Journal of Education and Learning; Vol. 7, No. 4; 2018
ISSN 1927-5250 E-ISSN 1927-5269
Published by Canadian Center of Science and Education

An Investigation of Reliability Coefficients Estimated for Decision Studies in Generalizability Theory

Ömer Kamış[1] & C. Deha Doğan[1]

[1] Department of Measurement and Evaluation, Ankara University, Turkey

Correspondence: C. Deha Doğan, Department of Measurement and Evaluation, Ankara University, Cemal Gürsel Cd., Cebeci 06590, Ankara, Turkey.

Received: February 15, 2018 Accepted: April 5, 2018 Online Published: May 15, 2018

doi:10.5539/jel.v7n4p103 URL: https://doi.org/10.5539/jel.v7n4p103

Abstract

This research aimed to compare the G and Phi coefficients estimated in Decision studies in Generalizability theory and obtained in actual cases for the same conditions of similar facets by using crossed design. The research was conducted as pure research on 120 individuals (students), six items and 12 raters. An achievement test composed of six open ended questions and a holistic rubric developed by the researcher were used in data collection. Data analysis included obtaining the G and Phi coefficients by creating actual cases for two, four and six raters followed by D studies conducted for other actual cases with different measurement conditions to make estimates. Finally, G and Phi coefficients obtained and estimated for two, four and six raters were compared separately. Findings show that G and Phi coefficients estimated in D studies by increasing the number of raters were sometimes greater than those obtained in actual cases although they were sometimes smaller. However, it was concluded that a pattern may not always exist between the G and Phi coefficients obtained in actual cases for same number of raters and those estimated in D studies.

Keywords: generalizability theory, decision studies, reliability, pure research

1. Introduction

Traits such as personality, interest, attitude, ability and achievement are considered as the subject of measurement in education and psychology. Psychological tests are used to measure these traits. Measurements performed by using psychological tests can be called psychological measurements (Cohen and Swerdlik, 2009). According to Erkuş (2014), psychological measures are mostly indirect measures. In other words, individual traits which are subject to measurement are measured by means of items presumed to measure the relevant trait and called stimulus.

Individuals receive scores for the responses they provide to these items and thus the observed scores are obtained which are used to determine the extent to which individuals have the psychological trait in question (Özgüven, 2012). Repeated applications of a measurement instrument may generate differences in the scores observed for the same individual due to various factors. These differences are regarded as errors in the context of measurement. Sources of these errors can generally be classified in three groups: the context in which the measurement is made (testing environment and the reason of the test including factors related to the implementer and the rater), the test taker and the test itself (Urbina, 2004).

Errors resulting from the three error sources cited above are of three types: fixed, systematic and random. These errors, which may interfere with the measurement process, reduce the validity and reliability of the scores obtained from the measurement process. While validity refers to the degree in which scores obtained from the measurement process serve its purpose; reliability, in short, is defined as the degree of freedom from random errors in the scores. Errors in the measurement process should be minimized to have high level of reliability.

Measurement theories such as Classical Test Theory (CTT), Generalizability Theory (GT) and Item Response Theory (IRT) can be used to determine errors in the measurement process and the reliability of measurement applications. A different source of error is examined in each of the reliability estimation methods in CTT. For example, time is considered as the error source in the test re-test method whereas the scope of the test is considered as the error source in the equivalent forms method. According to the CTT, it is necessary to analyze each error source separately when determining the reliability of the scores obtained from a measurement.

Foundations of GT were established in the early 1940s to overcome the limitations of CTT. GT, which was developed with the contributions of Cronbach, Rajaratnam, Gleser, Shavelson and Webb (Crocker and Algina, 2008), addresses all error sources in a measurement process in combination and sets the reliability coefficient by providing a variance value for all error sources with a single analysis.

Generalizability (G) analyzes are used not only to understand the relative importance of various error sources, but also to design reliable measurement processes. Another advantage of the GT is that a large number of error sources can be separately identified by a single analysis.

In GT, each factor that generates variance in the scores in measurement is called a source of variability. The universe defined by a source of variability is called a facet. Each level of a facet is called a condition (Shavelson and Webb, 1991).

