In this context, three experts were experts in both mathematics and measurement and evaluation areas, one of whom was a faculty member, and the other two were teachers at the graduate level. In addition, two academicians in the area of measurement and evaluation and one undergraduate mathematics teacher were also consulted. In line with expert opinions, revisions were made to the items based on content, form and item writing rules. After the revisions made by considering the expert opinions, Turkish language expert opinion was also taken. The experts chose one of the appropriate options from the expressions "appropriate" or "not appropriate" for each item while expressing their opinions. If the experts chose the same option for the same item, it was considered agreement, and if they chose different options, it was considered disagreement. In this study, inter-expert agreement was calculated to ensure validity and reliability. For this purpose, Miles and Huberman's (1994) reliability formula was used to determine the percentage of agreement between experts. According to the formula, the percentage of agreement is expressed as "Reliability = (Agreement / Agreement + Disagreement) * 100". Accordingly, the percentages of inter-expert agreement were calculated using Miles and Huberman's (1994) formula, and the average percentage of inter-expert agreement was found to be 85%. In order for the research to be considered reliable, the reliability estimates must be above 70% (Miles & Huberman, 1994). Therefore, the result obtained in this study indicates that inter-rater agreement was achieved. Experts were also asked to give their opinions on whether the open-ended and multiple-choice items had similar content. The experts expressed their opinions as "similar content", "partially similar content" and "not similar content". None of the experts chose the "not similar content" option. The percentage of agreement of the expert opinions on the similarity of the content was calculated, and the lowest was 75% and the highest was 100% and the average was 89%. Thus, the pilot application of the tests, which were decided to be appropriate in terms of content, language and expression, was started.

The pilot study aimed to determine the item difficulty and discrimination indices of the open-ended and multiple-choice test items. The open-ended items were scored independently by three mathematics teachers. The raters were given a brief explanation about the measured feature, item content, and rubric before scoring. The item statistics of the data obtained within the scope of the pilot application are shown in Table 1.

Table 1

Item Statistics of Tests in Pilot Study

		Item statistics for multiple-choice items		Item statistics for open-ended items	
Testlet	Item No	p	r	p	r
1	1	0,50	0,47	0,36	0,32
	2	0,46	0,46	0,33	0,71
	3	0,43	0,38	0,32	0,62
	4	0,33	0,24	0,17	0,40
2	5	0,58	0,62	0,47	0,84
	6	0,63	0,52	0,34	0,67
	7	0,37	0,57	0,09	0,22
	8	0,34	0,37	0,01	0,05
3	9	0,42	0,43	0,52	0,60
	10	0,18	0,19	0,03	0,10
	11	0,15	0,14	0	0
	12	0,33	0,45	0	0
4	13	0,48	0,48	0,14	0,41
	14	0,42	0,60	0,06	022
	15	0,36	0,55	0,08	0,27
	16	0,29	0,53	0,08	0,21

When deciding on the items to be used in the final application, the discrimination (r) of the multiple-choice items was taken into consideration and a selection was made accordingly. Generally, items with an item discrimination between 0.20 and 0.30 are considered usable in the test; items with an item discrimination between 0.30 and 0.40 are considered good; and items with an item discrimination higher than 0.40 are considered very good. It is recommended that items with discrimination lower than 0.20 should be corrected and improved (Özçelik, 2013). Since balanced designs were examined in this study, one item from the testlet was removed. Items with low discrimination (4, 8, 11 and 16) were removed from the test, and item 10 was corrected and included in the test. Since items 11 and 12 of the open-ended items were not responded by any student, the item statistics were zero. It is thought that this situation is caused by the fact that the responses to these items were prepared as free responses. Since the results differed, it was decided to take the opinions of two faculty members in the field of measurement and evaluation and use expert opinions that the items had similar content instead of comparing the item statistics one-to-one.

In line with this purpose, it was decided which items should be removed or revised, and the final achievement test forms were created with four testlets consisting of 3 items each.

Implementation Process

Since the achievement tests included the 8th-grade first-semester subjects, the pilot application was carried out after these subjects were covered. After the pilot application, the necessary analyses were made, and the final tests were created.

For the final application, one group was first administered an achievement test consisting of open-ended items, while the other group was administered an achievement test consisting of multiple-choice items. In this way, it was aimed to prevent an effect caused by the order of administration of tests consisting of different item formats. A ten-day break was given for the administration of the second test. After ten days, the test consisting of multiple-choice items was administered to the group to which the test consisting of open-ended items was administered first, and the test consisting of open-ended items was administered to the group to which the test consisting of multiple-choice items was administered.

