edam

ISSN 1303-0485 • **eISSN** 2148-7561
DOI 10.12738/estp.2015.4.2577
Copyright © 2015 EDAM • http://www.estp.com.tr
Educational Sciences: Theory & Practice • 2015 August • 15(4) • 969-980

Received | June 6, 2014
Accepted | March 25, 2015
OnlineFirst | May 15, 2015

The Effects of Testlets on Reliability and Differential Item Functioning*

Gulsen Tasdelen Teker[a]
Sakarya University

Nuri Dogan[b]
Hacettepe University

Abstract
Reliability and differential item functioning (DIF) analyses were conducted on testlets displaying local item dependence in this study. The data set employed in the research was obtained from the answers given by 1500 students to the 20 items included in six testlets given in English Proficiency Exam by the School of Foreign Languages of a state University in Turkey. One of the purposes of this study was to determine the influences of the tests composed of testlets on reliability, so the reliability coefficients obtained for cases where the influences of testlets were considered and those for cases where the testlet influences were not considered were compared. In consequence of the G theory analyses conducted in this context, it was found that the G and Phi coefficients estimated by not considering the testlet effects were higher than those estimated by considering the testlet effects. It was concluded that the reliability was estimated to be relatively higher when the influences of the testlet were not considered. Two methods were used in this study so as to determine the effects of testlets on differential item functioning and the results were compared. In the DIF-determining method considering the testlet effect, both the number of items displaying DIF at the significant and estimated levels of DIF were found to be higher than in the method not considering the testlet effect.

Keywords: Testlet • Reliability • Differential item functioning • Generalizability theory • Local item independence

* This study is based on a brief summary of the doctoral dissertation entitled "The Effects of Testlets on Reliability and Differential Item Functioning" prepared in the Educational Measurement and Evaluation Program, Hacettepe University, Turkey.

a **Corresponding author**
Gulsen Tasdelen Teker (PhD), Department of Educational Sciences, Faculty of Education, Sakarya University, Sakarya, Turkey
Research areas: Generalizability theory; Differential item functioning
Email: gtasdelen@sakarya.edu.tr

b Assoc. Prof. Nuri Dogan (PhD), Department of Educational Sciences, Faculty of Education, Hacettepe University, Ankara, Turkey
Email: nurid@hacettepe.edu.tr

Numerous large-scale international examinations (TOEFL, GRE, PISA, TIMMS, PIRLS, etc.), as well as national examinations (KPSS, YDS, ALES, YGS, LYS, etc.), are administered every year. On the basis of the results obtained in those examinations, assorted information is acquired in relation to individuals and to nations, and important decisions are made about them. Therefore, it is an indispensable necessity to present evidence on the reliability and validity for such examinations.

Writing the items, which are considered as the units constituting the tests, may be thought of as one of the most important steps in test preparation. Writing the items, which measure the intended properties and having the intended psychometric properties are quite demanding work (Özçelik, 2010). For this reason, groups of items called testlets (Wainer & Kiely, 1987) have been frequently used in standardized tests; these have been administered nationally and internationally for 20 years because of their advantages in time, cost, or measurement approach for a specific domain.

A testlet expresses a series of items related to only one common stimulus. The performance to be displayed in relation to each item constituting a testlet depends on both a general ability and specific ability related to the content or the topic (Rosenbaum, 1988 as cited in Wainer & Lewis, 1990). For instance, it may focus on a reading passage, a lab application, a graph question, or on a complex problem (DeMars, 2006).

The beneficial results yielded by the use of testlets in educational applications have led to the use of them in many large-scale tests. Yet, due to certain statistical features of testlets, their implementation in unidimensional measurement models has caused the emergence of some unfavorable circumstances (Fukuhara & Kamata, 2007; Yen, 1993).

One of the technical concerns caused by the use of testlets is the violation of the assumption of local item independence (Sireci, Thissen, & Wainer, 1991; Thissen, Steinberg, & Mooney, 1989; Wainer & Kiely, 1987; Wang & Wilson, 2005; Yen, 1993). There is a claim that the common stimuli of the items constituting the testlets have created dependence in individuals' responses to those items. For example, an individual previously informed of a theme in a reading passage would both be able to answer the questions about the text and be in a more advantageous position than those who are at the same level of ability, but who were not previously informed of the topic.

Local Item Independence

The assumption of local item independence is an important hypothesis for a number of measurement theories; therefore, the violation of the assumption, which results in local item dependence, may lead to undesired results. For example, based on the assessment of classical test theory, the standard error of the measurement may be underestimated and thus the reliability may be overestimated owing to the fact that the testlets violate the assumption of local item independence (Sireci et al., 1991; Wainer, 1995; Wainer & Thissen, 1996; Wainer & Wang, 2000; Yen, 1993). When looked at from the perspective of item response theory (IRT), the violation of the assumption results in overestimating the knowledge obtained from the test, and thus underestimating the standard error of the ability (Thissen et al., 1989; Yen, 1993). This also causes the misestimating of the item parameters (DeMars, 2006).

Reliability

The aim in measurement studies is to obtain observed scores as close to the actual scores as possible. Conversely, reliable measurement results—the measurement results that are as close as possible to the real scores—are only actualized on the condition that the random errors are minimized (Baykul, 2000). Item based estimation methods are generally used in determining reliability. This does not pose a problem if the tests are composed of items independent of each other. However, if the test that reliability is to be determined for is composed of testlets, which are defined as the items depending on a common stimulus, the fact that traditional reliability determination methods causes biased estimations stands as a reality, and has been common knowledge among researchers for numerous years (Lee & Frisbie, 1997, 1999; Lee & Park, 2012; Wainer, 1995). In consequence, on reviewing the relevant literature, it has been found that a great number of studies are available that conclude that the reliability of the scores for tests consisting of testlets can be higher than the ones determined through item based reliability estimation methods (Hendrickson, 2001; Lee & Frisbie, 1997; Lee & Park, 2012; Sireci et al., 1991; Wainer & Thissen, 1996).