In GT, the variability source for which the desired variance is obtained from among different variability sources is called the object of measurement (Güler, Kaya Uyanık and Taşdelen Teker, 2012; Suen, 1990). Relative and absolute decisions can be made about the object of measurement in GT by first calculating absolute and relative error variances and then obtaining the G and Phi coefficients. While the G coefficient is the reliability coefficient used to make the relative decisions, the Phi coefficient is the reliability coefficient used to make absolute decisions (Brennan, 2001).

Two studies (G and Decision (D) studies) are performed to calculate the reliability of a measurement in GT. While the purpose of G study is to provide as much information as possible about variability sources in a measurement, the purpose of D study is to determine the number of variability sources to produce the reliable measurement cases (Shavelson and Webb, 1991). G and D studies in GT can be used to calculate the reliability of many measurement applications at international and national level where open-ended items and raters or judges are included as variability sources. However, in schools, teachers prefer measurement tools that consist of open-ended item types to determine students' learning deficiencies and their level of recent learning. Likewise, measurement tools with open-ended item types are used to evaluate students' written expression skills in English preparatory classes at universities.

In institutional and classroom measurement applications, G and D studies in GT can be utilized in determining the reliability of the scores with these and similar measurement applications that are planned. Number of items and raters required to establish more reliable future applications can be determined by using D studies. Reliability (G and Phi) coefficients can be estimated for different measurement situations formed by changing the number of items and raters in D studies. Since it is difficult in practice to change the number of items due to content validity in D studies, the reliability is generally estimated by changing the number of raters. In this process, a new variance value for the new condition numbers in the facets is not calculated but the variance values obtained from the G study are used. In the estimation of relative and absolute error variances, only the new values of the condition are written instead of the condition numbers. The error variances estimated in this way are used to calculate the reliability coefficients for subsequent applications of the measurement process. However, it is an important question whether the reliability coefficients estimated in D studies are consistent with the values in the real cases.

It is important to know the degree to which the estimate made in D studies reflects actual cases. This information may contribute to a more accurate interpretation of results of the D study and to more qualified decisions. In addition, if one of the variability sources included in a measurement application is a rater, D study results can be utilized when the number of raters required for reliable scoring is determined. In order to make the right decision for the number of raters, it is important to know the degree to which the results obtained in D studies reflects actual cases.

Problems may occur when the reliability coefficients estimated in D studies differ from the values in actual cases. If the predicted reliability coefficients are lower than their actual values, the number of conditions will be increased to reduce the error related to variability sources. This may lead to a waste of resources. However, if the predicted reliability coefficients are higher than their actual values, the targetted reliability level cannot be achieved. In order to avoid these and similar problems, it is important to know how well the reliability coefficients predicted in D studies reflect actual cases. However, studies on this topic are limited.

When the literature is examined, several D studies in GT can be found (Anıl and Büyükkıdık, 2012; Büyükkıdık and Anıl, 2015; Can Aran, Güler, and Senemoğlu, 2014; Deliceoğlu and Çıkrıkçı Demirtaşlı, 2012; Han, 2016; Hoyt and Melby, 1999; Lin and Zhang, 2014; Nalbantoğlu Yılmaz and Gelbal, 2011; Wu et al., 2016; Yelboğa, 2008, 2012; Yelboğa and Tavşancıl, 2010). In these studies, first a measurement context was created followed by a G-study and then the reliability values estimated in the D study conducted by changing the rater or item

numbers were interpreted.

A limited number of studies (Arsan, 2012, Atılgan and Tezbaşaran, 2005, Gao and Brennan, 2001, Kamış and Doğan, 2017) were found in the literature comparing the predicted reliability values with the actual reliability values for the number of raters in D studies. In these researches in general, it was found that the reliability coefficients predicted in D studies differ from the reliability coefficients obtained in actual cases. In cases (Arsan, 2012) where the rater characteristics were similar, it was concluded that the estimated values in D studies were consistent with the values obtained in actual cases. However, some of these studies provide a limited amount of information on how the conditions on the surface were selected from relevant universes. In addition, there were no studies in which the raters were randomly selected from the universe. In the present study, a universe was defined for the raters and the raters were randomly selected from the universe. This way, the basic assumption of the G theory was better reflected and thus how well the reliability coefficients predicted in D studies reflected actual cases was more clearly presented.