Data Analyses

In order to obtain more reliable results in the analyses, three mathematics teachers served as raters. The mathematics teachers conducted their scoring independently. G and Phi coefficients were calculated in the tests consisting of open-ended and multiple-choice testlets within the framework of G theory. In addition, the D study was conducted by selecting the appropriate variables from the number of items, testlets, and raters within each design. EduG program was used in data analysis.

Results

Results Related to Sub-Problem 1

1.a. In the study, the responses of 128 8th-grade students to an achievement test consisting of 12 open-ended items were analyzed within the framework of G theory without considering the testlet effect. In this context, the results of the G study belonging to the p x i x r design in which person (p), item (i) and raters (r) were crossed with each other are given in Table 2.

Table 2

G Study Results for the p×i×r Design in which the Testlet Effect is not Handled in the Achievement Test with Open-Ended Items

Variance Source	Sum of Squares	Degree of Freedom	Mean Squares	Variance Value	Variance Proportion
p	1352,284	127	10,648	0,238	19,3
r	7,461	2	3,731	0,000	0,0
i	551,671	11	50,152	0,115	9,3
pr	91,816	254	0,361	0,007	0,6
pi	2773,607	1397	1,985	0,571	46,2
ri	92,247	22	4,193	0,031	2,5
pri,e	763,809	2794	0,273	0,273	22,1
Total	5632,895	4607			100%
G coefficient	0,81				
Phi coefficient	0,78				

When Table 2 is analyzed, the variance belonging to persons (p) explains 19.3% of the total variance and indicates the extent to which persons differ from each other. The value of this variance component is expected to be quite high. It is seen that most of the variability is explained by other sources of variability. The rater (r) variance component (0.000) indicates excellent consistency between the raters' ratings. The item (i) variance component accounts for 9.3% of the total variance and suggests that the difficulty levels of the items differ. The value obtained for the variance component of the person x rater (pr) interaction is 0.007, explaining 0.6% of the total variance. This value means that the scores given by the raters to the persons did not differ much between the raters. In other words, the raters gave similar scores to the persons, which is a desirable situation. When the person × item (pi) interaction variance component is analyzed, it accounts for 46.2% of the total variance and has the highest variance value. This value indicates that the difficulty levels of the items differ from one person to another. The value calculated for the variance component of the rater x item (ri) interaction is 0.031, explaining 2.5% of the total variance. This value indicates the variability of the scores given by the raters to the items. It is desirable that this value is low. The residual component accounts for 22.1% of the total variance, pointing that there are systematic or non-systematic error sources in this study with the interaction between persons, items and raters. Finally, in the analyses obtained as a result of the G theory study, it is seen that the G and Phi coefficient is calculated as 0.81 and 0.78, respectively. Since these values are well above 0.70, they are acceptable values.

1.b. The results of the D study for the p×i×r design in which the testlet effect was not considered in the achievement test consisting of open-ended items are shown in Table 3.

Table 3

D Study Results for the p×i×r Design in which the Testlet Effect is Not Handled in an Achievement Test Consisting of Open-Ended Items

Condition-1	Number of Items	G	Phi	Condition-2	Number of Raters	G	Phi
Number of persons: 128 Number of Raters: 3	6	0,68	0,64	Number of persons: 128 Number of Items: 12	2	0,79	0,76
	9	0,76	0,73		**3***	**0,81**	**0,78**
	12*	**0,81**	**0,78**		4	0,81	0,78
	15	0,84	0,81		5	0,82	0,79
	18	0,86	0,84		6	0,82	0,79
	21	0,88	0,86		7	0,82	0,79

* Refers to the case study data of the current study.

Table 3 shows that when the number of persons and raters is kept constant and the number of items was increased, the coefficients of G and Phi were also increased. When the number of persons and items were kept constant and the number of raters was increased, the reliability coefficients were increased, but they were not affected as much as the increase in the number of items.

1.c. The results of the G study for the p×(i:t)×r design in which the testlet effect was taken into account in the achievement test consisting of open-ended items are shown in Table 4.

