Examination of Testlet Effect in Open-Ended Items

Gülden Kaya Uyanık and Levent Ertuna

Abstract
The testlets are prevalently used in national and international standard tests and the teacher-made achievement tests thanks to their practicality. The purpose of the current study is to evaluate the testlet effect in test consisting the open-ended items and to identify the possible errors when this effect is ignored. The data consists of the scores given by three raters of the responses of 380 students in the teacher training program to two open-ended items in two testlets. The two designs with and without the testlet effect were modeled (totally crossed random design and the nested random design) and the differences between them were analyzed by Generalizability Theory (GT). The results showed that raters had a slight effect on the scores in both models tested, and the reliability of the scores was overestimated when the testlet effect was not ignored in such measurement tools.

Keywords
generalizability theory, open-ended item, testlet, testlet effect, rater effect

Introduction
Open-ended items (“constructed-response”) are question type used to collect data in many measurement tools such as achievement tests, interview forms, or questionnaires. Open-ended items used in standard or teacher-made achievement tests are the item type that student structures an answer by using the processes of applying to information, interpreting, comparing, organizing, and justifying information, and freely expresses this by gathering writing and language skills (Badger & Thomas, 1992; Downing, 2009; Gronlund, 1998; Haladyna & Rodriguez, 2013). Although open-ended items are mostly preferred for formative assessment, their use for assessment is also increasing. Because of its advantages, such as measuring higher order skills, testing companies especially encourage the use of this item type in large-scale and high-stakes testing (Haladyna & Rodriguez, 2013).

Open-ended items have advantages and disadvantages compared to other types. The first of these advantages is that the open-ended items are more effective in assessing high-level skills. This item type makes students think much more, it is mostly preferred in the measurement of the skills at analysis, synthesis, and evaluation level, such as problem-solving and critical thinking (Sanchez, 2013). Studies conducted to measure higher-order thinking have shown that these skills can be exhibited through open-ended items (Brookhart, 2010; Schraw & Robinson, 2011; Silva, 2009; Soland et al., 2013). Its second advantage is that this type of item decreases measurement errors by reducing chance success as there is no option (Bridgeman, 1992). Another advantage is that the answer can be partially graded, contrary to the dichotomously scored items. This allows flexible scoring of the information level of students (Ebel & Frisbie, 1991). Although they are easy to be prepared (Aiken & Groth-Marnat, 2005), their administration and scoring processes are laborious, which is one of their disadvantages. The greatest disadvantage of this type of item is that as a result of the partial scoring, it causes objectively scoring. Even if raters use a detailed rubric, the scores differ among them because of rating behavior (McMillan, 2017; Turgut & Baykul, 2015). This is the most common source of error in open-ended items.

The fact that response in open-ended items can be constructed in various format allows them to be used in context-dependent item sets. The item format structured based on a common content is referred to as testlet (Haladyna & Rodriguez, 2013). It is a set of items that are generally used in problem-solving, especially to determine the reactions of students to various stimuli in person tasks or exercises.

Sakarya University, Turkey

Corresponding Author:
Levent Ertuna, Department of Educational Sciences, Faculty of Education, Sakarya University, Hendek, Sakarya 54300, Turkey.
Email: leventertuna@sakarya.edu.tr
Wainer and Kiely (1987) proposed the testlets to eliminate problems such as the content balancing, context effect and the item order problem in computerized adaptive tests. Thanks to its benefits for practical applications in education, it is used in many large-scale assessments (PISA, PIRLS, TIMSS, SAT, NAEP, etc.) (Messick, 1994). The items can be mostly grouped in a general stimulus such as reading a passage, a laboratory scenario, a graphic, or a complex problem (DeMars, 2006). The testlets can measure not only the general ability but also a range of special cognitive information processing in complex tasks thanks to their structures (Rosenbaum, 1988). There is no limit in the number of the item within the testlets but Wainer and Lewis (1990) state that the structure of them should be small enough to manipulate and large enough to form its scope.

Testlets are preferred in various test administrations for different reasons. The most significant of them is that it saves time and energy for test developers, item writers, or other stakeholders since it allows the opportunity to answer many items depending on a stimulus (Thissen et al., 1989; Wainer et al., 2000). This is the answer that why the testlets are so much used in national and international large-scale tests. Another reason is that persons are more successful in following the predetermined path such as consecutive questions who are interested in test development processes. This study using real data presents new empirical evidence for how the reliability is affected by possible error sources if the testlet effect is ignored and composed of open-ended items.

**Background**

Although open-ended items measure higher-order thinking skills, they have some psychometric issues in the assessment processes. The scoring of these items brings along the risk of being subjective based on the rater judgments. The rater effect/error stem from have a tendency higher or lower scoring than s/he deserve (severity/leniency) (e.g., Eckes, 2005; Liu & Xie, 2014), use the middle categories in the rating scale more (central tendency) (e.g., Engelhard, 1994; Leckie & Baird, 2011; Myford & Wolfe, 2009), and effect from judgment regarding the overall test (halo effect) (e.g., Engelhard, 1994; Kim, 2020). These are a systematic source of variance in the observed ratings of open-ended items (Bimpel et al., 2020; Kim & Moses, 2013; Myford & Wolfe, 2003). Previous studies have shown that the rater effect cannot be ignored in large-scale assessment, this affects item parameter estimation and the validity of the results, and decreases reliability (Donoghue et al., 2006; Kim, 2009; Koretz et al., 1994; Wang & Yao, 2013).

