International Journal of Assessment Tools in Education
2021, Vol. 8, No. 3, 583–595
https://doi.org/10.21449/ijate.948677
Published at https://ijate.net/ https://dergipark.org.tr/en/pub/ijate **Research Article**

Comparison of G and Phi coefficients estimated in generalizability theory with real cases

Kaan Zulfikar Deniz [1,*], **Emel Ilican** [2]

[1]Ankara University, Faculty of Educational Sciences, Ankara, Turkey
[2]Republic of Turkey Ministry of National Education, Ankara, Turkey

ARTICLE HISTORY
Received: Oct.03, 2020
Revised: Jan. 16, 2021
Accepted: May 31, 2021

Keywords:
Reliability,
Generalizability theory,
Decision study,
Item difficulty index.

Abstract: This study aims to compare the G and Phi coefficients as estimated by D studies for a measurement tool with the G and Phi coefficients obtained from real cases in which items of differing difficulty levels were added and also to determine the conditions under which the D studies estimated reliability coefficients closer to reality. The study group for this research consisted of 80 seventh-grade students from various public and private secondary schools in the provinces of Ankara, Istanbul, and Adana in Turkey. Four raters who served as Turkish teachers in various public secondary schools in Ankara were included in this study. A data collection tool consisting of 12 tasks was prepared to measure the participating seventh grade students' written expression skills in Turkish. The equation of the G and Phi coefficients estimated in the D study and obtained through the real cases was observed only when six tasks with item difficulty indexes close to the mean difficulty of the test were added in such a way that the mean difficulty of the test never changed. In other cases, where the mean difficulty of the test changed because of the addition of easy or difficult tasks, it was determined that the reliability coefficients estimated in the D study and obtained in real cases were similar, but they had different values.

1. INTRODUCTION

The most important psychometric properties sought in a measurement tool are grouped under the concepts of reliability, validity, and usability. Reliability is defined as the ability to repeat measurements of a feature performed on the same individuals with the same measurement tool under similar conditions or to give consistent results (Baykul, 2015; Crocker & Algina, 1986; Nitko, 2004). According to the Classical Test Theory (CTT), reliability coefficient is to be estimated regarding reliability. While making this estimation, the effect of variable situations such as the content, construct and application of items and tests on test scores is examined using various reliability estimation methods (Aiken, 2009; Anastasi & Urbina, 1997).

In some cases, when utilising reliability estimation methods that are based on CTT, any single application of the CTT model cannot clearly differentiate among multiple sources of error. To find a solution to the limitations of CTT, the Generalizability theory (G) was developed, which

*CONTACT: Kaan Zulfikar Deniz zlfkrdnz@yahoo.com Ankara University, Faculty of Educational Sciences, Ankara, Turkey

e-ISSN: 2148-7456 /© **IJATE** 2021

allows for the calculation of reliability coefficients based on differing sources of variation (i.e., error) that may occur within a single study. G theory liberalizes classical theory by employing ANOVA methods that allow an investigator to untangle multiple sources of error (Brennan, 2001).

As a result, with G theory studies, any facet (i.e., source) of error such as rater, time, forms, and/or item is evaluated simultaneously and as a group in order to estimate a comprehensive and single reliability coefficient. The basic idea of G theory is that error variance derives from different sources of variability as well as from the interactions that take place between them. In other words, the superiority of G theory over CTT is that different error sources can be simultaneously estimated through a single analysis. This process is completed with the help of variance analysis that allows for multiple variance sources to be analysed through a single analysis, while at the same time a determination can be made regarding the size of each variance source (Brennan, 2001; Shavelson & Webb, 1991).

Also, G theory allows for the calculation of two differing reliability coefficients regarding both relative decisions; namely, those decisions based on individual performance and the absolute decisions of these individual performances. As a result, these are the generalizability coefficients that make up the relative evaluations and the Phi (Φ) coefficient for the absolute evaluations. Importantly, generalizability (G) and decision (D) studies are carried out in order to determine the reliability coefficients utilising G theory. Through the G study process, the variance components of scores and the interactions between them are estimated simultaneously through ANOVA. These estimated variance components are then utilised in the subsequent step of the D study. In a D study, in order to create measurement situations with sufficient reliability, measurements are organized so that the measurement error can be minimised (Brennan, 2001; Shavelson & Webb, 1991).

To explain, a D study is an estimation, use, and interpretation of variance components in order to formulate decisions according to already well-defined measurement processes (Crocker & Algina, 1986). For example, in a case where more than one rater scores a group of students` ability to solve mathematical problems, a G study that utilises three raters and 20 items is followed by a D study; as a result, differing numbers of raters and differing numbers of items can be estimated and through this process the G and Phi coefficients can also be estimated. However, in the results of the D study, the G and Phi coefficients are provided when adding or subtracting items from the measurement tool, yet no information is given in regard to the difficulty of these items. For example, in a D study, the G and Phi coefficients are estimated after at least three items have been added to a measurement tool, but to what extent these coefficients are sensitive to the item difficulty index (p_j) of the added items remains unknown, and whether the items are easy or difficult also remains undefined.