Table 4

G Study Results for the p×(i:t)×r Design Considering the Testlet Effect in an Achievement Test Consisting of Open-Ended Items

Variance Source	Sum of Squares	Degree of Freedom	Mean Squares	Variance Value	Variance Proportion
p	1352,284	127	10,648	0,199	16,0
r	7,461	2	3,731	-0,001	0,0
t	145,362	3	48,454	-0,005	0,0
i:t	406,309	8	50,789	0,119	9,6
pr	91,816	254	0,361	0,003	0,3
pt	1318,138	381	3,460	0,218	17,6
pi:t	1455,469	1016	1,433	0,392	31,6
rt	30,228	6	5,038	0,003	0,2
ri:t	62,019	16	3,876	0,028	2,3
prt	243,605	762	0,320	0,021	1,7
prt:i,e	520,203	2032	0,256	0,256	20,6
Total	5632,895	4607			100%
G Coefficient	0,67				
Phi Coefficient	0,65				

When Table 4 is examined, it is seen that the calculated value of the i:t variance component in which the items are nested in the testlets is 0.119, accounting for 9.6% of the total variance. This value indicates that there is a slight difference between the difficulty levels of the items in the testlets. The variance component of the person-testlet accounts for 17.6% of the total variance and indicates that there are differences due to the person-testlet interaction. The variance component of the person x item interaction within the testlet (pi:t) has the largest value (0.392). This value alone accounts for 31.6% of the total variance and indicates that the person-item interaction varies within the testlet. The G coefficient was 0.67 and the Phi coefficient was 0.65, and it is seen that the G and Phi coefficients are lower than when the testlet effect is not handled.

1.d. The results of the D study for the p×(i:t)×r design in which the testlet effect was taken into account in the achievement test consisting of open-ended items are given in Table 5.

Table 5

D Study Results for the p×(i:t)×r Design Handling the Testlet Effect in an Achievement Test Consisting of Open-Ended Items

Condition-1	Number of Teslets	G	Phi	Condition-2	Number of Items in a Testlet	G	Phi	Condition-3	Number of Raters	G	Phi
Number of persons: 128 Number of raters: 3 Number of items in the testlet: 3	3	0,61	0,58	Number of persons: 128 Number of raters: 3 Number of testlets: 4	1	0,53	0,49	Number of persons: 128 Number of testlets: 4 Number of items in the testlet: 3	2	0,66	0,64
	4*	**0,67**	**0,65**		2	0,63	0,60		**3***	**0,67**	**0,65**
	5	0,72	0,70		**3***	**0,67**	**0,65**		4	0,68	0,65
	6	0,75	0,73		4	0,69	0,67		5	0,68	0,66
	7	0,78	0,76		5	0,71	0,69		6	0,68	0,66
	8	0,80	0,78		6	0,72	0,71		7	0,68	0,66

* Refers to the case study data of the current study.

Table 5 shows that the coefficients of G and Phi increase when the number of testlets, the number of raters and the number of items in the testlets increase, respectively. It is observed that the G coefficient increases above 0.70 with four testlets consisting of five items each (total 20 items) when the number of items in the testlets increases, while it increases above 0.70 with five testlets consisting of three items each (total 15 items) when the number of testlets increases. It can be stated that the increase in the number of raters did not affect the reliability coefficients to a great extent.

Results Related to Sub-Problem 2

2.a. The results of the G study for the *p×i* design in which the testlet effect was not handled in the achievement test consisting of multiple-choice items are shown in Table 6.

Table 6

G Study Results for the p×i Design in Achievement Test Consisting of Multiple-Choice Items in which the Testlet Effect is not Handled

Variance Source	Sum of Squares	Degree of Freedom	Mean Squares	Variance Value	Variance Proportion
p	100,093	127	0,788	0,050	20,2
i	19,898	11	1,809	0,013	5,1
pi,e	258,852	1397	0,185	0,185	74,7
Total	378,843	1535			100%
G Coefficient	0,76				
Phi Coefficient	0,75				

Table 6 demonstrates that the estimated variance component for the individuals accounts for 20.2% of the total variance, with a value of 0.050. The fact that this variance component belonging to persons is quite high indicates that there is a systematic difference between persons and this is an expected situation. With a value of 0.013, the item variance component's calculated value accounts for 5.1% of the total variance. Here, it is possible to say that item difficulties do not differ much. The highest variance value belongs to the residual component with 0.185. This value accounts for 74.7% of the total variance. This is an indication that the person-item interaction and systematic or non-systematic error sources that could not be measured in this study were not controlled. This variance belonging to the residual component is expected to be quite low. The G coefficient was 0.76 and the Phi coefficient was 0.75, and it can be stated that the reliability coefficients are at an adequate level.