Other research on D studies concluded that different analyses programs produced similar results (Güler, 2009), a sample of at least 50 individuals should be selected in order to adequately represent the individual's universe (Atılgan, 2013) and bootstrap methods produced more reliable results when data did not have normal distribution (Özberk and Gelbal, 2014).

Knowing how well the reliability coefficients predicted in the D study in GT reflect actual cases will contribute to a more accurate interpretation of the results of the D study and will prevent possible problems. For this reason, the problem statement of this research is related to the examination of how well the reliability coefficients estimated in D studies by using the variance values obtained from the G studies in GT reflect the reliability coefficients obtained in actual cases.

1.1 Purpose

This research aimed to compare the G and Phi coefficients estimated in Decision studies in Generalizability theory and obtained in actual cases for the same conditions of similar facets by using crossed design. In line with this general purpose, answers were sought for the following questions.

1. When the object of measurement is selected as a person, an item and person-item respectively,

a) When the number of raters are 2 in actual cases, what are the G and Phi coefficients?

b) As a result of D study performed when the number of raters is actually 4, what are the G and Phi coefficients estimated for 2 raters?

c) As a result of D study performed when the number of raters is actually 6, what are the G and Phi coefficients estimated for 2 raters?

2. When the object of measurement is selected as a person, an item and person-item respectively,

a) When the number of raters are 4 in actual cases, what are the G and Phi coefficients?

b) As a result of D study performed when the number of raters is actually 2, what are the G and Phi coefficients estimated for 4 raters?

c) As a result of D study performed when the number of raters is actually 6, what are the G and Phi coefficients estimated for 4 raters?

3. When the object of measurement is selected as a person, an item and person-item respectively,

a) When the number of raters are 6 in actual cases, what are the G and Phi coefficients?

b) As a result of D study performed when the number of raters is actually 2, what are the G and Phi coefficients estimated for 6 raters?

c) As a result of D study performed when the number of raters is actually 4, what are the G and Phi coefficients estimated for 6 raters?

1.2 Assumptions

It was assumed that students in the study group reflected their actual performance in response to the achievement test developed in the framework of the research. It was also assumed that the raters in the study group acted rigorously and diligently in scoring students.

1.3 Limitations

An fully crossed design, one of the GT designs, was used in this study. Other designs in the theory were not

included in the study due to time and cost. This research was conducted "person", "item" and "rater" variability sources which are frequently used in GT. This research was limited to two, four and six raters, considered to be practically usable in terms of the number of raters. The reliability coefficients for the same number of raters that were obtained in actual cases and predicted in different D studies were descriptively compared in this study. Descriptive comparisons were made due to lack of scientifically agreed statistical methods to test the significance of differences between these coefficients.

1.4 Definition

Student Input: Student input refers to students' responses given to the items in the achievement test developed within the scope of this research. In this context, student input is obtaiend from their exam papers.

2. Method

2.1 Research Model

This research was conducted as pure research. These type of studies are designed to advance knowledge on ambiguous issues about current or future applications in a subject area. Researchers conducting pure research examine the concepts and assumptions of the relevant subject area (Bailey, 1994). The main purpose of pure research is to add new information to existing information (Karasar, 2006). This research investigated whether or to what rate using the variance values for variance sources obtained in G studies in GT for D studies reflects actual cases.

2.2 Study Group

This research was conducted on two different study groups. The first study group was composed of 149 undergraduate and pedagogical formation program students attending the Faculty of Educational Sciences at Ankara University in 2016/2017 academic year and taking "Measurement and Evaluation" course. The second study group consisted of raters. The raters were selected by simple random sampling among the instructors with PhD. or higher degrees employed in Departments of Measurement and Evaluation in Turkish Universities. In this context, a total of 12 raters (first two, then four and finally six raters) were selected from among 102 raters. The R computer program was used in the selection process.