Reliability, which is defined in classical test theory (CTT) as the consistency of the scores obtained through measurements, can vary according to the source to which the error is connected. Therefore, the errors in the measurement results are considered

as the errors coming from only one source of variability, and this emerges as a restriction of CTT. One of the most important assumptions of item response theory (IRT) is local item independence. If researchers employ an IRT model different from this one, the results obtained will change and the researchers will need to meet the strong assumptions (unidimensionality and local item independence) of IRT (Lee & Frisbie, 1997).

Generalizability Theory

Generalizability (G) theory is a statistical theory related to the reliability of performance measurements (Shavelson & Webb, 1991). The G theory can deal with the problems mentioned above that are encountered when both the CTT and the IRT based methods are used. Since the methods in G theory can consider several sources of error simultaneously, estimations can be made more accurately than the ones in CTT. As it is not necessary to meet very strong assumptions like in IRT, it can be used for the estimation of reliability of tests composed of testlets.

An important situation faced in the G theory is the type of designs and whether they are balanced or unbalanced. The number of observations in balanced designs is equal at each level for the source of variability (Brennan, 2001a). To cite an example, consider a test containing 10 individuals (s), 20 items (i), and 5 testlets (h). In this example, considering the individuals as the object of measurement, and the items and the testlets as the facets, the design can be represented as the *sx(i:h)* design where the items are nested in the testlets and the individuals are crossed with them. An equal number of items are included in each testlet in this example. That is to say, each of the five testlets has four items, thus there is a balanced data structure in which all individuals answer a 20-item test.

On examining the tests composed of testlets, it is frequently observed that the number of items in testlets can differ, and that this causes the emergence of unbalanced data sets. The *sx(i:h)* design mentioned above will be considered differently at this point. The example considered will turn into an unbalanced data set when the equal number of items in the testlets differs because the observations obtained for each variable in unbalanced designs are not equal. The reasons for this are: (1) the availability of the missing data, and (2) the real variability of the number of observations concerning the levels of variables, as in this example (Brennan, 2001a).

Differential Item Functioning (DIF)

Regardless of the type of items constituting the test, the most important psychometric property that it should have is validity. The significant elements of threat in relation to validity are item bias and test bias (Clauser & Mazor, 1998). Neutrality against those taking a test is a property that should be satisfied in tests. The differing performances displayed in relation to an item by individuals being at equal levels of ability but in different groups can be accounted for by differential item functioning (DIF). This can also be defined as the differentiation of the probability of individuals who are similar in the measured property but in different sub-groups such as gender and socio-economic level to answer an item correctly (Hambleton, Swaminathan, & Rogers, 1991).

The increase in the use of testlets in standardized tests in the field of education over the last 20 years has led to an increase in studies concerning how to score and analyze them. Traditionally, the effects of testlets were ignored and each item constituting the testlet was scored as if they were independent items (Bradlow, Wainer, & Wang, 1999; Lee, Kolen, Frisbie, & Ankenmann, 2001; Sireci et al., 1991; Wainer, 1995; Wainer & Wang, 2000; Yen, 1993). The DIF analyses performed in these studies are considered at item level (Sireci et al., 1991; Wainer & Thissen, 1996). Other studies calculating the sums or averages of the items constituting testlets, and then obtaining the scores at the testlet level are also available in the literature (Lee, Dunbar, & Frisbie, 2001; Lee & Frisbie, 1999; Sireci et al., 1991).

Another method of scoring is *testlet scoring* where testlets are considered as one single item and are polytomously scored; polytomous item response models, especially Bock's nominal response model, are used in this process (Sireci et al., 1991; Thissen et al., 1989; Wainer, 1995; Wainer & Thissen, 1996; Yen, 1993). However, the DIF derived from this method is at testlet level rather than item level. In other words, through this method, the differential testlet functioning (DTF) is derived. Therefore, specific items which lead to DIF cannot be determined (Fukuhara & Kamata, 2011). Since constructing a testlet is demanding, time-consuming work, taking the testlet displaying DTF out of the item pool would not be a desirable case. Instead, identifying and regulating the problematic items in the testlet would be advantageous in re-using the testlet.

The items constituting the testlet are dependent on each other is a positive situation, but in doing so there is the loss of information in relation to the response design of the individuals responding

to those items (Wainer, Bradlow, & Du, 2000). To eliminate this negative outcome, attaching the random effect of the testlets effects to the original IRT models is also considered as a strategy (Bradlow et al., 1999; Li, Bolt, & Fu, 2006; Wainer et al., 2000; Wang, Bradlow, & Wainer, 2002). This strategy is called the Testlet Response Theory (TRT).

Many testlet models consider the testlet effects as a random result that is available in the TRT (Bradlow et al., 1999; Li et al., 2006; Wainer et al., 2000; Wang et al., 2002). All TRT models recommend the testlet parameter, in addition to the traditional IRT parameters, in order to determine the amount of local item dependence. All TRT models that are developed are adapted from the IRT models or from the previously suggested TRT models.