Several methods that test the rater effect are discussed in different measurement models (Engelhard, 2013). Classical Test Theory (CTT) investigates this problem by calculating the simple percentage of the rater agreement and inter-rater reliability and rater agreement statistic (Cohen’s kappa, Fleiss’ kappa, Krippendorff’s Alpha, etc.). These methods are based on only one error source (Myford & Wolfe, 2000).
Generalizability Theory (GT), which is an extension of CTT and variance analysis, has been proposed to eliminate this disadvantage. Apart from these, many methods based on Item Response Theory (IRT) have been presented to determine and correct or control the rater effect (Nieto & Casabianca, 2019; Wang & Yao, 2013): Many Facet Rasch Model (MFRM; Linacre, 1994) (the most popular and known among them), the hierarchical rater model (HRM; Patz, 1996), the rater bundle model (Wilson & Hoskens, 2001). In MFRM, like GT, it handles all sources of variability that affect the scoring together and estimates scores by separating the effects of each variable unlike other methods (Eckes, 2019).

Testlet is another concept addressed in the study, and its sources of error should be explained to understand the framework of the study. The advantage of testlet that structuring a subset of items by sharing the common stimulus also brings along the problem of violation of the local independence assumption of the IRT (Li, 2017; Sirici et al., 1991; Wainer & Thissen, 1996; Yen, 1993). Mainly, local independence means that a response to an item in the test should not be affected by other items when an ability is conditioned/controlled (Hambleton & Swaminathan, 1985). Previous studies have demonstrated that violation of local independence leads to imprecise person ability and item parameter estimation, overestimated test reliability scores, and errors in scaling, test equating, and calculation of standard error of measurement (Jiao & Zhang, 2015; Jiao et al., 2012; Sirici et al., 1991; Tao & Cao, 2016; Wainer et al., 2007). Lee and Park (2012) stated that the degree of overestimated test reliability scores differed in previous studies, the practical importance of this was discussed and this overestimation causes crucially errors of interpretation in high stake tests.

To overcome this issue, different methods were used under GT and IRT theories in testlet literature. GT suggested a solution with the help of using the concepts of crossed and nested design together: persons (p) crossed with items (i) nested in testlet (t) design [px (i: t)] (Lee & Frisbie, 1999). Thanks to GT, this problem can be overcome by using raw scores (Li, 2017). Another solution is to employ a polytomous IRT model by considering all items within the testlet as a single multi-scored item (Lee et al., 2001). However, this procedure brings with its possible disadvantages such as loss of information at the item level (Wainer et al., 2000, 2007; Yen, 1993). Apart from that, the local dependencies of the item within the testlet were modeled as a general random effect in the IRT approach (Li, 2017), for example, bi-factor model (Gibbons & Hedeker, 1992), the multilevel model (Jiao et al., 2005), testlet response theory (TRT) models (Bradlow et al., 1999; Wainer et al., 2007).

Various studies have been carried out to evaluate the rater effect with different models based on measurement theories in scoring open-ended items. (Cor & Peeters, 2015; Güler, 2014; Leckie & Baird, 2011; Nieto & Casabianca, 2019; Toffoli et al., 2016). In addition, the performances of the methods on this subject were compared with each other, especially GT and IRT models (e.g., Kim & Wilson, 2009; Lee & Cha, 2016; Sudweeks et al., 2004). Kim and Wilson (2009) compared two well-known and frequently used analyzing methods (GT and MFRM) deal with the rater effect. They concluded that both methods have pros and cons, and they are worthwhile in obtaining more reliable scores. Regardless of the model and theory, it was found that the raters affect the scoring and this also affects the reliability of the scores. For example, Toffoli et al. (2016) study evaluated the scoring quality of the responses given by 350 students to two open-ended items scored by 42 raters using the Many Facet Rasch model. The results of the study showed that although the raters received training course, the scoring severity of the raters differed significantly. Similarly, in recent studies addressing this subject with the GT, it has been observed that raters and their interaction with other factors (e.g., task, items, time) is a variance component of varying magnitude in scoring. These studies also revealed that more raters are needed for higher reliability estimation (Atilgan, 2019; Bouwer et al., 2015; Kim et al., 2017; Zhao & Huang, 2020).

In the literature on testlet, a number of studies were documented in which the item parameters of testlets were estimated and the reliability of the scores obtained from them was examined by employing the above-mentioned different approaches (Eckes, 2014; Eckes & Baghaei, 2015; Li, Li et al., 2010; Paap et al., 2015; Ravand, 2015; Shaw et al., 2020). Also, these issues have been addressed in many studies within the scope of GT (Chien, 2008; Kaya Uyanik & Gelbal, 2018; Lee, 2000; Lee & Frisbie, 1999; Lee & Park, 2012). Among them, a key study by Lee and Park (2012) compared the methods under GT and IRT framework using simulation data in calculating the reliability of the test scores constructed testlets. These methods are classified as item score (item-based), testlet score, and item nested-testlet approaches for both measurement theories. The results reported that the item-based methods produced the highest error and the methods in IRT overestimated more than those in GT under the same approaches. Starting from this result, it can be stated that GT makes the most accurate estimation when multiple error sources are considered together in scores. Therefore, GT proposing a solution to both the testlet effect with nested and crossed designs and the rater effect on open-ended items are the main reasons for its use in this study.

GT allows to making relative and absolute decisions for the measurement object. The variance sources of the measurement object are considered as relative and absolute sources. G and Phi coefficient is used for relative and absolute evaluation, respectively (Brennan, 2001). G coefficient regarded as the ordering of the measurement object and other measurement objects in line with each other. In this sense, it is similar to CTT’s reliability statistics. However, Phi coefficient is more robust than G coefficient, and it is calculated by considering all of the error variances (Güler, 2009).
Reliability in GT is examined with two different phases called Generalizability (G) and Decision (D) study. In G study, it is dealt with generalizations starting from the universe from allowable observations to the whole universe, so it is aimed to provide information on the sources of variance in the sample as much as possible. In D study, scenarios are formed in line with a specific purpose by using the information obtained in G study and it is aimed to make some decisions about these scenarios (Brennan, 2001; Güler et al., 2012).