2.3 Data Collection

An achievement test composed of six open ended questions that focus on assessing the knowledge and skills taught in the Measurement and Assessment Course and a holistic rubric for scoring the items in the achievement test were used in data collection. The holistic rubric was preferred since the performance required by students to respond to each item used in the achievement test could not be divided into sub areas. The achievement test developed by the reseracher was given to 149 individuals who constituted the student study group. A total of 29 student tests with missing or insufficient data were excluded and were not sent to raters because they would not reveal the differences between individuals. Therefore, 120 student tests were used in the research.

Student tests were numbered from 1 to 120 and the resulting name-number matches were recorded. Then, among the 120 student tests, three separate 60 student tests were randomly selected for two, four, and six rater groups. The findings of Atılgan's (2013) study were taken into consideration to identify the number of students tests that were sent to raters. In this study, it was emphasized that a sample of at least 50 individuals should be used in order to provide an unbiased estimate for a universe of 480691. R program commands were utilized in selecting student tests.

The first group of 60 student tests selected as a result of the first sample selection command was sent to the group consisting of two raters. Then the next group of 60 student tests determined as a result of the second sample selection command was sent to the group consisting of four raters. Finally, the last group of 60 student tests determined as a result of the third sample selection command was sent to the group consisting of six raters. Student tests were sent to raters via cargo. The raters scored student tests using the holistic rubric developed within the framework of this research.

2.3.1 Achievement Test

The achievement test was developed by following the steps in the test development process. The taxonomy established by Haladyna (1997) was used to define the cognitive level dimension for the learning outcomes. The identified learning outcomes and the relevant cognitive levels are presented in the Table of Specifications.

After the Table of Specifications were created, items regarding the identified learning outcomes were written. Since the final form of the achievement test was planned to include six open ended items, a total of 13 items were written in the first stage, with at least two items for each learning outcome. The items in the achievement

test were designed to target high-level cognitive processes related to daily life. The items were submitted for expert review to be examined for content validity, language and expression and scientific accuracy. The trial form was prepared by revising the items in line with the suggestions from the experts.

Content validity index proposed by Hambleton and Rovinelli (1977) (cited in Crocker and Algina, 2008) was calculated to to determine the fit of the items on the trial form with the learning outcomes the items aimed to measure. Views of a total of five experts (three doctoral students and two experts with PhD degrees in the field of Measurement and Evaluation) were sought to calculate this index. Calculations pointed to + 1 fit indice, the highest level, between each item and the learning outcome that the item aimed to measure. This value indicates that each item accurately measures the learning outcome that it aims to measure. After the calculations were performed, trial application started.

Trial application was conducted on a group of 76 individuals with similar composition to the group of individuals that would take part in the main application. Trial application checked whether the items were clear and intelligible for students and whether there were items that were not answered. It was determined that the items items were clear and intelligible for students in general. Only three items were revised for clarity in expression. Finally, 6 items were selected from among the trial items and the achievement test was finalized.

2.3.2 Holistic Rubric

A separate holistic rubric was developed for each item to score the items in the achievement test. The steps identified by Kutlu, Doğan and Karakaya (2014) were followed while developing the rubric. First of all, the correct answer for each item and score for this answer were determined. Then, distant correct answers that are closest to correct answers and their scores were identified for each item by taking into account the cognitive processes that students should use in responding to these items. Students who answered the item using higher level cognitive processes were given higher scores. The draft form of the rubric was submitted to two experts with doctorates in the Departmant of Measurement and Evaluation for review and necessary revisions were made in line with the opinions received.

The trial application for the revised achievement test was scored using the rubric. The reliability analysis provided a Cronbach alpha value of 0.59. Since this reliability value was low (Diederich, 1973) for teacher-made tests, the rubric was improved to allow distinguishing student responses more clearly. The Cronbach alpha value of 0.71 was obtained after the second trial application with the revised rubric carried out on a group of 44 people. The obtained value of reliability was regarded to be sufficient considering the number of items (Diederich, 1973) and the final form of the rubric and the achievement test were used in the main application.