Li et al. (2006), and De Mars (2006) suggest testlet models based on polytomous item response theory. The representation of the bi-factor polytomous testlet response theory model is shown in Equation 1:

$$P(y_{ij} = 1) = a_{i1}\theta_j - b_i + a_{i2}\gamma_{d(i)j} \quad (1)$$

In Equation 1, the ability component, represented by θ_j, expresses the first dimension, and the random effect component for the testlets, represented by $\gamma_{d(i)j}$, expresses the second dimension. Both have standard normal distribution, and are independent of each other. a_{i1} and a_{i2} express the discrimination parameters of item j. a_{i1} expresses the relation between ability and item j whereas a_{i2} expresses the relation between testlet effect and item j. That is to say, separate discrimination parameters are available for the two dimensions considered. And, b_i is the item difficulty parameter.

In this testlet model, the testlet effect is interpreted differently from 2P-TRT and from 3P-TRT. If the a_{i2} parameter is estimated higher than the a_{i1} parameter, it may be said that the second dimension representing the testlets is more dominant. Both, Li et al. (2006) and DeMars (2006) concluded that the bi-factor polytomous testlet theory models fit better than the TRT models for testlets in data containing local item dependence caused by testlets.

As mentioned previously there are many models for determining DIF, some of them are based on IRT and some are not. However, when applying the IRT-based DIF-determining models to tests containing testlets, the value of the DIF can be biased (Fukuhara & Kamata, 2007). Hence, it is necessary to use the models that consider the local item dependence caused by testlets to estimate more accurately the DIF value in those cases.

The model suggested by Fukuhara and Kamata (2011) and used in this study is basically the developed form of the bi-factor multidimensional item response theory model for testlets (bi-factor MIRT) suggested by Li et al. (2006) and DeMars (2006), and can be represented mathematically as in Equation 2:

$$\ln\left(\frac{P(y_{ij} = 1)}{P(y_{ij} = 0)}\right) = a_i(\beta_\theta G_j + \zeta_j - b_i + \gamma_{d(i)j} - \beta_i G_j) \quad (2)$$

In Equation 2, β_θ, represents the group effect (G_j), in ability θ_j and considers the average ability difference between the focus ($G_j = 1$) and the reference ($G_j = 0$) groups. The ζ_j is considered a residual for individual j, and β_i represents the DIF magnitude of item *i*.

This study employs the IRT-based DIF-determining method which does not consider the testlet effect in order to control the DIF-determining model based on the bi-factor MIRT model for testlets. This IRT-based DIF-determining model is the developed form of the two-parameter logistic TRT model. It is believed that since there might be differences between the discriminations of the items in the study, this model will determine the DIF better than the Rasch TRT model does.

This model considers the group covariates in determining DIF. The mathematical expression of the model is shown in Equation 3:

$$\ln\left(\frac{P(y_{ij} = 1)}{P(y_{ij} = 0)}\right) = a_i(\beta_\theta G_j + \zeta_j - b_i + \gamma_{d(i)j} - \beta_i G_j) \quad (3)$$

Equation 3 is the same as Equation 2 except for the testlets random effect ($\gamma_{d(i)j}$). Consequently, an attempt was made at the second stage of this study to compare the effects of testlets in DIF-determining through these two methods.

Purpose of the Study

This study in general aims at determining the effects of testlets on reliability and on differential item functioning, which is an issue of validity, by using a test composed of testlets. Although many theses, articles, and reports are available with regard to test reliability and to differential item functioning in Turkey, no studies concerning the testlet effects have been encountered. Because local item dependence is a property capable of affecting reliability and DIF analyses stem from the characteristic properties of the testlets, which are frequently used in tests like the ALES and YDS, it is considered an important topic to examine. Therefore, in discussing reliability and DIF-determining methods, an attempt was made in this research to exhibit the differences between methods considering and not considering the testlet effects.

Although there has been an increase in the number of research studies conducted in Turkey in relation to generalizability (G) theory, it is observed that balanced designs are used in the majority of them. The research studies performed by Nalbantoğlu Yılmaz and Uzun Başusta (2012), and by Nalbantoğlu Yılmaz (2012) can be given as examples for the limited number of studies performed in Turkey in which an unbalanced design is used. Because the G theory analyses are performed in unbalanced designs, which are more appropriate for real case data, it is believed that this study will contribute to the field.

Problem Statement and Sub-problems

This study seeks answers to the problem question "do testlets have any effects on reliability and differential item functioning" as well as to the sub-problems listed below in accordance with the purpose mentioned:

1. Are there any differences between the reliability coefficients estimated at the level of item and testlet without including the testlet effect and the reliability coefficients estimated by including the testlet effect?
2. What are the results for generalizability theory of the unbalanced *sx(i:h)* design in which items (i) nested in testlets (h) and individuals (s) are crossed with them?
3. Do the number of items displaying DIF and the estimated level(s) of DIF differ according to gender when the testlet effects are considered and not considered in the tests composed of testlets?

Method

Since the aim of the study is to determine the effects of testlets on test reliability and differential item functioning, it is a descriptive study.

Study Group

The study group was composed of 1500 students responding to 20 items, scored as 1-0, included in the reading comprehension/ text comprehension, part of the English Proficiency Exam given at the School of Foreign Languages of a state University. Of the study group, 661 (44.07%) participants were female and 839 (55.93%) were male.

Descriptive Test Statistics

The descriptive statistics for the test considered in the study are shown in Table 1.