In the literature, there are many studies in which items related to various measurement and evaluation practices have been considered a source of variability and reliability studies based on G Theory (Choi & Wilson, 2018; Çakıcı Eser & Gelbal, 2013; Deliceoğlu & Çıkrıkçı Demirtaşlı, 2012; Demir, 2016; Doğan & Anadol, 2017; Doğan & Bıkmaz Bilgen, 2017; Güler, 2011; Güler et al., 2014; Gülle et al., 2018; Hathcoat & Penn, 2012; Hill et al., 2012; Scherbaum et al., 2018; Solano-Flores & Li, 2013; Yılmaz Nalbantoğlu & Gelbal, 2011). Furthermore, in some of these studies (Doğan & Anadol, 2017; Scherbaum et al., 2018; Yılmaz Nalbantoğlu & Gelbal, 2011) comparisons were also made regarding the use of crossed and nested research designs within the scope of G theory. In other studies (Doğan & Bıkmaz Bilgen, 2017; Güler et al., 2014; Gülle et al., 2018; Hathcoat & Penn, 2012; Solano-Flores & Li, 2013) it was observed whether the reliability of performance-based measures could be examined through G theory. In addition, there are several studies (Çakıcı Eser & Gelbal, 2013; Deliceoğlu & Çıkrıkçı Demirtaşlı, 2012; Demir, 2016; Güler, 2011) in which the reliability of measurements was examined through methods other than G theory. Apart from these studies, there are few studies in which the G and Phi coefficients estimated through a D study were compared with the reliability coefficients in real cases. Atılgan and Tezbaşaran (2005) compared the G and Phi coefficients acquired from D studies and real situations from a number of different raters by using data from two successive years of special skill selection exams conducted from a student selection program. In another study, the G and Phi coefficients estimated for two, three, and four raters from real cases in which it was not possible to randomly select raters from a population universe, were compared with the results from relevant D studies (Kamış & Doğan, 2017). However, there was no identified study that compared the predicted G and Phi coefficients in the D studies as well as the obtained G and Phi coefficients from real cases in which there were items of varying difficulty levels added and/or removed from the measuring tool.

While test items are considered as a source of variability and reliability in which studies based on G theory have been carried out, there can be a determination made to change the number of test items in order to obtain the reliability coefficients that have previously been predicted in the D study. At this stage, it is believed that knowing the difficulty level of items and under which conditions the D study accurately estimates the reliability coefficients in real cases will ultimately contribute to a more meaningful interpretation of D studies. In addition, this information is expected to facilitate the selection of items as a way of obtaining reliability coefficients as estimated in the D study as well as supporting the efficient completion of reliability studies.

As a result, the aim of this study was to compare the G and Phi coefficients as estimated by D studies as well as the G and Phi coefficients obtained in real cases in which the items of differing difficulty levels were added and to also determine the conditions under which the D studies estimated the reliability coefficients more in line with the real situation. In this respect, easy, moderate or difficult items were added to a measuring tool and these additional items were meant to reflect two conditions, both modifying and not-modifying the mean difficulty of the test. The sub-objectives determined for the general purpose of this study are as follows:

a) To compare the G and Phi coefficients estimated by the D studies and the G and Phi coefficients obtained by increasing the total number of tasks to 18 that change the mean difficulty of the test: with six easy tasks; with six moderate tasks; and with six difficult tasks.

b) To compare the G and Phi coefficients estimated by the D studies and the G and Phi coefficients obtained by increasing the total number of tasks to 18 that did not change the mean difficulty of the test: with two easy, two moderate, and two difficult tasks; and with six moderate tasks.

c) To determine whether there were any significant differences between the G and Phi coefficients estimated by D studies and the G and Phi coefficients obtained in various real cases, where the total number of tasks was increased to 18.

2. METHOD

This section indicates the research design used in the study, the study group, the data collection, and the analysis of the data.

2.1. Research Design

This study followed a survey research model in which attempts were made to define a situation under a set of circumstances without changing and/or influencing that situation in any way. In addition, since this research was aimed at generating information, it was prepared and carried out in a basic manner (Büyüköztürk et al., 2012; Fraenkel et al., 2015; Karasar, 2016).

2.2. Study Group

The study group for this research consisted of 80 seventh grade students (ages 12-13) studying in various public and private secondary schools located in Ankara (n=30, 37.5%), Istanbul (n=25, 31.5%), and Adana (n=25, 31.5%), Turkey during the 2016-17 academic year. Of the students in the study group, 34 (42.5%) were male and 46 (57.5%) were female. Students for the study were selected from 26 schools, 20 of which were public and 6 were private. The study group of the research was selected from the sample of a study conducted by the Republic of Turkey Ministry of National Education that aimed to evaluate the Turkish written expression skills of students from various grade levels. The students who were applied one of the seventh grade test forms used in the study and raters assigned for item scoring were included in the study group of this research study. Four raters worked as Turkish language teachers in various public secondary schools. These teachers had previously received training on item scoring and were also informed about the use of rubrics prepared for this study. Importantly, the students and teachers included in this study group were selected from different schools.