2.b. The results of the D study for the *p*×*i* design in which the testlet effect was not handled in the achievement test consisting of multiple-choice items are shown in Table 7.

Table 7

D Study Results for the p×i Design in which the Testlet Effect is Not Handled in an Achievement Test Consisting of Multiple-Choice Items

Condition-1	Number of Items	G	Phi
Number of persons: 128	6	0,62	0,60
	9	0,71	0,70
	12*	0,76*	0,75*
	15	0,80	0,79
	18	0,83	0,82
	21	0,85	0,84

* Refers to the case study data of the current study.

When Table 7 is examined, it is observed that in achievement tests consisting of multiple-choice items, the G and Phi reliability coefficients obtained in the p×i design in which persons (*p*) are crossed with items (*i*) increase as the number of items increases. When the number of items increased from 6 to 15, the G coefficient increased from 0.62 to 0.80.

2.c. The results of the G study for the p×(i:t) design in which the testlet effect is taken into account in the achievement test consisting of multiple-choice items are presented in Table 8.

Table 8

G Study Results for the p×(i:t) Design Handling the Testlet Effect in an Achievement Test Consisting of Multiple-Choice Items

Variance Source	Sum of Squares	Degree of Freedom	Mean Squares	Variance Value	Variance Proportion
p	100,093	127	0,788	0,047	18,8
t	7,929	3	2,643	0,003	1,1
i:t	11,969	8	1,496	0,010	4,2
pt	86,154	381	0,226	0,019	7,5
pi:t,e	172,698	1016	0,170	0,170	68,3
Total	378,843	1535			100%
G Coefficient	0,71				
Phi Coefficient	0,70				

When Table 8 is analyzed, the variance belonging to the main effect of persons (*p*) accounts for 18.8% of the total variance. This value is the second largest percentage of variance in the table 8, indicating that there is a systematic difference between persons. With a value of 0.003, the variance estimated for the testlet (t) main effect explains 1.1% of the overall variance. This value means that the difficulty levels of the testlets do not differ from each other. With a value of 0.010, the variance component of the i:t effect, which involves items nested within testlets, accounts for 4.2% of the total variance. The fact that this value is close to zero indicates that the difficulty levels of the items in the same testlet are close to each other. The estimated variance of person testlet (pt) accounts for 7.5% of the total variance. Here, it can be stated that there are differences due to person-testlet interaction. In this design, the largest variance value belongs to the residual component (pi:t,e), accounting for 68.3% of the total variance. This value is lower than the accounted percentage of the residual component (74.7%) obtained when the testlet effect is not handled. This value shows that since there are more variance sources when the testlet effect is handled, the percentage of accounted total variance is divided into these variance sources and the percentage of variance belonging to the residual component decreases. When the reliability coefficients are analyzed, it is seen that the G coefficient is 0.71 and the Phi coefficient is 0.70.

2.d. The results of the D study for the p×(i:t) design in which the testlet effect is taken into account in the achievement test consisting of multiple-choice items are given in Table 9.

Table 9

D Study Results for the p×(i:t) Design Handling the Testlet Effect in an Achievement Test Consisting of Multiple-Choice Items

Condition-1	Number of Testlets	G	Phi	Condition-2	Number of Items in a Testlet	G	Phi
.Number of persons: 128 Number of items in a testlet: 3	2	0,55	0,53	Number of persons: 128 Number of Testlets: 4	2	0,64	0,63
	3	0,65	0,63		**3***	**0,71**	**0,70**
	4*	**0,71**	**0,70**		4	0,75	0,74
	5	0,76	0,74		5	0,78	0,76
	6	0,79	0,77		6	0,80	0,78
	7	0,81	0,80		7	0,81	0,80

* Refers to the case study data of the current study.

When Table 9 is analyzed, it is seen that the G and Phi coefficients increase when the number of persons and items in the testlets are kept constant and the number of testlets is increased. In the second case, the reliability coefficients increased when the number of individuals and testlets were kept constant and the number of items in the testlet was increased. Reliability coefficients are found to be more impacted by an increase in testlet count than by an increase in testlet item count. This result is similar to the results obtained with testlets consisting of open-ended items. Obtaining the same G coefficient with more items indicates that the increase in testlets is more effective.