Table 1
Descriptive Statistics of the Test

Variable	Value
Number of items	20
Number of students	1500
Mean	10.566
Standard deviation	4.089
Skewness	0.242
Kurtosis	-0.788
Mean difficulty	0.541
Mean Discrimination	0.475
Cronbach a	0.776

According to Table 1, the score distribution had a skewness value partly to the right and it partly had kurtosis. On the other hand, the results of the test were at a medium level of difficulty, and the discrimination indexes of the items that were obtained from the value of mean discrimination were sufficient in general. The reliability coefficient for the test was estimated at approximately .78. These statistics can be interpreted as that there are no obstacles to the test that need to be worked on.

Data Analysis

The G-String (Bloch & Norman, 2011) program that functions on the basis of urGENOVA (Brennan, 2001b) was used in the generalizability theory analyses; it was performed by including (at item and testlet levels) and not including the testlet effect in determining the reliability of the results of test composed of testlets.

In tests composed of testlets, the DIF-determining method considering the testlet effects based on the "bi-factor multidimensional IRT model for testlets with covariates" and the DIF-determining method not considering the testlet effects based on the "two parameter logistic IRT model with covariates" were used to determine the items displaying DIF. The analyses were done using the WinBUGS 1.4 (Spiegelhalter, Thomas, & Best, 2003) program. Moreover, opinions were obtained from 10 experts in relation to the differing functions in DIF displaying items. One of the experts held a doctorate degree in the field of measurement and evaluation, and focused on DIF. One expert was preparing a doctoral dissertation in the field of measurement and evaluation and had research studies focused

on DIF. Three of the experts were preparing their doctoral dissertation(s) in the field of measurement and evaluation. The remaining five experts were in the field of English language teaching — one was a professor, two were studying for their doctoral degrees, and two were English teachers-.

Findings

Findings Concerning the First Sub-problem

The data from the 1500 students (s) taking the English Proficiency Test resulted in 20 items (i) that nested in six testlets (h). Because the number of items included in each testlet is different in this data set, the researcher worked with an unbalanced data set concerning the *sx(i:h)* design. The distribution of items into testlets is shown in Table 2.

Table 2
The Distribution of Items into Testlets

Testlet number	Item numbers in testlets	Number of item in a testlet
1	1, 2	2
2	3, 4, 5	3
3	6, 7, 8	3
4	9, 10, 11	3
5	12, 13, 14	3
6	15, 16, 17, 18, 19, 20	6
		Total: 20 items

The G theory analyses performed without considering the testlet effects were done in two different ways:

1. The random design (*sxi*) in which individuals (s) were crossed with items (i) at the level of items: A data set composed of 1500 students and 20 items was used.
2. The random design (*sxh*) in which individuals (s) were crossed with testlets (h) at the level of testlets: A data set composed of 1500 students and six testlets was used. Each score of testlets was found by calculating the average for the items constituting the testlets.

In the G theory analyses that were performed considering the testlet effect, the random design [*sx(i:h)*] where the answers given by 1500 students to the nested 20 items in the six testlets were crossed and were used. Since the number of items in each testlet was different, the data was unbalanced and the analyses were done by taking this into consideration.

The values obtained in consequence of the analyses are shown in Table 3.

Table 3
Reliability Coefficients

		The situations in which the effect of testlets were considered		The situation in which the effect of testlets was not considered [*sx(i:h)*]
		Item level (*sxi*)	Testlet level (*sxh*)	
Reliability Coefficient	G	0,776	0,762	0,761
	Phi	0,757	0,712	0,711

On examining Table 3, it was found that the G coefficient obtained in consequence of the generalizability theory analyses performed at the item level on the 20-item test data was 0.78 while the Phi coefficient was 0.76. The G coefficient obtained by calculating the averages for the items in each testlet was 0.76 whereas the Phi coefficient was 0.71. In this case, the G coefficient estimated at the item level was about 0.02 higher than the one at the testlet level. The Phi coefficient estimated at the item level, however, was found to be 0.07 higher than the one at the testlet level. The lowest reliability values read in Table 3 were estimated in a case where the testlet effects were considered.

The G and the Phi coefficients that were estimated at the testlet level and by considering the testlet effects were found to be very close. Yet, the analyses were performed at the testlet level as if there were six items in the test. In other words, the averages were calculated with the six items in the testlets, and the reliability analyses were done with the scores for the six testlets. On the other hand, on including the testlet effects, the analyses were conducted with the 20-item data set. The fact that the reliability values were very close when numerous items are so different, (6 versus 20 items) may be interpreted as that one of the estimations was higher than the actual one. It may be possible that the reliability values that had been closely estimated to the ones for the 20-item data were estimated higher than the actual ones for the six items.

Consequently, according to Table 3, it is clear that the G and Phi coefficients estimated without considering the testlet effects are higher than the ones estimated by considering the testlet effects. The estimation of reliability without considering the testlet effects as higher than the actual is compatible with the findings obtained in studies in literature (Hendrickson, 2001; Lee & Frisbie, 1997, 1999; Sireci et al., 1991; Wainer & Thissen, 1996). Results similar to the ones obtained in the cited studies were gathered in this research.

Findings Concerning the Second Sub-problem

The analysis results for the generalizability theory performed for the unbalanced *sx(i:h)* design where items (i) nested in the testlets (h) and individuals (s) were crossed with them are given below as the findings obtained in generalizability (G) and decision (D) studies.

The G Study Findings: The variance components estimated through the G study performed with the unbalanced *sx(i:h)* design, and the total variance explanation ratios are given in Table 4.