2.4 Data Analysis

2.4.1 Preparing Data for Analysis

Data generated by raters after scoring student tests were entered in the Excel program. First, data from each rater were saved in separate Excel files. Subsequently, data of the raters in the same group were combined. An example of the data structure was provided in Table 1 for the group composed of two raters.

Table 1. Example of the data structure for two raters in a fully crossed design

Students	I1		I2		I3		I4		I5		I6	
	R1	R2	R1	R2	R1	R2	R1	R2	R1	R2	R1	R2
1	x	x	x	x	.	.	.	.	.	.	x	x
2	x	x	x	x	.	.	.	.	.	.	x	x
.	.	.	.	.							.	.
.	.	.	.	.							.	.
.	.	.	.	.							.	.
60	x	x	x	x							x	x

Table 1 shows that the first column of the data matrix is for students, the first row is for items and the second row is for raters. This sequence also includes information as to in which order the facets in this research will be entered to the Edu G program used in the analysis of the data. In this context, the order of the facets to be entered in the program was as follows: students (person)-item-rater. Following this prepation, data analysis was carried out.

2.4.2 Operations Performed in Data Analysis

G and Phi coefficients were obtained by generating the actual measurement cases in the manner cited above for the two, four and six raters according to the fully crossed design in order to answer the research questions included in the sub-goals of this research. Then, D study was conducted using the actual measurement cases with four and six raters according to the crossed design and the G and Phi coefficients for two raters were estimated. Likewise, the D study was conducted using the actual measurement cases with two and six raters for four raters and also two and four raters for six raters, and G and Phi coefficients for four and six raters were estimated. Finally, G and Phi coefficients that were obtained and estimated for the two, four and six raters were compared separately.

While the G and Phi coefficients were calculated for two, four and six raters in actual cases, first the variance values related to the sources of variability in each data set were calculated. Then, using these calculated variance values, relative and absolute error variances were obtained for each case where the object of measurement is person, item and person-item respectively. Finally, G and Phi coefficients were calculated by using variance values related to relative and absolute error variances and objects of measurement. In the actual cases, after the G and Phi coefficients were obtained, D studies were performed for each situation in which the number of raters were two, four, and six. Therefore, G and Phi coefficients were estimated for each situation.

The variance values that were obtained in the case including four and six raters were used in estimating the G and Phi coefficients for two raters. Only the relative and absolute error variances were estimated by taking the number of raters as 2, and then the G and Phi coefficients were estimated by using these estimated error variances. Similarly, the variance values obtained from two and six raters were used while the D study was performed for four raters and the number of raters were taken as 4 in the estimation of error variances. Similar procedures were repeated to estimate G and Phi coefficients for six raters from two and four raters. Edu G 6.1-e program was used in all these calculations and estimates.

3. Results

3.1 Findings: When the Object of Measurement Was a Person

In order to answer questions regarding the object of measurement as a person, the data collected from two, four and six raters were examined. By performing the G study; G and Phi coefficients were obtained for cases where the number of raters was actually two, four and six. Then, D studies were performed using the collected data for each of the other rater numbers for each rater number and the G and Phi coefficients were estimated for cases where rater number was actually two, four and six. Table 2 provides the reliability coefficients obtained in actual cases and estimated in D studies when the number of raters were two, four and six.

Table 2. When the object of measurement was selected as a person, The G and Phi coefficients obtained and estimated for the two, four and six raters

Design	**Object of Measurement**	**Rater Number**	**Actual Cases**		**D Studies** (estimated for the number of raters in the actual case)	
			G	**Phi**	**G**	**Phi**
bxmxp	P	2	0.57	0.46	-	-
		4	-	-	0.58	0.44
		6	-	-	0.46	0.43
bxmxp	P	2	-	-	0.58	0.48
		4	0.61	0.48	-	-
		6	-	-	0.49	0.47
bxmxp	P	2	-	-	0.59	0.49
		4	-	-	0.63	0.50
		6	0.50	0.49	-	-

When the findings obtained when the object of measurement was selected as a person were taken into consideration, it can be stated that there were differences between the reliability coefficients obtained in actual cases and estimated in D studies when the number of raters were two, four and six. The differences between the G coefficients varied between 0.01 and 0.13, while the differences between the Phi coefficients varied between 0.00 and 0.03. In addition, no patterns were found between the estimated/actually obtained reliability coefficients and increasing or decreasing the number of raters.