Conclusion and Discussion

When the findings of the achievement tests consisting of open-ended items were analyzed, the reliability coefficients that were estimated differed when the testlet effect was not taken into account and when it was taken into account. This result can be stated as an expected situation in testlet consisting of open-ended items with low objective scoring (Kaya Uyanık & Ertuna, 2022). In the case where the testlet effect is not handled, it overestimates the reliability value of the test due to the correlation between the items in the testlet (Lee & Park, 2012). As a result of the D studies, the reliability coefficients were increased when the number of items increased by keeping the number of persons and raters constant in achievement tests consisting of open-ended items. The increase in the number of items also increases the reliability of the test (Baykul, 2000; Turgut, 1992). However, it was observed that the reliability coefficients increased to a certain number of items, and after a certain number of items, the increase did not contribute as much as before. For this reason, it should first be decided whether the test to be applied will be high-stakes testing or low-stakes testing. If it is a high-stakes test, it is recommended that the reliability should be 0.80 and above (Nunnally, 1967; as cited in Henson, 2001). Then, choosing the number of items according to the reliability coefficients at the level that will serve the test's purpose would be appropriate. In tests consisting of open-ended items, it was discovered that adding more testlets to the test was more beneficial than adding more items to the testlets. When the increase in the number of testlets is compared with the increase in the number of items, it can be stated that the increase in the number of testlet has a higher contribution to the increase in the number of items in the test. One of the reasons the rater variance was very low in the analyses according to both different designs in the tests with open-ended items is the use of a detailed rubric. Since the use of rubrics can increase objectivity (Moskal & Leydens, 2000), it may have reduced the error caused by the rater. The decision study on the number of raters determined that the increase in the number of raters did not have much effect on the reliability value. The study concluded that two raters would be sufficient due to the difficulty and inconvenience of finding a large number of raters and in terms of time and practicality. This result is similar to the results of Kaya Uyanık and Ertuna (2022). Taşdelen Teker et al (2016) obtained sufficient reliability value with two raters as a result of the D study conducted in their study in which students' communication skills were evaluated with a 5-point rating scale.

In the test consisting of multiple choice items, as in the test with open-ended ones, similar results were obtained: the reliability coefficients were estimated higher when the testlet effect was not taken into account, and the reliability coefficients were estimated lower when the testlet effect was taken into account. This result was similar with the results of studies examining the testlet consisting of multiple choice items and its reliability effect (Hendrickson, 2001; Lee & Park, 2012; Sireci et al, 1991; Taşdelen Teker, 2014; Thissen et al, 1989; Wainer, 1995). Situations, where the testlet effect is not handled may cause bias in the results and a higher estimate of the reliability value. A high correlation between items within the same testlet will also contribute to homogeneity. If the testlet effect is taken into account, the different contents of the testlets will provide heterogeneity. This may lead to a lower estimate of reliability when the testlet effect is taken into account than when it is not taken into account. It is seen in the D studies that when the number of individuals is kept constant, reliability will increase more when testlets are increased rather than when items are increased. In their simulation studies, Kaya Uyanık and Gelbal (2018) obtained a higher reliability value when the number of items increased if the testlets were equal, similar to the results of this study. In the event that each testlet had the same number of items, higher reliability was obtained as the number of testlets increased. In short, higher reliability is achieved when the total number of items increases.

In their study with dummy-coded SAT data, Sireci et al.'s (1991) determined that not taking into account the relationship between items in the same testlet led to a 10-15% overestimation in both the CTT-based and the IRT-based reliability estimation. In this study, higher G and Phi coefficients were obtained by ignoring the testlet effect in both the test consisting of open-ended items and the test consisting of multiple-choice items. If the testlet effect was taken into account in the test with open-ended items, it caused a decrease in the G coefficient (difference of 0.14) and Phi coefficient (difference of 0.12). The similar difference is greater than the difference in G coefficient (difference of 0.05) and Phi coefficient (difference of 0.05) in the achievement test, which includes testlets of multiple-choice items. Therefore, it can be stated that the item types are effective in the change in G and Phi coefficients.

Within the scope of the research, an achievement test for mathematics courses was developed. The results of the research in different fields can be examined. Since the number of items in the testlets was equal in this research, the studies were conducted on balanced designs. Research can be conducted in unbalanced designs where the number of items in the testlets varies. In addition, the results can be examined by conducting studies with different designs in which raters, which are not included in the scope of this research, are nested within persons or items are nested within raters.