On examining the variance values estimated through the G study performed with the unbalanced *sx(i:h)* design and the total variance explanations given in Table 4, it is found that the variance for the individual (s) as the main effect explains 13.1% of the total variance with a value of 0.033. This is the second highest value read in the table. This is an indicator of the fact that interpersonal differences can be exhibited, which is a desirable case.

The variance component estimated for the testlet (h) main effect accounts for 5.2% of the total variance with a value of 0.013. The smallness of this value indicates that the difference between the difficulty levels from one testlet to another is not very different.

The numbers of items included in each testlet are different in the study, and the variance component for the i:h effect, where items nested in testlets, accounts for 4.3% of the total variance with a value of 0.011. The variance component for the i:h effect is composed of the variance components for item (i) main effect, and item*x*testlet (ih) common effect (Brennan, 2001a). The smallness of this value may be interpreted as that the individuals' perceptions about the difficulties of the items in testlets are not very different.

The variance component for the individual x testlet (sh) common effect accounts for 1.8% of the total variance with a value of 0.005. It is clear from Table 4 that this estimated variance value is the smallest value read. The interpretation for this variance value may be that the differences stemming from individual-testlet interaction are small.

The variance component for the residual effect (si:t,e) accounts for 75.6% of the total variance with a value of 0.191. This is the highest value in Table 4 in terms of accounting for the total variance. It is impossible in the nested sx(i:h) design to obtain the individual-item (si) common interaction separately. Therefore, this interaction constitutes a part of residual variance. The fact that the residual variance is so high indicates that the effects could be caused by individual-item common interaction (si), individual-item-testlet common interaction (sit), or that other unexplained sources of variability could be high. On reviewing field literature, it was found that residual variance was estimated to be high in other studies (Hendrickson, 2001; Nalbantoğlu Yılmaz, 2012). This may stem from the fact that the design displays both an unbalanced and a nested structure.

The D Study Findings: Since there were differing numbers of items for each testlet, the D study was conducted by increasing and decreasing the numbers of items in the unbalanced sx(i:h) design and in each testlet. The G and the Phi coefficients obtained are thus shown in Table 5.

As is clear from Table 5, reducing the number of testlets (nh = 3) on the condition that the total number of items ($ni+$ = 20) in the testlets remain constant, varying numbers of items were obtained in each testlet, thus the G coefficient was estimated as 0.749, and the Phi coefficient as 0.671 ($ni+$ = 20, nh = 3, $\check{n}h$ = 2.89). On the condition that the total number of items in testlets ($ni+$ = 20) remains constant, the number of testlets (nh = 8) was increased, thus the differing numbers of items were included in each testlet, and the G coefficient was estimated as 0.766, and the Phi coefficient as 0.728 ($ni+$ = 20, nh = 8, $\check{n}h$ = 7.69). The G and the Phi coefficients derived by reducing the number of testlets decreased by 0.011 in terms of the G coefficient and by 0.040 in terms of the Phi coefficient. The G and the Phi coefficients derived by increasing the number of testlets increased by 0.0051 in terms of the G coefficient and by 0.017 in terms of the Phi coefficient. In conclusion, it was found that decreasing or increasing the number of

Table 4
G Study Results of Unbalanced sx(i:h) Design

Source of Variance	Sum of Squares	*df*	Mean of Squares	Values of Variances	Percentages of Variances
s	1305,54992	1499	0,87095	0,03314	13,13
h	403,12864	5	80,62573	0,01323	5,24
i:h	228,05199	14	16,28943	0,01073	4,25
sh	1540,90457	7495	0,20559	0,00457	1,81
si:h,e	4003,61443	20986	0,19078	0,19078	75,57
Total	7481,24955	29999	98,18248	0,25245	100

Table 5
The Results of the D Study of the Unbalanced sx(i:h) Design

ni+	*ňh*	*ni*:*h*	*ňh*	*G*	Phi
16	3	4, 5, 7	2,84	0,710	0,637
16	6	2, 2, 2, 3, 3, 4	5,57	0,722	0,677
16	6	2, 2, 3, 3, 3, 3	5,82	0,723	0,679
16	8	2, 2, 2, 2, 2, 2, 2, 2	8,00	0,726	0,691
20	3	5, 7, 8	2,89	0,749	0,671
20	**6**	**2, 3, 3, 3, 3, 6**	**5,26**	**0,761**	**0,711**
20	6	2, 3, 3, 4, 4, 4	5,71	0,762	0,715
20	8	2, 2, 2, 2, 3, 3, 3, 3	7,69	0,766	0,728
40	3	12, 13, 15	2,97	0,840	0,750
40	6	4, 4, 6, 6, 9, 11	5,22	0,854	0,797
40	6	3, 5, 7, 7, 9, 9	5,44	0,855	0,800
40	8	2, 2, 4, 4, 6, 6, 8, 8	6,67	0,859	0,811

ni+: number of item; *nh*: number of testlet; *ni*:*h*: number of item in testlets; $\check{n}h$: $n_{i+}^2 / \sum_h n_{i:h}^2$

testlets on the condition that the number of items remained constant did not greatly affect the G and the Phi coefficients.