Another issue that should be known in GT, is the design of facets (crossed or nested). The crossed design means that all of the conditions of a facet are associated with all of the conditions of the other facet. The nested design is the one that a condition of a facet is paired with some conditions of the other facet. The presentation of these designs is different well: the crossed and nested design are represented by “x” and “;,” respectively (Shavelson & Webb, 1991).

In literature, there are many studies that the reliability of the scores of the test composing of open-ended items are estimated by GT, and the performance of these is compared with CTT (correlation coefficient, agreement statistic, etc.) and IRT based methods (Attilgan, 2019; Bouwer et al., 2015; Cor & Peeters, 2015; Donnon & Paolucci, 2008; Zhao & Huang, 2020). Moreover, there are several studies that the parameters of testlets are obtained by using GT (Chien, 2008; Kaya Uyanık & Gelbal, 2018; Lee & Frisbie, 1999; Lee & Park, 2012; Tsai et al., 2012). As stated in the aforementioned studies, although many studies examining the testlet and rater effect are handled in various models in the framework of measurement theories, there was little, if any, the studies that investigate the testlet effect and estimate the parameters of the open-ended testlets with real data, by GT. Therefore, the motivation of the current study is to analyze the testlet effect in the whole above-discussed framework. This study has two values that should make an important contribution to the educational measurement and testing field. The first is that, unlike most studies, it uses real data instead of simulation data to examine the rater and -in particular- the testlet effect. The other one draws special attention to how the reliability of scores is changed/affected by ignoring the testlet effect and including the rater effect based on an example in a non-high-stakes achievement test for higher education.

**Table 1.** Demographic Characteristics of Participants.

| Variables                        | Frequency (%) |
|----------------------------------|---------------|
| **Gender**                      |               |
| Female                           | 306 (80.5)    |
| Male                             | 74 (19.5)     |
| **Department**                  |               |
| Science Education                | 66 (17.3)     |
| Pre-School Education             | 121 (31.8)    |
| Guidance and Psychological Counseling | 143 (37.6) |
| Turkish Language Education       | 50 (13.3)     |
| **Total**                       | 380 (100)     |

Instrument and Procedures

This study data was collected with an achievement test (totally four items) structured as two open-ended within two testlets. This test was developed by writers to measure higher-order thinking ability and include a unit of the course, Measurement and Evaluation in Education. This unit is suitable for preparing items that allow applying the acquired knowledge to real context, analyzing, and evaluating the problem, and offering creative solutions. The details of test development phases are as follows;

1. Two lecturers, who are specialists on the field of measurement and evaluation, write many items in form of testlets. It was assumed that these items, which measure higher-order thinking skills, may have some error sources and these sources were eliminated as much as possible to include another error source in the scores.
2. These items were reviewed for clarity, the suitability of content, and after the necessary improvements were made, it was decided to include these two testlet in the test.
3. The test form comprised of two testlets was applied with pilot administration.
4. After this, item difficulty and item discrimination values were calculated for each item in the testlets. The discrimination index computed was D-index. These values are given in Table 2.
5. When Table 2 was examined, it was seen that the four items had moderate difficulty and their discrimination is above 0.30. Depending on these findings, it was decided to include all of the items in the achievement test.
6. The final test formed of four items was obtained.

The final test was administrated to 380 students whose demographic characteristics were given above. Then, the
three raters polytomously scored the responses of all students to four open-ended items with an analytical rubric. These raters are experts in the subject, Measurement and Evaluation in Education, and lectured this course for at least two semesters. This rubric was designed by three psychometricians to score each item in the testlet between 0 and 10. It is used in the same way in both pilot and final administration.

### Data Analysis

The data obtained from the study were analyzed within the framework of generalizability theory. The data consists of the scores given by three raters of the responses of 380 students to two open-ended items in two testlets. Since each rater scored all items, the data are also suitable for the crossed design. Also, there is no missing value in the data matrix. With generalizability theory, pxrxi, crossed random design, which ignored testlet effect, was formed and pxrx (i:t), nested random design, which considered testlet effect, was formed. The two designs formed were compared for main and common effects, variance values, and test reliability. Furthermore, prospective scenarios were made by performing Decision (D) studies. As a result of the applications, EDUG program (Swiss Society for Research in Education Working Group, 2006) was used to make all the generalizability theory analyses that include the calculation of the reliability of scores with G and Φ (Phi) coefficients and in prediction of variance values for main and common effects based on generalizability theory. It is a user-friendly program developed by Jean Cardinet in 1996 for performing generalizability theory analyzes by allowing the use of raw scores or sums of squares (Cardinet et al., 2010)

### Results

#### Findings for Crossed Random Design (pxrxi)

The results obtained from pxrxi crossed random design are given in Table 3.

When the predicted variance and rates of variances accounted for the total variance as a result of pxrxi crossed random design G study are examined, the variance component (7.44) predicted for the main effect of person (p) explains 66% of the total variance. In generalizability studies, main effect of person is evaluated as universe score variance and states difference between persons in terms of the trait measured (Brennan, 2001; Güler et al., 2012; Kaya Uyanık & Güler, 2016; Shavelson & Webb, 1991). The fact that the rate of the predicted variance for students in the total variance is great is a wanted status. This is an indicator that some differences may appear between persons at the dimension obtained with measurement (Brennan, 2001; Kaya Uyanık & Güler, 2016). According to the result obtained in this study, it can be said that measurement process performed with open-ended items can greatly reveal the difference between persons. The variance component predicted for rater main effect (0.29) explains 2.6% of the total variance. This value is the lowest third value. Rater main effect is resulted from the inconsistency between the scorings of the raters. Thus, the fact that this effect is low is a wanted status. The variance component (0.0019) predicted from G study for item (i) main effect explains 0.01% of the total variance. Item main effect displays the differentiation grade of each measurement unit (item) difficulty level in the open-ended items. In line with the result obtained, it can be interpreted that the difficulty levels of the items applied to the students are very close to each other.