Declarations

Author Contribution: Serpil KOCAOĞLU: conceptualization, methodology, analysis, writing & editing, visualization. Melek Gülşah ŞAHİN: conceptualization, methodology, writing-review & editing, supervision.

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

Ethical Approval: The study was ethically approved by the Gazi University Ethics Commission dated 30.12.2022 and numbered E.548756.

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

References

Attali, Y., Laitusis, C., & Stone, E. (2016). Differences in reaction to immediate feedback and opportunity to revise answers for multiple-choice and open-ended questions. *Educational and Psychological Measurement, 76*(5), 787-802. https://journals.sagepub.com/doi/10.1177/0013164415612548

Baykul, Y. (2000). *Eğitimde ve psikolojide ölçme: Klasik test teorisi ve uygulaması.* ÖSYM.

Berberoğlu, G. (2006). *Sınıf içi ölçme değerlendirme teknikleri*. Morpa Kültür.

Brennan, R. L. (2001). *Generalizability theory.* Springer-Verlag.

Bridgeman, B. (1992). A comparison of quantitative questions in open-ended and multiple-choice formats. *Journal of Educational Measurement, 29*(3), 253-271. https://doi.org/10.2307/145138

Doğan, N. (2009). Yazılı yoklamalar. *In H. Atılgan (Ed.), Eğitimde ölçme ve değerlendirme* (p.148). Anı.

Doğan, N. (2019a). Geleneksel ölçme ve değerlendirme teknikleri I: Yanıtı seçmeyi gerektiren ölçme araçları. *In N. Doğan (Ed.), Eğitimde ölçme ve değerlendirme* (pp. 113-138). Pegem Akademi.

Doğan, N. (2019b). Geleneksel ölçme ve değerlendirme teknikleri II: Yanıtı yapılandırmayı gerektiren ölçme araçları. *In N. Doğan (Ed.), Eğitimde ölçme ve değerlendirme* (pp:140-179). Pegem Akademi.

Downing, S. M. (2006). Twelve steps for effective test development. *In S. M. Downing & T. M. Haladyna (Eds.), Handbook of test development* (pp. 3-26).

Gessaroli, M. E., & Folske, J.C. (2002). Generalizing the reliability of tests comprised of testlets. *International Journal of Testing, 2*(3-4), 277-295. https://doi.org/10.1080/15305058.2002.9669496

Güler, N., Kaya Uyanık, G., & Taşdelen Teker, G. (2012). *Genellenebilirlik Kuramı.* Pegem Akademi.

Haladyna, T. M. (2004). *Developing and validating multiple-choice test items*. Taylor & Francis Group. https://ebookcentral.proquest.com/lib/gazi-ebooks/detail.action?docID=255610

Hendrickson, A. B. (2001). *Reliability of scores from tests composed of testlets: A comparison of methods.* Paper presented at the Annual Meeting of the National Council on Measurement in Education.

Henson, R. K. (2001). Understanding internal consistency reliability estimates: A conceptual primer on coefficient alpha. *Measurement and Evaluation in Counseling and Development, 34*(3), 177-189. https://www.tandfonline.com/doi/abs/10.1080/07481756.2002.12069034

Karasar, N. (1994). *Bilimsel Araştırma Yöntemi.* 3A Araştırma Eğitim Danışmanlık.

Karatoprak Erşen, R., & Gündüz, T. (2023). Seçme ve katkı gerektiren maddelerin yazımı ve düzenlenmesi için kontrol listeleri. *Dokuz Eylül Üniversitesi Buca Eğitim Fakültesi Dergisi* (58), 2473-2493. https://doi.org/10.53444/deubefd.1279240

Kaya Uyanık, G., & Ertuna, L. (2022). Examination of testlet effect in open-ended items. *SAGE Open*, 1-12. https://doi.org/10.1177/21582440221079849

Kaya Uyanık, G., & Gelbal, S. (2018). Madde tepki modellemesinde genellenebilirlik ile iki yüzeyli desenlerin incelenmesi. *Journal of Measurement and Evaluation in Education and Psychology, 9*(1), 17-32. https://doi.org/10.21031/epod.349718

Ko, M. H. (2010). A comparision of reading comprehension tests: Multiple-choice vs. open-ended. *English Teaching, 65*(1), 137-159. doi:10.15858/engtea.65.1.201003.137