The G coefficient obtained by reducing the number of items in each testlet and also the total number of items (*ni* = 16) on the condition that the number of testlets (*nh* = 6) remained constant was 0.722; the Phi coefficient estimated thus was 0.677 (*ni*+ = 16, *nh* = 6, *ňh* = 5.57). The G coefficient obtained by increasing the number of items in each testlet and also the total number of items (*ni* = 40) on the condition that the number of testlets (*nh* = 6) remained constant was 0.855; the Phi coefficient estimated thus was 0.800 (*ni*+ = 40, *nh* = 6, *ňh* = 5.44). When the number of testlets stayed constant and the number of items decreased, the G coefficient decreased by 0.039 and the Phi coefficient decreased by 0.034. When the number of items increased, the G coefficient increased by 0.094 and the Phi coefficient increased by 0.089. In conclusion, it was found that decreasing or increasing the number of items in the testlets on the condition that the number of testlets remained constant did not greatly affect the G and the Phi coefficients.

On increasing the number of testlets and the number of items used in the testlets (*ni*+ = 40, *nh* = 8, *ňh* = 6.67), the G coefficient was estimated as 0.0859 and the Phi coefficient as 0.811. Consequently, it was found that in increasing the number of testlets and items in the testlets together there was an increase of 0.098 in the G coefficient and an increase of 0.100 in the Phi coefficient. From this, it may be concluded that increasing the number of both the testlets and the items in the testlets caused an increase in the reliability coefficient.

On reducing the difference between the number of items between testlets (*ni*+ = 20, nh = 6, *ňh* = 5.71) compared to the initial difference by keeping the number of testlets and the total number of items, the G coefficient was estimated as 0.762, and the Phi coefficient as 0.715. Thus, it may be interpreted that reducing the variability between testlets in terms of the number of items in the testlets by keeping the number of testlets and the total number of items in the testlets constant caused an increase in the G and the Phi coefficients. This is also true for when the number of testlets is constant and the total number of items is 16 and 40. When the number of items is 16, the G and the phi coefficients estimated where the variability between testlets is small (*ni*+ = 16, *nh* = 6, *ňh* = 5.82) is higher than those estimated where the variability is high (*ni*+ = 16, *nh* = 6, *ňh* = 5.57). When the number of items is 40, the G and Phi coefficients estimated where the variability between testlets is small (*ni*+ = 40, *nh* = 6, *ňh* = 5.44) is greater than those estimated in case where the variability is big high (*ni*+ = 40, *nh* = 6, *ňh* = 5.22).

In accordance with the findings obtained from the D study, which was performed on the unbalanced sx(i:h) design and for which the results are shown in Table 5, it was observed that increasing the number of testlets and number of items in the testlets influenced reliability in a positive way. Yet, the G and Phi coefficients estimated by keeping the number of items constant and by increasing the number of testlets rose more than the G and the Phi coefficients estimated by keeping the number items constant and increasing the number of items in each testlet. This was an important finding parallel to results obtained found in the literature (Hendrickson, 2001; Lee & Frisbie, 1999). That is to say, the increase in the number of testlets contributed more to reliability than the increase in the number of items in the testlets did.

The local item dependence observed in testlets can increase in parallel to the increase in the number of

items in a testlet. Therefore, it is normally expected that the contributions from an increase in the number of items in a testlet to reliability is lower than those made by an increase in the number of testlets.

Findings Concerning the Third Sub-problem

The DIF-determining study was performed on the items which were considered within the scope of the study and which contained testlets. Since the aim was to determine the effects of tests composed of testlets on DIF, two methods, one of which considered testlet effects and one of which did not consider testlet effects, were used in this study. The methods used were recommended by Fukuhara and Kamata (2011), and were based on IRT. In determining the DIF levels in this study, the measure of 0.426 was considered to be significant at B level whereas the measure of 0.638 was considered to be significant at C level for both methods (Vaughn, 2006). The findings obtained from the DIF analysis performed on the tests through the WinBUGS program based on both considering and without considering the testlet effects are shown in Table 6.

As it is clear from Table 6, significant levels of DIF were determined in five items (items 6, 7, 10, 11, and 14) through the IRT DIF model. Besides, the significant level of DIF determined in those five items, analyses were performed through the bi-factor MIRT DIF model. In addition to that, significant DIF was not estimated for two items (items 8 and 16) in the IRT DIF model. However, significant DIF was estimated for the testlets and analysis was performed through the bi-factor MIRT DIF model. In other words, based on the analysis performed through the bi-factor MIRT DIF model, it was found that the number of DIF items estimated significantly was two more than the number of DIF items estimated significantly with the IRT DIF model.

On examining the DIF levels estimated in the DIF analyses performed on gender basis, it becomes evident that there are differences between the levels obtained through the models used, as is clear in Table 6. The DIF levels estimated for items 6, 7, and 10 were estimated at the C, B, and C levels respectively in both models. On the other hand, the DIF that was estimated at the A level in the IRT DIF model for items 8 and 16 was estimated at the B level in the bi-factor MIRT DIF model. In a similar vein, the DIF estimated at the B level in the IRT DIF model for items 11 and 14 was estimated at the C level in bi-factor MIRT DIF model.