2.3. Data Collection

The data of this research were obtained from the Ministry of National Education by official correspondence for research permission. In the data collection process of the study, the students and raters from the study group were briefly informed about the study process. A skill test consisting of 12 tasks was first prepared and then applied in order to measure the students` Turkish written expression skills. Then, four raters scored the skill test independently and the data were collected for analysis. Through the application of student tasks, each student answered the same 12 tasks and the four raters via a scoring rubric prepared for the test scored each student's responses. Thus, the research design for this study can be considered to follow a fully crossed (sxtxr) design.

2.3.1. *Data collection tool*

The test utilised in this study consisted of 12 tasks prepared to measure the Turkish written expression skills of seventh grade students. In completing the tasks included in the test, the participating students had to create sentences and paragraphs with a variety of characteristics. In the first task of sentence knowledge, the students were asked to select at least five words from a word pool provided and then form a sentence consisting of a minimum of eight words in total. In the second task, these students were asked to form a sentence consisting of a minimum of eight words in accordance with a visual prompt. In the third task, a dialogue was provided to the students and they were asked to complete the dialogue with an appropriate sentence consisting of at least five words. The subsequent four tasks of the test were related to a persuasion paragraph and then the remaining final five tasks involved writing a petition. Rubrics that can be scored from 0 to 4 were developed for each task of the test in this study. The experts in the study team formed by the Ministry of National Education developed these rubrics. As a result, the highest score a student could receive from task scoring was 48 and the lowest possible score was 0.

2.4. Data Generation and Analysis

In this study initially, variance sources were estimated from the G study of 12 tasks. Then, D study was conducted by using these variance sources and increasing the number of tasks to 18. The G and Phi coefficients were estimated for 18 tasks within the test through the D study. These coefficients were compared with the reliability coefficients estimated from the real situation of 18 tasks subsequent to adding tasks of various difficulty indexes, which ultimately changed the test`s mean difficulty for some of the cases but not all. Since all of the 12 tasks initially included in the scale were rated at a moderate level of difficulty, there were randomly selected tasks from the scale that were reused by adding moderate tasks to the test. The easy tasks added to the test were produced by increasing the points of the easiest tasks in the test by two points each except for those with full points. The difficult tasks added to the test were artificially created by dividing the scores of the most difficult tasks in the test into three and then decreasing the scores downward. Finally, the significance of the differences between the estimated G and Phi coefficients as well as the G and Phi coefficients obtained in various real cases was examined through a Fisher`s z' test. Variance sources and the G and Phi coefficients were estimated in the analysis performed through crossed design (sxtxr) obtained by grading 80 students by four raters for 12 tasks. The EduG 6.1-e program was utilised in analysing the data obtained from this study.

3. RESULTS

In the results section, first, those results related to the estimated variance of the sources of variability from the fully crossed design are provided for different cases where the number of tasks was either 12 or 18. Second, in accordance with the sub-objectives of this study, the findings from the comparison of the G and Phi coefficients estimated through the D study as well as the G and Phi coefficients obtained in real cases were interpreted. In regards to the analysis findings, results related to the estimated variance components are provided in Table 1.

Table 1. *Analysis of variance results and variance component estimates for students, tasks, raters, and their interactions.*

	Number of Tasks	Source of Variance	df	MS	Variance Component Estimates	Percentage of Total Variance Estimates
Actual status	12	s	79	97.63	1.66	22.50
		t	11	70.75	0.10	1.40
		r	3	278.77	0.25	3.40
		st	869	11.67	2.39	**32.40**
		sr	237	8.43	0.53	7.10
		tr	33	28.70	0.33	4.50
		str	2607	2.11	2.11	28.60
The mean difficulty of the test changes	18 (Six easy tasks added)	s	79	142.62	1.69	21.20
		t	17	502.85	1.48	18.60
		r	3	394.57	0.25	3.20
		st	1343	9.06	1.80	22.60
		sr	237	13.63	0.65	8.20
		tr	51	21.19	0.24	3.00
		str	4029	1.84	1.84	**23.10**
	18 (Six moderate tasks added)	s	79	187.59	2.28	26.30
		t	17	92.97	0.16	1.90
		r	3	306.76	0.19	2.20
		st	1343	14.84	3.21	**37.10**
		sr	237	10.85	0.49	5.70
		tr	51	27.98	0.32	3.80
		str	4029	2.00	2.00	23.10
	18 (Six difficult tasks added)	s	79	84.87	0.98	13.30
		t	17	747.30	2.24	**30.50**
		r	3	207.02	0.12	1.70
		st	1343	9.05	1.87	25.50