Person x rater (pr) common effect (1.55) explains 13.7% of the total variance and this value is the highest second value. The interaction between person and rater states that the raters have some inconsistencies about generosity-rigidity in scoring for some persons. Although the raters generally give consistent results, there are some differences between scorings in some students. Person x item (pi) common effect (1.07) shows 9.5% of the total variance and it is the highest third variance. This displays that the relative situations of some students differ from one item to the other. Similarly, it can be said that to answer the open-ended items correctly differs from one student to another. Rater x item (ri) common effect states 0.1% of the total variance. This value is the lowest second variance. It can be interpreted that there is no difference between the raters according to items for this result. The explanation rate of the variance component (0.94) of person x rater x item (residual) common effect for the total variance is 8.3%. The fact that the residual variance is high, is an indicator that person, rater, item interaction, immeasurable variability sources and/or random errors are great. When the values obtained are examined, it is observed that the random error rate is low for this study.

Variance values calculated over the data used in G study are employed for the decision to be made in D study. D study allows to respectively predict G and Phi coefficients for relative and absolute decisions by decreasing and increasing facet conditions in G universe. At the same time, if G and Phi

### Table 2. Item Statistics.

| Testlet  | Item difficulty | Item discrimination |
|----------|-----------------|---------------------|
| Testlet 1|                 |                     |
| 1st Item | 0.52            | 0.46                |
| 2nd Item | 0.48            | 0.54                |
| Testlet 2|                 |                     |
| 3rd Item | 0.44            | 0.58                |
| 4th Item | 0.56            | 0.49                |
coefficients are calculated over the facets including condition at the same number with facet conditions in G study, reliability values of the data in G study are obtained as well. In Table 4, G and Phi coefficient values calculated by decreasing and increasing rater number and holding item facet still in D study, and G and Phi coefficient values calculated by decreasing and increasing item number and holding rater facet still.

In Table 4, G and Phi coefficients of the present status of the study are displayed in a bold. In the study, there are four items in total in the achievement test formed of the open-ended items, and these items are scored by three raters. In this case, G coefficient is 0.895 and Phi coefficient is 0.885. It can be said that the measurement instrument used is reliable. When the scenarios are examined, it is observed that as rater number increases, reliability value increases as well when item number is held still and rater number is changed. Similarly, when rater number is held still and item number is increased, reliability value increases as item number increases. The highest reliability value was obtained from the scenario that the item number was four and the rater number was six.

**Findings for Nested Random Design pxrx (i:t)**

The results obtained in the nested random design pxrx (i:t) are given in Table 5. As a result of pxrx(i:t) nested random design G study given in Table 5, the predicted variance and total variance explanation rates are examined, variance component predicted for person (p) main effect (7.58) explains 65% of the total variance. As it is previously revealed, in generalizability studies, the fact that the rate of the predicted variance for person main effect in the total variance is high, is a wanted case. This displays that differences between persons can be revealed at the dimension obtained with measurement (Brennan, 2001; Kaya Uyanık & Güler, 2016). According to the result obtained in this study, it can be said that the measurement process performed with the open-ended items can greatly show the difference between persons. In addition, when the variance value obtained in consideration of testlet effect is higher than the one obtained when the variance value is disregarded. Accordingly, it can be stated that the testlet effect can disclose person differences very much. The variance component predicted for rater main effect (0.28) explains 2.4% of the total variance. The fact that this effect is low, can be interpreted that the inconsistency between scorings of the raters in measurement instruments with testlet. The variance component predicted from G study performed for testlet (t) main effect (0.0027) reveals 0.001% of the total variance. The main effect of the testlet refers to the measure of the change in the difficulty of the testlet. It can be seen from Table 5, this effect can be interpreted as exceedingly small since their difficulty levels very close to each other.

### Table 3. Estimated Variance Values by G Study on pxrxi Crossed Random Design and Percentage of Estimated Variances Accounted for Total Variance.

| Source of variation | df | Sum of squares | Mean squares | Variance | %   |
|---------------------|----|----------------|--------------|----------|-----|
| p                   | 379| 37755.64386    | 99.61911     | 7.43809  | 65.8|
| r                   | 2  | 901.72149      | 450.86075    | 0.28912  | 2.6 |
| i                   | 3  | 31.92807       | 10.64269     | 0.00195  | 0.0 |
| pr                  | 758| 5414.27851     | 7.14285      | 1.54991  | 13.7|
| pi                  | 1137| 4732.57193     | 4.16233      | 1.07305  | 9.5 |
| ri                  | 6  | 31.17325       | 5.19554      | 0.01119  | 0.1 |
| pri,e               | 2274| 2144.82675     | 0.94320      | 0.94320  | 8.3 |
| Total               | 4559| 51012.14386    |              | 100      |     |

### Table 4. Results of D-Study for pxrx Crossed Random Design.

| Number of raters | G Coefficient | Φ Coefficient |
|------------------|---------------|---------------|
| 2                | 0.864         | 0.850         |
| 3                | **0.895**     | **0.885**     |
| 4                | 0.912         | 0.904         |
| 5                | 0.922         | 0.915         |
| 6                | 0.929         | 0.923         |

| Number of items  | G Coefficient | Φ Coefficient |
|------------------|---------------|---------------|
| 3                | 0.883         | 0.873         |
| 4                | **0.895**     | **0.885**     |
| 5                | 0.903         | 0.892         |
| 6                | 0.908         | 0.897         |
| 7                | 0.912         | 0.907         |

Note. The number of items and raters for the current study’s condition are presented in bold.
Table 5. Estimated Variance Values by G Study on pxrx (i:t) Nested Random Design and Percentage of Estimated Variances Accounted for Total Variance.