Expert opinion was consulted so as to exhibit the source of DIF in items for which significant DIF was

Table 6
The DIF Magnitudes and Levels Obtained from Using Two Models by Means of Gender

Item	IRT DIF Model (Without considering the testlet effect)			Bi-factor MIRT DIF Model (Considering the testlet effect)		
	DIF Magnitude (β)	DIF Level	Advantageous group	DIF Magnitude (β)	DIF Level	Advantageous group
1	0.0023			-0.01213		
2	0.0168			0.00307		
3	0.0235			0.00883		
4	0.4100			0.36539		
5	0.0326			0.01159		
6	-0.6461 ***	C	Male	-0.79144 ***	C	Male
7	-0.4702 **	B	Male	-0.58357 **	B	Male
8	-0.3497 *	A	Male	-0.45286 **	B	Male
9	-0.0635			-0.08365		
10	0.7096 ***	C	Female	0.82231 ***	C	Female
11	-0.4955 **	B	Male	-0.76745 ***	C	Male
12	-0.0449			-0.07591		
13	-0.2459			-0.29884		
14	-0.5311 **	B	Male	-0.8571 ***	C	Male
15	-0.2145			-0.36132		
16	-0.2923 *	A	Male	-0.44404 **	B	Male
17	0.1441			0.22508		
18	0.1700			0.25345		
19	-0.0386			-0.08705		
20	-0.2393			-0.31315		

*$|\beta| < 0.426$ (A level or no DIF); **$0.426 \leq |\beta| < 0.638$ (B level DIF); ***$|\beta| \geq 0.638$ (C level DIF)

estimated through analyses performed in the bi-factor MIRT DIF model. The following was concluded:

- The reading passage to which items 6, 7, and 8, containing DIF in favor of male students, was related with connection to the Internet, social media and emotions. Current experts understand that there are no distinctions in these issues in relation to gender basis, and therefore it was pointed out by almost all of the experts that those items did not have bias towards male students.
- The reading passage to which item 10, containing DIF in favor of female students, and item 11, containing DIF in favor of male students, was related with machines, industry, chemistry, and engineering. The experts stated that the topics in the text were interesting to the male students. In addition, seven of the experts agreed that because the detail asked in item 10 was related with cloth, it might display bias towards female students. A close examination of the expert opinions concerning item 11, which was in favor of male students, showed that half of the experts contended that the topic of the testlet was within the interest of males and therefore there might be biased towards male students, while the other half disagreed.
- The reading passage to which item 14, containing DIF in favor of male students, was related with mechanical vehicles (such as ferries) and their technical faults. Therefore, it was pointed out by six of the experts that this item might be biased towards men. The remaining four experts said that they did not expect a gender bias in this item.
- The reading passage to which item 16, containing DIF in favor of male students, was related with brain drain, counties' economies, and education. It was stated by the majority of the experts that the topic would not be gender biased.

Discussion and Recommendations

First, the generalizability theory was employed in estimating the reliability of the tests composed of testlets. The reliability of the tests containing testlets was estimated at item and testlet levels where testlet effects were not considered, and for the nested design in which testlet effects were considered. Of those values estimated, the reliability coefficient obtained at the item level was higher than the other coefficients. The G coefficient estimated in one-facet designs in the G theory is found to be equal to Cronbach alpha (Brennan, 2001a). In consequence, when the testlet effects are not considered in tests containing testlets, a biased result, which is equal to Cronbach alpha, is obtained. When the testlet effects are not considered, it is also possible to estimate a reliability value so high as to reach a value when the number of items is doubled (Wainer, Sireci, & Thissen, 1991).

It is very common to encounter the use of testlets in tests owing to their several advantages in the test development process. On the basis of the results obtained in this study, it was concluded that an increase in the number of testlets was more influential than an increase in the number of items constituting the testlet for the purposes of raising reliability. This is supported by other G studies performed within the framework of the G theory of this study (see Table 4). For example, the effect of testlets on the total variance is greater than the one obtained by the nesting of the items in testlets.

The variation in the number of items in testlets in a test influences the reliability obtainable from the test results. If the number of items constituting the testlets differs greatly from one testlet to another within a research study, it affects reliability in a negative way.

It was found that the number of items, which was determined to display DIF at significant levels through the DIF-determining method, considering the testlet effects was greater than the number of items which was determined through the method not considering the testlet effect.

Since testlets violate the assumption of local item dependence, the DIF that was estimated through methods based on IRT was estimated as smaller than the actual. Moreover, on reviewing the DIF studies available in literature, it was concluded that the DIF was found to be trivial at the item level with analyses without considering the testlet effects and became clear with analyses considering the testlet effects (Fukuhara, 2009; Fukuhara & Kamata, 2011; Sedivy, 2009; Wainer et al., 1991). This does not lead to serious differences when there is large enough data sample ; but as the sample and the DIF becomes smaller, the methods considering the testlet effects displays better performance.

In conclusion, the increase in the frequency of the use of testlets in test development activities makes it obligatory to analyze those instruments through right statistical means in terms of accurately interpreting the results to be obtained from those tests.

Recommendations Based on Conclusions

- In estimating the reliability of the tests containing testlets, the estimation methods that take local item dependence into consideration should be considered in order to reach unbiased estimations.
- Since increasing the number of testlets instead of increasing the number of items in testlets is more effective, this may be preferred in order to achieve more reliable measurements of tests composed of testlets. In other words, when increasing the number of testlets prepared, depending on a common stimulus is more influential.
- So as to increase the reliability of the tests containing testlets, the closeness of the number of items from one testlet to another should be assured.

Recommendations to Researchers

- The method based on the bi-factor MIRT DIF model, which may be used in estimating the DIF in tests composed of testlets, can also be used in other tests. Many cases in education are capable of leading to local item dependence (Yen, 1993). For instance, a standardized science test is composed of such sub-themes as physics, chemistry and biology, and in such a test, the sub-themes can be grouped together and be considered as testlets. The methods and analyses used in this study can be employed in determining the DIF.
- The fact that this research was performed with a data set, and thus with a small number of items, can be thought of as a possible restriction. Yet, in research studies conducted in relation to the effects of testlets on reliability (Hendrickson, 2001; Lee & Frisbie, 1997, 1999; Sireci et al., 1991; Wainer, 1995) and on differential item functioning (Fukuhara & Kamata, 2011; Wainer, 1995) similar cases were encountered. Performing research with different data sets (composed of just testlets or independent items and testlets together) and with data sets containing more items will be beneficial in obtaining more reliable results.
- Analyses were done in this research by using real data. Simulation can be done in order to evaluate the testlet effects in differing measurement situations in more details.
- To conclude, targets should be set for tests containing testlets to achieve more unbiased results through methods which do not ignore cases of local item dependence while estimating reliability and validity. These issues have been considered in this research.