Table 1. *Continued.*

		sr	237	7.06	0.31	4.20
		tr	51	23.75	0.28	3.80
		str	4029	1.55	1.55	21.10
The mean difficulty of the test remains unchanged	18 (Six moderate tasks added)	s	79	149.38	1.77	23.80
		t	17	65.58	0.09	1.20
		r	3	405.34	0.26	3.40
		st	1343	11.79	2.45	**32.90**
		sr	237	12.09	0.56	7.50
		tr	51	27.01	0.31	4.20
		str	4029	2.01	2.01	27.00
	18 (Six easy, moderate and difficult tasks added)	s	79	134.02	1.59	19.50
		t	17	602.77	1.77	21.70
		r	3	457.06	0.29	3.50
		st	1343	9.00	1.80	**22.10**
		sr	237	12.24	0.58	7.10
		tr	51	29.69	0.35	4.30
		str	4029	1.78	1.78	21.80

Table 1 illustrates that in a majority of the cases studied; the ST interactive variance component value had the highest rate of total variance. Accordingly, it can be stated that the difficulty levels of the tasks differed from one student to another in the cases examined. In addition, when six difficult tasks were added to the test and the mean difficulty of the test decreased, it was determined that instead of ST, the T variance component value (2.24) had the highest rate (30.5%) in the total variance. Thus, after adding difficult tasks, it can be said that the tasks in the test become very different from each other in terms of their difficulty level. Among all the cases examined, it was observed that when six items with moderate difficulty were added to the test and the average difficulty of the test varied, the ST-interactive variance component value (3.21) was found to have the highest value. Here, it was the source of variability that explained the total variance with the highest rate (37.1%). The second highest rate in total variance generally belongs to residual component. Accordingly, it can be said that there is interaction between students, tasks, and raters and there are systematic or unsystematic sources of variability that cannot be measured in this study. In these cases, the variance component for students was generally high in total variance. This result demonstrated that the measured characteristics of students differed from each other; as a result, the measurement process proved successful in distinguishing students from one another according to their results. Finally, in all of the cases, it can be stated that the raters generally provided consistent scores because the overall rater ratio variance in total was negligible.

Table 2 illustrates the G and Phi coefficients obtained when the number of tasks in the test was actually 12 and then an estimate was produced for 18 tasks in the D study.

Table 2. *D study results.*

Number of Tasks	G	Phi
12	0.82	0.79
18	0.85	0.82

Table 2 displays that the G and Phi coefficients obtained from real cases where the number of tasks in the test was 12 were 0.82 and 0.79. Furthermore, according to the results of the decision study, in which the number of tasks was 18, the G and Phi coefficients were 0.85 and 0.82.

3.1. Results for the First Sub-Objective

The mean difficulty of the test and the G and Phi coefficients obtained in the G and D studies from the cases where the number of test taks was increased to 18 and the test mean difficulty changed are provided in Table 3.

Table 3. *G and phi coefficients obtained in cases where the test mean difficulty changed.*

Number of Tasks	Actual Status			Decision Studies (estimated for 12 tasks)	
	Mean difficulty of the test	G	Phi	G	Phi
12[a]	0.51	0.82	0.79	-	-
18[b]	0.60	0.85	0.79		
18[c]	0.48	0.87	0.85	0.85	0.82
18[d]	0.38	0.83	0.73		

[a]Original scale
[b]Added six easy tasks
[c]Added six moderate tasks
[d]Added six difficult tasks

As can be seen in Table 3, when six tasks of moderate difficulty (p_j = 0.41-0.58) were added and the mean difficulty of the test was least varied, the G and Phi coefficients were 0.87 and 0.85 for the first case. In addition, when the test had 12 tasks, the G and Phi coefficients estimated for the 18 tasks within the D study were 0.85 and 0.82. As a result, it can be stated that the G and Phi coefficients estimated for the 18 tasks from the D study were relatively smaller than those obtained in the real case where six moderate tasks had been added to the test.

The G and Phi coefficients were 0.85 and 0.79 for the second case where six easy tasks (p_j = 0.76-0.80) were added to the test and the test mean difficulty had changed more than the first case. Through the analysis results it was recognised that the G coefficient estimated in the D study for 18 tasks was equal to the G coefficient obtained in the real case where six easy tasks had been added to the test. Also, the Phi coefficient obtained after adding easy tasks to the test was less than the estimated Phi coefficient (0.82) from the D study with 18 tasks.

Finally, when six difficult tasks (p_j = 0.12-0.13) were added to the test, it was recognised that the mean difficulty of the test decreased/increased considerably compared to the first two cases. In this case, the G and Phi coefficients were acquired as 0.83 and 0.73 for the real situation in which difficult tasks had been added to the test, and as a result, the values were smaller than the G and Phi coefficients estimated in the D study for 18 tasks. In addition, these values (G = 0.83 and Phi = 0.73) differed from the G (0.85) and Phi (0.82) coefficients estimated in the D study as compared to the other two cases where the mean difficulty of the test had changed less.