| Source of Variation | df | Sum of squares | Mean squares | Variance | %  |
|---------------------|----|----------------|--------------|----------|----|
| p                   | 379| 37755.64386    | 99.61911     | 7.58571  | 64.6|
| r                   | 2  | 901.72149      | 450.86075    | 0.28511  | 2.4 |
| t                   | 1  | 19.47456       | 19.47456     | 0.00273  | 0.0 |
| i:t                 | 2  | 12.45351       | 6.22675      | 0.00014  | 0.0 |
| pr                  | 758| 5414.27851     | 7.14285      | 1.45963  | 12.4|
| pt                  | 379| 1043.02544     | 2.75205      | 0.44286  | 0.0 |
| rt                  | 2  | 23.30833       | 11.65417     | 0.01203  | 0.1 |
| Residual Effects    |    |                |              |          | 20.4|
| p(i:t)              | 758| 3689.54649     | 4.86748      | 1.36828  | (11.6)|
| r(i:t)              | 4  | 7.86491        | 1.96623      | 0.00317  | (0.0)|
| prt                 | 758| 988.69167      | 1.30434      | 0.27086  | (2.3)|
| pr(i:t)             | 1516| 1156.13509   | 0.76262      | 0.76262  | (6.5)|
| Total               | 4559| 51012.14386    |              | 100%     |     |

Item: testlet (i:t) common effect (0.0001) explains % 0.001 of the total variance. This value shows the differentiation of the nested items in the testlet and according to the values obtained, this differentiation is not observed. Person x rater (pr) common effect (1.45) discloses 12.4% of the total variance and this value is the highest second one. The interaction between person and rater states that the raters have some inconsistencies about generosity-rigidity in scoring for some persons. Although the raters generally give consistent results, there are some differences between scorings in some students. Person x testlet (pt) common effect (0.44) shows 0.001% of the total variance. This displays that the relative situations of some students do not differ from one item to the other. Rater x testlet (rt) common effect (0.01) explains 0.1% of the total variance. It can be interpreted that there is no difference according to testlets between the raters for this result.

The total of the four common effects in pr (i:t) design that the raters score all of the students and each student answers each testlet and the items are nested in different testlets, is called residual effect. The explanation rate of the common effect variance component for the total variance is 20.4%. The fact that the residual variance is high, is an indicator that person, rater, testlet, item interaction, immeasurable variability sources and/or random errors are high. When the values obtained are examined, it is observed that there are random errors for the study performed.

In Table 6, there are G and Phi coefficients obtained from the scenarios given above.

In Table 6, G and Phi coefficients about the present status of the study are displayed in bold. In pxr x (i:t) nested random design, there are four items in total in the achievement test comprised of the open-ended items and these items are nested in 2 testlets, all of the testlets are scored by three raters. In this case, G coefficient is 0.850 and Phi coefficient is 0.840. It can be said that the measurement instrument including testlet is reliable. When the scenarios made are examined, it is observed that when the rater and item number is held still and the testlet number is changed, reliability value increases as the testlet number increases. Similarly, when the rater and testlet number is held still and the item number is increased, reliability value increases when the item number increases. When the item number and testlet number is held still and the rater number changes, reliability increases as the rater number increases. The highest reliability value was obtained from the scenario in which the testlet number was two, the item number was two and the rater number was six (0.882).

When the data set used in the study is compared to the reliability coefficients obtained, it is seen that G (0.895) obtained with p x r x i design and Phi (0.885) coefficients are higher than G obtained with p x r x (i:t) design (0.850) and Phi (0.840) coefficients. This shows that the reliability calculated by ignoring testlet in measurement instruments with testlets will be higher than the real reliability value, therefore, it is required to consider the testlet effect in the calculations of reliability.

Discussion and Conclusion

In this study, it is aimed to examine the reliability coefficients of the open-ended questions in testlets. In line with this aim, two designs were formed in which the testlet effect was ignored and not ignored, and the differences between them
were examined through Generalizability theory. With this theory, pxrx crossed random design, which disregard the testlet effect, and pxrx (i:t) nested random design, which considered the testlet effect, were formed. The two designs formed were compared in terms of variance values and the reliability of the test for main and common effects. Furthermore, progressive scenarios were created by making Decision (D) studies.

When the two designs were examined for variance sources, it was seen that the largest source of variance was the person. This clearly shows that the achievement test is sufficient in measuring the abilities. Considering the use of open-ended items in the study, this result indicates that such items are an effective option to assess the abilities and achievements. Coincidentally the effectiveness of open-ended questions in revealing how the information is formed in the students and their success in measuring high-order skills is often emphasized in the literature (Haladyna, 1997; Lee et al., 2011; McMillan, 2017).

Similarly, another important source of variance for both of the designs is the common effect of rater and person. The interaction between person and rater indicates that the severity/leniency of raters in scoring for some people is inconsistent. In this case, it was concluded that the raters scored some students unstable. This result revealed the possible disadvantage in rating open-ended items. These results support previous research examining the rater effect on constructed response items (Cor & Peeters, 2015; Leckie & Baird, 2011; Temizkan & Sallabaş, 2011; Toffoli et al., 2016). The results of Temizkan and Sallabaş (2011) study support this finding, and they explained that when scoring open-ended items, features that are not actually measured can interfere with scoring. In the studies that especially other item types and open-ended item are compared, there are results to support this finding (Beller & Gafni, 2000; Birgili, 2014; Hancock, 1994; Rauch & Ve Hartig, 2010). As stated in the Background section, the reasons for this are the increased use of the middle rates in scoring and the judgments regarding the whole test. To minimize these errors of rater, the practices such as training possible scorers or scoring via computer interface are recommended (Swartz et al., 1999; Wolfe & McVay, 2010; Wolfe et al., 2010).