Lee, G., & Frisbie, D. A. (1999). Estimating reliability under a generalizability theory model for test scores composed of testlets. *Applied Measurement in Education, 12*(3), 237–255. https://doi.org/10.1207/S15324818AME1203_2

Lee, G., & Park, I.-Y. (2012). A comparison of the approaches of generalizability theory and item response theory in estimating the reliability of test scores for testlet-composed tests. *Asia Pacific Education Review, 13*(1), 47-54. https://doi.org/10.1007/s12564-011-9170-0

Lee, G., Brennan, R. L., & Frisbie, D. A. (2000). Incorporating the testlet concept in test score analyses. Educational Measurement: Issues and Pratice, 19(4), 9-15. https://doi.org/10.1111/j.1745-3992.2000.tb00041.x

Moskal, B. M., & Leydens, J. A. (2000). Scoring rubric development: Validity and reliability. *Practical Assessment, Research and Evaluation, 7*(10), 1-6. https://doi.org/10.7275/q7rm-gg74

Miles, M. B., & Huberman, A. M. (1994). *Qualitative data analyses: An expanded sourcebook.* CA: Sage Publications.

Nitko, A. J., & Brookhart, S. M. (2014). *Educational assessments of students* (6th ed.). Essex: Pearson International.

Özçelik, D.A. (2013). *Test hazırlama kılavuzu*. Pegem Akademi.

Popham, J.W. (2014). Selected-response tests. *In Classroom assessment: What teachers need to know* (7th ed, pp. 155-180). Pearson Education Ltd.

Russell, M. & Airasian, P.(2008). Designing, administering, and scoring achievement tests. *Classroom assessment: Concepts and applications* içinde (7th ed, pp. 144-175). McGrawHill Higher Education.

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: a primer*. Sage Publicatons.

Sireci, S. G., Thissen, D., & Wainer, H. (1991). On the reliabilty of testlet-based tests. *Journal of Educational Measurement, 28*(3), 237-247. https://doi.org/10.1111/j.1745-3984.1991.tb00356.x

Taşdelen Teker, G. (2014). *Madde takımlarının güvenirlik ve değişen madde fonksiyonu üzerine etkisi.* Doctoral Dissertation, Hacettepe Üniversitesi Eğitim Bilimleri Enstitüsü, Ankara.

Taşdelen Teker, G., Şahin, M. G., & Baytemir, K. (2016). Using generalizability theory to investigate the reliability of peer assessment. *Journal of Human Sciences, 13*(3), 5574-5586. https://doi.org/10.14687/jhs.v13i3.4155

Tekin, H. (2009). *Eğitimde ölçme ve değerlendirme.* Yargı.

Thissen, D., Steinberg, L., & Mooney, J. A. (1989). Trace lines for testlets: A use of multiple-categorical-response models. *Journal of Educational Measurement*, *26*(3), 247-260. https://doi.org/10.1111/j.1745-3984.1989.tb00331.x

Turgut, M. F. (1992). *Eğitimde ölçme ve değerlendirme metotları.* Saydam.

Wainer, H., & Kiely, G. L. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement, 24*(3), 185-201. https://doi.org/10.1111/j.1745-3984.1987.tb00274.x

Wainer, H. (1995). Precision and differential item functioning on a testlet-based test: The 1991 law school admissions test as an example. *Applied Measurement in Education, 8*(2), 157-186. https://doi.org/10.1207/s15324818ame0802_4

Wainer, H., & Thissen, D. (1996). How is reliability related to the quality of test scores? What is the effect of local dependence on reliability? Educational Measurement: Issues and Practice, *15*(1), 22-29. http://doi: 10.1111/j.1745-3992.1996tb00803.x

Wainer, H., Bradlow, E. T., & Du, Z. (2000). Testlet response theory: An analog for the 3PL model useful in testlet-based adaptive testing. *In W. J. van der Linden & G. A. W. Glas (Eds.)*, *Computerized adaptive testing: Theory and practice* (pp. 245–269). Springer. https://doi.org/10.1007/0-306-47531-6_13

Yaman, S. (2016). Çoktan seçmeli madde tipleri ve fen eğitiminde kullanılan örnekleri. *Gazi Eğitim Bilimleri Dergisi*, *2*(2), 151-170. https://dergipark.org.tr/tr/pub/gebd/issue/35205/390659