3.2. Results for the Second Sub-Objective

The mean difficulty of the test and G and Phi coefficients obtained from G and D studies where cases that had the number of test tasks increased to 18 and the test mean difficulty did not change are provided in Table 4.

Table 4. *G and phi coefficients obtained in cases where the mean difficulty of the test did not change.*

Number of Tasks	Actual Status			Decision Studies (estimated for 12 items)	
	Mean difficulty of the test	G	Phi	G	Phi
12[a]	0.51	0.82	0.79	-	-
18[b]	0.51	0.85	0.78	0.85	0.82
18[c]	0.51	0.85	0.82		

[a]Original scale
[b]Two of the six tasks added were easy, two were moderate and two were difficult.
[c]Added six moderate tasks

As Table 4 presents the G and Phi coefficients were 0.85 and 0.78 in the first case when two easy (p_j = 0.78 and p_j = 0.80), two moderate (p_j = 0.58), and two difficult tasks (p_j = 0.12 and p_j = 0.13) were added to the test and the mean difficulty of the test (p_j = 0.51) remained unchanged. Very close to these values, next, in the second case the values remained close with the G and Phi coefficients obtained at 0.85 and 0.82 when six moderate tasks (p_j = 0.41-0.58) were added and the mean difficulty (p_j = 0.51) of the test again remained unchanged. As a result, the G coefficients acquired in both real cases were found to be equal to the G coefficient that had been estimated in the D study for the 18 tasks. In addition, the Phi coefficient (0.78) obtained in the first case was less than the Phi coefficient (0.82) estimated in the D study for the 18 tasks. Importantly, among all of the cases examined, only within the second case was the obtained G (0.85) and Phi (0.82) coefficient equal to the G (0.85) and Phi (0.82) coefficient estimated in the decision study for 18 tasks.

3.3. Results for the Third Sub-Objective

In order to determine whether the differences between the G and Phi coefficients estimated by the D studies in this research and those obtained through real cases were significant, all of the G and Phi coefficients were converted to z scores through the Fisher Z-transformation test. Accordingly, the G and Phi coefficients obtained and as well as the Fisher`s z' scores calculated are provided in Table 5.

Table 5. *G and phi coefficients obtained in cases where the number of test tasks were 12 or 18 including the Fisher z' scores.*

		Decision Studies (estimated for 12 tasks)		Actual Status	
	Number of tasks	G	Phi	G (Fisher z')	Phi (Fisher z')
The mean difficulty of the test changes	18[a]	0.85	0.82	0.85 (0)	0.79 (0.20)
	18[b]			0.87 (-0.18)	0.85 (-0.23)
	18[c]			0.83 (0.16)	0.73 (0.54)
The mean difficulty of the test remains unchanged	18[d]	0.85	0.82	0.85 (0)	0.78 (0.26)
	18[e]			0.85 (0)	0.82 (0)

[a]Added six easy tasks
[b]Added six moderate tasks
[c]Added six difficult tasks
[d]Two of the six tasks added were easy, two were moderate and two were difficult
[e]Added six moderate tasks

When the Fisher`s z' test results provided in Table 5 are examined, it can be recognised that all of the z' values calculated were between -1.96 and +1.96 (Kenny, 1987). As a result, this finding shows that there was not a significant difference between the reliability coefficients estimated in the D studies and those obtained in real cases.

4. DISCUSSION and CONCLUSION

As a result of the analyses conducted in this study, it was observed that the reliability coefficients predicted in the D studies and those obtained in real cases were different; however, in general, they remained quite close to each other. When the differentiated Phi coefficients were examined, it was also found that the values estimated in the D studies and obtained through real cases were different for four of the five cases examined. Next, the values obtained in real cases for the G coefficient were equal to the estimated G coefficient from the D study in three of the five cases studied. As a result, it can be stated that the reliability coefficients in the case where items were added to the measurement tool and estimated through the D studies, the Phi coefficient was more sensitive to the difficulty level of the added items in comparison to the G coefficient. This result is thought to be related to the fact that the item variance considered when calculating the Phi coefficient increased more than the G coefficient when the easy or difficult items were added to the measuring instrument (Brennan, 2001). In this study, it was observed that the task variance, which has the smallest value in the total variance in the real situation, increases when tasks with different difficulty levels are added to the test. Added easy or difficult tasks caused the Phi coefficient to decrease as the task variance and absolute error variance increased. As a result, although it was estimated that the Phi coefficient would increase if the number of tasks was increased from 12 to 18, it was instead recognised that the Phi coefficient did not change and/or decrease from the addition of either any easy and/or difficult tasks to the test. Furthermore, the relative error variance utilised in determining the G coefficient was acquired with the interactive variance components that included the students and was ultimately less affected by the change in variance that arose from the test tasks and was generally close in value to those predicted in the D studies (Güler et al., 2012). When the literature for this study was examined, it was determined that there were findings which increased the number of items that had a positive effect of ensuring the desired quality of reliability as well as that reliability would increase as the number of items increased (Ankenmann & Stone, 1992; Bıkmaz Bilgen & Doğan, 2017; Güler & Yetim, 2008; Hulin et al., 1982; Tavşancıl, 2005). In other previous studies, it was concluded that low reliability was in effect due to the low number of substances (i.e., items) included (Güler & Yetim, 2008; Kaya, 2005). This is important because in research where D studies were conducted based on G theory, it was concluded that reliability would increase if the number of items in the test increased (Deliceoğlu & Çıkrıkçı Demirtaşlı, 2012; Demir, 2016; Doğan & Bıkmaz Bilgen, 2017; Gülle et al., 2018; Hathcoat & Penn, 2012). However, as was determined in this study, an increase in the number of items may in effect not actually provide a higher reliability coefficient in all cases. Similarly, the research study by Giray and Şahin (2012) revealed that solely reducing the number of items did not in itself lead to a decrease in reliability.