While in the crossed design, in which the testlet effect is disregarded, the residual variances 8.3%, in the nested design, in which the testlet effect is considered, this variance value is 20.4%. The fact that the residual variance is high, displays that the non-measurable variability sources and/or random errors are high. When the values obtained are examined, random error is low if the testlet effect is ignored and random error is high if it is considered. This is supported by reliability coefficients. The reliability coefficients obtained for three facet crossed design are (G = 0.895; Phi = 0.885) and the reliability coefficients obtained for four facet nested design are (G = 0.850; Phi = 0.840). In the model in which the testlet effect is not considered, the random error of the scores is underestimated, so the reliability is calculated higher than it should be. This finding broadly supports the previous studies in this area examining the effect between testlet and reliability. (Hendrickson, 2001; Lee, 2000; Lee & Park, 2012; Teker & Dogan, 2015). As reported in the Background section, Lee and Park (2012) similarly found that item-based methods have greater overestimation than item nested testlet methods in both IRT and GT. They explained one of the reasons for this situation as the dependency relationships among the items within the testlet rises, the size of the overestimation for the reliability of the scores increases in item-based methods.

| Table 6. Results of D-Study for pxrx(i:t) Nested Random Design. |
|----------------------------------------------------------|
| Number of testlets | G coefficient | Φ coefficient |
|---------------------|---------------|---------------|
| Number of raters: 3 | Number of items: 2 |
| 1                   | 0.805         | 0.796         |
| 2                   | 0.850         | 0.840         |
| 3                   | 0.866         | 0.856         |
| 4                   | 0.874         | 0.864         |
| 5                   | 0.879         | 0.869         |
| Number of raters: 3 | Number of items: 2 |
| 2                   | 0.850         | 0.840         |
| 3                   | 0.864         | 0.849         |
| 4                   | 0.872         | 0.861         |
| 5                   | 0.876         | 0.865         |
| 6                   | 0.879         | 0.868         |
| Number of raters: 3 | Number of items: 2 |
| 2                   | 0.820         | 0.806         |
| 3                   | 0.850         | 0.840         |
| 4                   | 0.866         | 0.858         |
| 5                   | 0.876         | 0.869         |
| 6                   | 0.882         | 0.877         |

Note. The number of items, raters and testlets for the current study’s condition study are presented in bold.
In the study, Decision studies were performed for pxrx (t:r) nested design in which the testlet effect was considered. As a result of the study, it was observed that the reliability increased as the testlet number and item number increased. Accordingly, it is recommended that in the open-ended tests including the testlets, many more testlets and items should be used. These results are in agreement with the findings of Kaya Uyanik and Gelbal's (2018) study which showed that the reliability of score increased depending on the number of items. In addition, they suggested that the number of items in testlet and testlets should be as high as possible.

When the findings of the study are carefully examined, it is observed that the highest reliability value is obtained from the scenario (0.882) that the testlet number is two, the item number is two and the rater number is six. In this case, it can be asserted that the most significant element that affects the reliability for the open-ended questions is the rater. In this sense, it is highly recommended to use many raters in scoring the open-ended questions. However, the increase in the number of raters does not mean that the rater effect on the scores has been completely eliminated. Toffoli et al. (2016) study showed that despite scoring by 42 raters, rater scores differ significantly.

Finally, in this study, it is clearly presented that the testlet effect cannot be ignored in the reliability estimates of the scores for the test composed testlets. Hence, this result concerns it concerns stakeholders who prepare and use these tests. As Paap et al. (2015) point out, if item authors and test developers plan to minimize this effect when using testlets, it would be useful to carefully evaluate the properties of the common stimulus in this testlet. In addition, the dependencies of the items in the testlet can be considered during the item writing and test design stages. Since the Decision (D) study conducted in this study shows that the increase in the number of items and testlets rises the reliability, this situation can be taken into consideration while developing a test. If item sets consisting of open-ended items are used, the practices (such as rater training) that will eliminate the rater effect in scoring should be administrated. Beyond these, models that conceive the testlet effect should be employed for the estimation based on these items. Although it is a straightforward method, the use of item-based methods should be avoided. The current study has some limitations. First, in this study, two testlets consisting of four items in total were applied. The number of items within testlet (since nested design is used) and testlets was limited. Second, the achievement test used in this study content only one unit. Therefore, it is a limitation that the items consist of similar theme. Last, most importantly, the gender ratio of the participants in the study is unbalanced. The gender effect may be a concern for impacting results in this study. Therefore, future research should consider using a measurement tool that is a more comprehensive achievement test consisting of more items and testlets, and having balanced demographic characteristics of the participants may be designed.

**Declaration of Conflicting Interests**
The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

**Funding**
The author(s) received no financial support for the research, authorship, and/or publication of this article.

**ORCID iDs**
Gülden Kaya Uyanik https://orcid.org/0000-0002-8100-6994
Levent Ertuna https://orcid.org/0000-0001-7810-1168