References

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

Bloch, R., & Norman, G. (2011). *G-String 4 user manual (Version 6.1.1)*. Hamilton, Ontario, Canada: Ralph Bloch & Geoff Norman.

Bradlow, E. T., Wainer, H., & Wang, X. (1999). A Bayesian random effects model for testlets. *Psychometrika*, *64*, 153-168.

Brennan, R. L. (2001a). *Generalizability theory*. New York, NY: Springer-Verlag.

Brennan, R. L. (2001b). *Manual for urGENOVA (Version 2.1)* (Iowa Testing Programs Occasional Paper Number 49). Iowa City, IA: Iowa Testing Programs, University of Iowa.

Clauser, B. E., & Mazor, K. M. (1998). Using statistical procedures to identify differentially functioning test items. *Educational Measurement: Issues and Practice*, *17*(1), 31-47.

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.

Fukuhara, H. (2009). *A differential item functioning model for testlet-based items using a bı-factor multidimensional item response theory model: a Bayesian approach* (Doctoral dissertation, Florida State University College of Education). Retrieved from http://diginole.lib.fsu.edu/cgi/viewcontent.cgi?article=1573&context=etd

Fukuhara, H., & Kamata, A. (2007, November). *DIF detection in a presence of locally dependent items*. Paper presented at the annual meeting of the Florida Educational Research Association, Tampa.

Fukuhara, H., & Kamata, A. (2011). A bi-factor multidimensional item response theory model for differential item functioning analysis on testlet-based items. *Applied Psychological Measurement*, *35*(8), 604-622.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. California, CA: Sage.

Hendrickson, A. B. (2001, April). *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, Seattle.

Lee, G., Dunbar, S. B., & Frisbie, D. A. (2001). The relative appropriateness of eight measurement models for analyzing scores from test composed of testlets. *Educational and Psychological Measurement*, *61*(6), 958-975.

Lee, G., & Frisbie, D. A. (1997, March). *A generalizability approach to evaluating the reliability of testlet-based test scores*. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago.

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 test composed of testlets. *Applied Psychological Measurement*, *25*, 357-372.

Lee, G., & Park, I. (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.

Li, Y., Bolt, D. M., & Fu, J. (2006). A comparison of alternative models for testlets. *Applied Psychological Measurement*, *30*(1), 3-21.

Nalbantoğlu Yılmaz, F. (2012). *Genellenebilirlik kuramında dengelenmiş ve dengelenmemiş desenlerin karşılaştırılması* (Doctoral dissertation, Ankara University, Turkey). Retrieved from https://tez.yok.gov.tr/UlusalTezMerkezi

Nalbantoğlu Yılmaz, F., & Uzun Başusta, B. (2012, September). *Genellenebilirlik kuramıyla dikiş atma ve alma becerileri istasyonu güvenirliğinin değerlendirilmesi*. Paper presented at III. Ulusal Eğitimde ve Psikolojide Ölçme ve Değerlendirme Kongresi, Bolu, Turkey.

Özçelik, D. A. (2010). *Test hazırlama kılavuzu*. Ankara: Pegem Akademi.

Sedivy, S. K. (2009). *Using traditional methods to detect differential item functioning in testlet data* (Doctoral dissertation, University of Wisconsin-Milwaukee). Retrieved from http://gradworks.umi.com/33/73/3373884.html

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer*. Newbury Park, CA: Sage.

Sireci, S. G., Thissen, D., & Wainer, H. (1991). On the reliability of testlet-based tests. *Journal of Educational Measurement, 28*, 237-247.

Spiegelhalter, D., Thomas, A., & Best, N. (2003). *WinBUGS 1.4*. Cambridge, UK: MRC Biostatistics Unit, Institute of Public Health.

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.

Vaughn, B. K. (2006). *A hierarchical generalized linear model of random differential item functioning for polytomous items: A Bayesian multilevel approach* (Doctoral dissertation, The Florida State University College of Education). Retrieved from http://diginole.lib.fsu.edu/cgi/viewcontent.cgi?article=5595&context=etd

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*, 157-186.

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 & C. A. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 245-269). Dordrecht: Kluwer Academic Publishers.

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., Sireci, S. G., & Thissen, D. (1991). Differential testlet functioning: definitions and detection. *Journal of Educational Measurement*, *28*(3), 197-219.

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.

Wainer, H., & Wang, C. (2000). Using a new statistical model for testlets to score TOEFL. *Journal of Educational Measurement*, *37*, 203-220.

Wang, W. C., & Wilson, M. (2005). Assessment of differential item functioning in testlet-based items using the Rasch testlet model. *Educational and Psychological Measurement*, *65*(4), 549-576.

Wang, X., Bradlow, E. T. & Wainer, H. (2002). A general Bayesian model for testlets: theory and application. *Applied Psychological Measurement*, 26(1), 109-128.

Yen, W. M. (1993). Scaling performance assessment: strategies for managing local item dependence. *Journal of Educational Measurement, 30*, 187-213.