In this present study, the equality of both the G and Phi coefficients obtained in the real situation as well as estimated in the D study could only be witnessed when six tasks of moderate difficulty (p_j=0.41-0.58) were added to the test but did not change the mean difficulty of the test (p_j=0.51). In addition, it was also determined that the difference between reliability coefficients, especially the Phi coefficient, which was obtained for the real cases and estimated in the D study, increased more when the mean difficulty of the test changed as a result of adding items. Accordingly, it can be stated that when the reliability coefficients estimated in the D study from the addition of items to the test were expected to be obtained in a real case, it would be beneficial in future research to select items that do not change the mean difficulty of the test or items with the difficulty indexes closest to the mean difficulty of the existing test. On the other hand, it was also determined that there were no significant differences between the G and Phi coefficients obtained in various situations when the number of tasks in the test was actually 18 and the G and Phi coefficients estimated as a result of the D studies were made with 12 tasks. However, it is recommended that this situation be re-examined by utilising different measurement tools when added items can be changed in the mean difficulty of the test. In addition, it can be stated that these examinations may be useful for a test where the percentage of item variance in the total variance is greater. This is recommended because the G studies conducted in this study generally showed that item variance made up a small percentage of the total variance. While, in studies by Demirel and Epçaçan (2012) and Katrancı and Yangın (2003), very easy and very difficult items were removed from the test for a similar purpose, and as a result, sufficient KR-20 reliability coefficients were obtained. Similarly, for other studies (Çakır & Aldemir, 2011; Kaplan & Duran, 2016), some test items were excluded in order to obtain a higher reliability coefficient, but unfortunately no information was provided regarding the item difficulty index of the extracted items.

Also, previous research studies have pointed out that in decision studies with G theory the G and Phi coefficients will increase if the number of items and raters are increased (Güler et al., 2012). However, in this current study, it was determined that the Phi coefficient could remain the same or even decrease if easy or difficult tasks were added to a moderate scale. As a result, it was recognised that the Phi coefficients obtained from the real case where the number of items was increased might be smaller than the estimated Phi coefficients for the number of items in the D study. In addition, Kamış and Doğan (2017) revealed that even though the number of raters increased in their study, the reliability coefficients could possibly decrease and could even be lower values than the predicted values from the related D studies. Furthermore, Atılgan and Tezbaşaran (2005) determined that the reliability coefficients obtained in real cases where the number of raters was increased were smaller than the reliability coefficients predicted within the D studies. In this current study, it is discovered that if the number of tasks increased, the G and Phi coefficients obtained for real situations may be larger, smaller, or equal to the G and Phi coefficients estimated in the D studies. As a result of this finding, it is believed that the difference between the findings of the two previous studies may be a result of whether or not the items/raters have been randomly selected from the population universe or the ratio of the item/rater variance in regards to the total variance within the study.

Finally, the significant findings of this study may show that since the reliability of real situations cannot be estimated completely and/or systematically through the utilisation of D studies in G theory, then it is recommended that these factors be taken into consideration when interpreting the results of future D studies. Since the scores on easy and/or difficult tasks were artificially produced in this study, future researchers are recommended that they perform similar studies utilising a real pool of items, in which the easy or difficult items can be added to a test at any point and with no concern of its effect on the outcome and/or validity of the test.

Declaration of Conflicting Interests and Ethics

The authors declare no conflict of interest. This research study complies with research publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the author(s).

Authorship Contribution Statement

Kaan Zülfikar Deniz: Investigation, research design, literature review, supervision and writing the manuscript. **Emel Ilican**: Research design, literature review, methodology, data collection, data analysis, and writing the manuscript.