**References**
Aiken, L. R., & Groth-Marnat, G. (2005). Psychological testing and assessment (12th ed.). Pearson Education.
Atilgan, H. (2019). Reliability of essay ratings: A study on generalizability theory. Eurasian Journal of Educational Research, 19(80), 1–18.
Badger, E., & Thomas, B. (1992). Open-ended questions in reading. Practical Assessment, Research & Evaluation, 3(4), 03.
Beller, M., & Gafni, N. (2000). Can item format (multiple choice vs. open-ended) account for gender differences in mathematics achievement? Sex Roles, 42(1), 1–21.
Bimpeh, Y., Pointer, W., Smith, B. A., & Harrison, L. (2020). Evaluating human scoring using generalizability theory. Applied Measurement in Education, 33(3), 198–209.
Birgili, B. (2014). Open ended questions as an alternative to multiple choice: Dilemma in Turkish examination system. [Master’s thesis, Middle East Technical University]. OpenMETU. http://etd.lib.metu.edu.tr/upload/12617861/index.pdf
Bouwer, R., Béguin, A., Sanders, T., & van Den Bergh, H. (2015). Effect of genre on the generalizability of writing scores. Language Testing, 32(1), 83–100.
Bradlow, E. T., Wainer, H., & Wang, X. (1999). A Bayesian random effects model for testlets. Psychometrika, 64(2), 153–168.
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.
Brookhart, S. M. (2010). How to assess higher-order thinking skills in your classroom. ASCD.
Cardinet, J., Johnson, S., & Pini, G. (2010). Applying generalizability theory using EduG. Routledge – Taylor & Francis Group.
Chien, Y. M. (2008). An investigation of testlet-based item response models with a random facets design in generalizability theory. The University of Iowa.
Cor, M. K., & Peeters, M. J. (2015). Using generalizability theory for reliable learning assessments in pharmacy education. Currents in Pharmacy Teaching and Learning, 7(3), 332–341.
DeMars, C. E. (2006). Application of the Bi-factor multidimensional item response theory model to testlet-based tests. Journal of Educational Measurement, 43(2), 145–168.
Donnon, T., & Paolucci, E. O. (2008). A generalizability study of the medical judgment vignettes interview to assess students’ noncognitive attributes for medical school. BMC Medical Education, 8(1), 58.
Donoghue, J. R., McClellan, C. A., & Gladkova, L. (2006). Using rater effects models in NAEP. Unpublished manuscript.
Downing, S. M. (2009). Written tests: Constructed-response and selected-response formats. In S. M. Downing & R. Yudkowsky (Eds.), *Assessment in health professions education* (pp. 149–175). Routledge.

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

Eckes, T. (2005). Examining rater effects in TestDaF writing and speaking performance assessments: A many-facet Rasch analysis. *Language Assessment Quarterly: An International Journal*, 2(3), 197–211.

Eckes, T. (2014). Examining testlet effects in the TestDaF listening section: A testlet response theory modeling approach. *Language Testing*, 31(1), 39–61.

Eckes, T. (2019). Implications for rater-mediated language assessment. In V. Aryadoust & M. Raquel (Eds.), *Quantitative data analysis for language assessment volume I: Fundamental techniques* (pp. 153–175). Routledge.

Engelhard, G. (1994). Examining rater errors in the assessment of written composition with a many-faceted Rasch model. *Journal of Educational Measurement*, 31(2), 93–112.

Engelhard, J. R. G. (2013). Invariant measurement: Using Rasch models in the social, behavioral, and health sciences. Routledge.

Gibbons, R. D., & Hedeker, D. R. (1992). Full-information item bifactor analysis. *Psychometrika*, 57(3), 423–436.

Gronlund, N. E. (1998). *Assessment of student achievement*. Allyn & Bacon Publishing.

Güler, N. (2009). Generalizability theory and comparison of the results of G and D studies computed by SPSS and GENOVA packet programs. *Education Sciences*, 34(154), 93.

Güler, N. (2014). Analysis of open-ended statistics questions with many facet rasch model. *Eurasian Journal of Educational Research*, 55, 73–90.

Güler, N., Kaya Uyanık, G., & Teker, G. T. (2012). *Genelenebilirlik kurami [generalizability theory]*. Pegem Akademi.

Haladyna, T. M. (1997). Writing test items to evaluate higher order thinking. Allyn & Bacon Publishing.

Haladyna, T. M., & Rodriguez, M. C. (2013). *Developing and validating test items* (1st ed.). Routledge.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Kluwer Academic Publishers.

Hancock, G. R. (1994). Cognitive complexity and the comparability of multiple-choice and constructed-response test formats. *Journal of Experiential Education*, 62(2), 143–157.

Hendrickson, A. B. (2001). (April). Reliability of scores from tests composed of testlets: A comparison of methods [Paper presentation]. Annual meeting of the National Council on Measurement in Education, Seattle.

Jiao, H., Kamata, A., Wang, S., & Jin, Y. (2012). A multilevel testlet model for dual local dependence. *Journal of Educational Measurement*, 49(1), 82–100.

Jiao, H., Wang, S., & Kamata, A. (2005). Modeling local item dependence with the hierarchical generalized linear model. *Journal of Applied Measurement*, 6(3), 311–321.

Jiao, H., & Zhang, Y. (2015). Polytomous multilevel testlet models for testlet-based assessments with complex sampling designs. *British Journal of Mathematical and Statistical Psychology*, 68(1), 65–83.

Kaya Uyanık, G., & Gelbal, S. (2018). Investigation of two facets design with generalizability In item response modeling. *Journal of Measurement And Evaluation In Education And Psychology-Epod*, 9(1), 17–32.

Kaya Uyanık, G., & Güler, N. (2016). Kavram Haritası Puanlarının Güvenilirliğini İncelenmesi: Genellenebilirlik Kuramında Çaprazlamış Karışık Desen Örnegi [Investigation of Concept Map Scores’ Reliability: Example of Crossed Mixed Design in Generalizability Theory]. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi [Hacettepe University Journal of Education]*, 31(1), 97–111.

Kim, G. Y., Schatschneider, C., Wanzek, J., Gatlin, B., & Al Otaiba, S. (2017). Writing evaluation: Rater and task effects on the reliability of writing scores for children in grades 3 and 4. *Reading and Writing*, 30(6), 1287–1310.

Kim, H. (2020). Effects of rating criteria order on the halo effect in L2 writing assessment: A many-facet rasch measurement analysis. *Language Testing in Asia*, 10(1), 1–23.

Kim, S., & Moses, T. (2013). Determining when single scoring for constructed-response items is as effective as double scoring in mixed-format licensure tests. *International Journal of Testing*, 13(4), 314–328.