ORCID
Kaan Zülfikar Deniz https://orcid.org/0000-0003-0920-538X
Emel Ilican https://orcid.org/0000-0003-4244-6441

5. REFERENCES

Aiken, L., R. (2009). *Psychological testing and assessment* (Twelfth ed.). Pearson.

Anastasi, A., & Urbina, S. (1997). *Psychological testing.* Pearson.

Ankenmann, R. D., & Stone, C. A. (1992, April). *A monte carlo study of marginal maximum likelihood parameter estimates fort he graded model*. Paper presented at the Annual Meeting of the Council on Measurement in Education, San Francisco, CA.

Atılgan, H., & Tezbaşaran, A. A. (2005). Genellenebilirlik kuramı alternatif karar çalışmaları ile senaryolar ve gerçek durumlar için elde edilen G ve Phi katsayılarının tutarlılığının incelenmesi [An investigation on consistency of G and Phi coefficients obtained by generalizability theory alternative decisions study for scenarios and actual cases]. *Eurasian Journal of Educational Research*, *18*, 236-252.

Baykul, Y. (2015). *Eğitimde ve psikolojide ölçme: Klasik test teorisi ve uygulaması.* Pegem.

Bıkmaz Bilgen, Ö., & Doğan, N. (2017). Çok kategorili parametrik ve parametrik olmayan madde tepki kuramı modellerinin karşılaştırılması [Comparison of polytomous parametric and nonparametric item response theory models]. *Journal of Measurement and Evaluation in Education and Psychology, 8*(4), 354-372.

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

Büyüköztürk, Ş., Çakmak, E.K., Akgün, Ö.E., Karadeniz, Ş. ve Demirel, F. (2012). *Bilimsel Araştırma Yöntemleri.* Pegem.

Choi, J., & Wilson, M. R. (2018). Modeling rater effects using a combination of generalizability theory and IRT. *Psychological Test and Assessment Modeling*, *60*(1), 53-80.

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

Çakıcı Eser, D., & Gelbal, S. (2013). Genellenebilirlik kuramı ve lojistik regresyona dayalı hesaplanan puanlayıcılar arası tutarlılığın karşılaştırılması [Comparison of interrater agreement calculated with generalizability theory and logistic regression]. *Kastamonu Education Journal, 21*(2), 421-438.

Çakır, M., & Aldemir, B. (2011). İki aşamalı genetik kavramlar tanı testi geliştirme ve geçerlik çalışması [Developing and validating a two tier mendel genetics diagnostic test]. *Mustafa Kemal University Journal of Social Sciences Institute*, *8*(16), 335-353.

Deliceoğlu, G., & Çıkrıkçı Demirtaşlı, N. (2012). Futbol yetilerine ilişkin dereceleme ölçeğinin güvenirliğinin genellenebilirlik kuramına ve klasik test kuramına dayalı olarak karşılaştırılması [The comparison of the reliability of the soccer abilities' rating scale based on the classical test theory and generalizabilty theory]. *Hacettepe Journal of Sport Sciences, 23*(1), 1-12.

Demir, B. P. (2016). Vee diyagramından elde edilen puanların güvenirliğinin klasik test kuramı ve genellenebilirlik kuramına göre incelenmesi [The examination of reliability of vee diagrams according to classical test theory and generalizability theory]. *Journal of Measurement and Evaluation in Education and Psychology, 7*(2), 419-431.

Demirel, Ö., & Epçaçan, C. (2012). Okuduğunu anlama stratejilerinin bilişsel ve duyuşsal öğrenme ürünlerine etkisi [Effects of reading comprehension strategies on cognitive and affective learning outcomes]. *Kalem International Journal of Education and Human Sciences, 2*(1), 71-106.

Doğan, C. D., & Anadol, H. Ö. (2017). Genellenebilirlik kuramında tümüyle çaprazlanmış ve maddelerin puanlayıcılara yuvalandığı desenlerin karşılaştırılması [Comparing fully crossed and nested designs where items nested in raters in generalizability theory]. *Kastamonu Education Journal, 25*(1), 361-372.

Doğan, N., & Bıkmaz Bilgen, Ö. (2017). Using generalizability theory in reliability estimation of measurements of higher-order cognitive skills. *The Journal of Academic Social Science, 44*, 1-9.

Fraenkel, J. R., Wallen, N. E., & Hyun, H. H. (2015). *How to design and evaluate research in education.* McGraw Hill Education.

Giray, M. D., & Sahin, D. N. (2012). Algılanan örgütsel, yönetici ve çalışma arkadaşları desteği ölçekleri: Geçerlik ve güvenirlik çalışması [Perceived organizational, supervisor and co-worker support scales: A study for validity and reliability]. *Turkish Psychological Articles, 15*(30), 1-9.

Güler, M., & Yetim, Ü. (2008). Ebeveyn rolüne ilişkin kendilik algısı ölçeği: Geçerlik ve güvenirlik çalışması [Self-perception of parental role (SPPR) scale: Validity and reliability study]. *Turkish Psychological Articles*, *11*(22), 34-43.

Güler, N. (2011). Rasgele veriler üzerinde genellenebilirlik kuramı ve klasik test kuramına göre güvenirliğin karşılaştırılması [The comparison of reliability according to generalizability theory and classical test theory on random data]. *Education and Science, 36*(162), 225-234.

Güler, N., Eroğlu, Y., & Akbaba, S. (2014). Genellenebilirlik kuramına göre ölçüt bağımlı ölçme araçlarında güvenirlik: Yemek yeme becerileri örneğinde bir uygulama [Reliability of criterion-dependent measurement tools according to generalizability theory: Application in the case of eating skills]. *Abant İzzet Baysal University Journal of Faculty of Education, 14*(2), 217-232.

Güler, N., Kaya Uyanık, G., & Taşdelen Teker, G. (2012). *Genellenebilirlik kuramı [Generalizability theory].* Pegem.

Gülle, A., Uzun, N. B., & Akay, C. (2018). Ortaokul öğrencilerine yönelik blok flüt icra performansı dereceli puanlama anahtarının güvenirliğinin genellenebilirlik kuramı ile incelenmesi [The study on the reliability of the grading key measuring the performance of the block flute performance of the secondary school students via generalizability theory]. *Elementary Education Online, 17*(3), 1463-1475.

Hathcoat, J. D., & Penn, J. D. (2012). Generalizability of student writing across multiple tasks: A challenge for authentic assessment. *Research & Practice in Assessment, 7*, 16-28.

Hill, H. C., Charalambous, C. Y., & Kraft, M. A. (2012). When rater reliability is not enough: Teacher observation systems and a case for the generalizability study. *Educational Researcher*, *41*(2), 56-64.

Hulin, C. L., Lissak, R. I., & Drasgow, F. (1982). Recovery of two and three parameter logistic item characteristic curves: A monte carlo study. *Applied Psychological Measurement*, *6,* 249-260.

Kamış, Ö., & Doğan, C. D. (2017). Genellenebilirlik kuramında gerçekleştirilen karar çalışmaları ne kadar kararlı? [How consistent are decision studies in g theory?]. *Gazi University Journal of Gazi Educational Faculty, 37*(2), 591-610.

Kaplan, A., & Duran, M. (2016). Ortaokul öğrencilerine yönelik matematiksel üstbiliş farkındalık ölçeği: Geçerlik ve güvenirlik çalışması [Mathematical metacognition awareness inventory towards middle school students: Validity and reliability study]. *Journal of Kazım Karabekir Education Faculty, 32*, 1-17.

Karasar, N. (2016). *Bilimsel Araştırma Yöntemi.* Nobel.

Karlsson, J. (2017). *Generalizability theory and a scale measuring emotion knowledge in preschool children* [Master's thesis, Stockholm University]. http://www.diva-portal.org/smash/get/diva2:1065849/FULLTEXT01.pdf

Katrancı, M., & Yangın, B. (2012). Üstbiliş stratejileri öğretiminin dinlediğini anlama becerisine ve dinlemeye yönelik tutuma etkisi [Effects of teaching metacognition strategies to listening comprehension skills and attitude toward listening]. *Adiyaman University Journal of Social Sciences*, *2013*(11), 733-771.

Kaya, A. (2005). Çocuklar için yalnızlık ölçeğinin Türkçe formunun geçerlik ve güvenirlik çalışması [The validity and reliability study of the Turkish version of the children`s loneliness scale]. *Eurasian Journal of Educational Research*, *19,* 220-237.

Kenny, D.A. (1987). *Statistics for the social and behavioral science*. Little, Brown.

Nitko, A. (2004). *Educational asssessments of students.* Pearson.

Scherbaum, C., Dickson, M., Larson, E., Bellenger, B., Yusko, K., & Goldstein, H. (2018). Creating test score bands for assessments involving ratings using a generalizability theory approach to reliability estimation. *Personnel Assessment and Decisions*, *4*(1), 1-8. https://doi.org/10.25035/pad.2018.001

Solano-Flores, G., & Li, M. (2013). Generalizability theory and the fair and valid assessment of linguistic minorities. *Educational Research and Evaluation*, *19*, 245-263. https://doi.org/10.1080/13803611.2013.767632

Tavşancıl, E. (2005). *Tutumların ölçülmesi ve SPPS ile veri analizi [Measurement of attitudes and data analysis with SPPS]*. Nobel.

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

Yılmaz Nalbantoğlu, F., & Gelbal, S. (2011). İletişim becerileri istasyonu örneğinde genellenebilirlik kuramıyla farklı desenlerin karşılaştırılması [Comparison of different designs in accordance with the generalizability theory in communication skills example]. *Hacettepe University Journal of Education, 41*, 509-518.