Kim, S. C., & Wilson, M. (2009). A comparative analysis of the ratings in performance assessment using generalizability theory and the many-facet Rasch model. *Journal of Applied Measurement*, 10(4), 408–423.

Kim, Y. (2009). Combining constructed response items and multiple choice items using a hierarchical rater model. Columbia University.

Koretz, D., Stecher, B., Klein, S., & McCaffrey, D. (1994). The Vermont portfolio assessment program: Findings and implications. *Educational Measurement Issues and Practice*, 13(3), 5–16.

Leckie, G., & Baird, J. A. (2011). Rater effects on essay scoring: A multilevel analysis of severity drift, central tendency, and rater experience. *Journal of Educational Measurement*, 48(4), 399–418.

Lee, G. (2000). A comparison of methods of estimating conditional standard errors of measurement for testlet-based test scores using simulation techniques. *Journal of Educational Measurement*, 37(2), 91–112.

Lee, G., Brennan, R. L., & Frisbie, D. A. (2000). Incorporating the testlet concept in test score analyses. *Educational Measurement Issues and Practice*, 19(4), 9–15.

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.

Lee, G., Kolen, M. J., Frisbie, D. A., & Ankenmann, R. D. (2001). Comparison of dichotomous and polytomous item response models in equating scores from tests composed of testlets. *Applied Psychological Measurement*, 25(4), 357–372.

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.

Lee, H. S., Liu, O. L., & Linn, M. C. (2011). Validating measurement of knowledge integration in science using multiple-choice and explanation items. *Applied Measurement in Education*, 24(2), 115–136.
Lee, M., & Cha, D. (2016). A comparison of generalizability theory and many facet rasch measurement in an analysis of mathematics creative problem solving test. *Journal of Curriculum Evaluation, 19*(2), 251–279.

Li, F. (2017). An information-correction method for testlet-based test analysis: From the perspectives of item response theory and generalizability theory. *ETS Research Report Series, 2017*(1), 1–25.

Linacre, J. M. (1994). *Many-facet rasch measurement*. MESA Press.

Liu, J., & Xie, L. (2014). Examining rater effects in a WDCT pragmat-ics test. *International Journal of Language Testing, 4*(1), 50–65.

Li, Y., Li, S., & Wang, L. (2010). Application of a general polytomous testlet model to the Reading section of a large-scale English language assessment. *ETS Research Report Series, 2010*(2), i–34.

McMillan, J. H. (2017). *Classroom assessment: Principles and practice that enhance student learning and motivation*. Pearson.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher, 23*(2), 13–23.

Myford, C. M., & Wolfe, E. W. (2000). Monitoring sources of variability within the test of spoken English assessment system. *ETS Research Report Series, 2000*(1), i–51.

Myford, C. M., & Wolfe, E. W. (2003). Detecting and measuring rater effects using many-facet rasch measurement: Part I. *Journal of Applied Measurement, 4*(4), 386–422.

Myford, C. M., & Wolfe, E. W. (2009). Monitoring rater performance over time: A framework for detecting differential accuracy and differential scale category use. *Journal of Educational Measurement, 46*(4), 371–389.

Nieto, R., & Casabianca, J. M. (2019). Accounting for rater effects with the hierarchical rater model framework when scoring simple structured constructed response tests. *Journal of Educational Measurement, 56*(3), 547–581.

Paap, M. C. S., He, Q., & Veldkamp, B. P. (2015). Selecting testlet features with predictive value for the testlet effect. *Sage Open, 5*(2), 2158244015581860.

Patz, R. J. (1996). Markov chain Monte Carlo methods for item response theory models with applications for the National Assessment of Educational Progress [Doctoral dissertation]. Carnegie Mellon University.

Rauch, D. P., & Ve Hartig, J. (2010). Multiple-choice versus open-ended response formats of reading test items: A twodimensional IRT analysis. *Psychological Test and Assessment Modeling, 52*(4), 354–379.

Ravand, H. (2015). Assessing testlet effect, impact, differential testlet, and item functioning using cross-classified multilevel measurement modeling. *Sage Open, 5*(2), 2158244015585607.

Rosenbaum, P. R. (1988). Items bundles. *Psychometrika, 53*(3), 349–359.

Sanchez, W. B. (2013). Open-ended questions and the process standards. *Mathematics Teacher, 107*(3), 206–211.

Schraw, G., & Robinson, D. H. (eds.) (2011). *Assessment of higher order thinking skills*. IAP.

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer*. SAGE.
Wainer, H., & Kiely, G. L. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement, 24*(3), 185–201.

Wainer, H., & Lewis, C. (1990). Toward a psychometrics for testlets. *Journal of Educational Measurement, 27*(1), 1–14.

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.

Wang, Z., & Yao, L. (2013). The effects of rater severity and rater distribution on examinees’ ability estimation for constructed-response items. *ETS Research Report Series, 2013*(2), 1–22.

Wilson, M., & Hoskens, M. (2001). The rater bundle model. *Journal of Educational and Behavioral Statistics, 26*(3), 283–306.

Wolfe, E. W., Matthews, S., & Vickers, D. (2010). The effectiveness and efficiency of distributed online, regional online, and regional face-to-face training for writing assessment raters. *The Journal of Technology, Learning and Assessment, 10*(1), 1–22.

Wolfe, E. W., & McVay, A. (2010). Rater effects as a function of rater training context. http://images.pearsonassessments.com/images/tmrs/tmrs_rg/RaterEffects_101510.pdf

Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement, 30*(3), 187–213.

Zhao, C., & Huang, J. (2020). The impact of the scoring system of a large-scale standardized EFL writing assessment on its score variability and reliability: Implications for assessment policy makers. *Studies in Educational Evaluation, 67*, 100911.