3.2 Findings: When the Object of Measurement Was an Item

The variance value for the source of item variability was calculated as 0 (zero) in the case where the number of raters was six. This led to the calculation and estimation of the G and Phi coefficients as 0, both in the actual case where the number of raters was six and in estimations for the other raters from six raters. For this reason, no comparison was made for the six raters when the object of measurement was an item and only the reliability coefficients, obtained in actual case and estimated in D studies for two and four raters, were compared.

In order to answer the reserach questions in which the object of measurement was an item, the object of measurement was selected as the item by considering the data collected from two and four raters. Then by performing the G study, G and Phi coefficients were obtained for the situations where the number of raters was actually two and four. Later, D studies were performed using the collected data for each of the other rater number for each rater number and the G and Phi coefficients were estimated for situations where the rater number was actually two and four. Table 3 provides the reliability coefficients obtained in the actual cases and estimated in the D studies when the number of raters were two and four.

Table 3. The G and Phi coefficients obtained and estimated for the two and four raters when the object of measurement was selected as an item

Design	**Object of Measurement**	**Rater Number**	**Actual Cases**		**D Studies** (estimated for the number of raters in the actual case)	
			G	**Phi**	**G**	**Phi**
bxmxp	I	2	0.90	0.87	-	-
		4	-	-	0.86	0.78
bxmxp	I	2	-	-	0.93	0.91
		4	0.91	0.87	-	-

It can be stated that the reliability coefficients estimated in D studies were different from reliability coefficients actually obtained for the same rater numbers in the case where the item was selected as the object of measurement and the number of raters was two and four. The differences between the G coefficients varied between 0.02 and 0.04, while the differences between the Phi coefficients varied between 0.04 and 0.09. However, it was found that the reliability coefficients estimated for four raters were larger than their actual values in D study performed using two raters and the reliability coefficients estimated for the two raters were smaller than their actual values in the D study performed using the four raters.

3.3 Findings: When the Object of Measurement Was Person-Item

In order to answer questions regarding the object of measurement as person-item, the data collected from two, four and six raters were examined. Then, by performing the G study, G and Phi coefficients were obtained for cases where the number of raters was actually two, four and six. D studies were performed for each rater number by using the collected data for each of the other rater number and G and Phi coefficients were estimated for cases where the rater number was actually two, four and six. Table 4 provides the reliability coefficients obtained in the actual cases and estimated in D studies when the number of raters were two, four and six.

Table 4. When the object of measurement was selected as person-item, The G and Phi coefficients obtained and estimated for the two, four and six raters

Design	**Object of Measurement**	**Rater Number**	**Actual Cases**		**D Studies** (estimated for the number of raters in the actual case)	
			G	**Phi**	**G**	**Phi**
bxmxp	PI	2	0.90	0.89	-	-
		4	-	-	0.84	0.82
		6	-	-	0.83	0.81
bxmxp	PI	2	-	-	0.95	0.94
		4	0.91	0.90	-	-
		6	-	-	0.90	0.90
bxmxp	PI	2	-	-	0.96	0.96
		4	-	-	0.94	0.93
		6	0.93	0.93	-	-

It can be stated that reliability coefficients actually obtained and the reliability coefficients estimated in D studies for the same rater numbers were the same in terms of the reliability level but different in terms of numerical value in the case where the object of measurement was selected as person-item and the number of raters were two, four and six. The differences between the G coefficients varied between 0.01 and 0.07, while the differences between the Phi coefficients varied between 0.00 and 0.08. However, it was observed that the reliability coefficients estimated by decreasing the number of raters were lower than their actual values, and the reliability coefficients estimated by incraesing the number of raters were higher than their actual values.

4. Discussion

When the findings obtained for three research questions were assessed in general, it was seen that the G and Phi coefficients obtained in actual cases were different from the ones estimated in the D studies when the object of measurement was selected as the person, the item and the person-item respectively for the two, four and six raters. This finding is consistent with the findings of Atılgan and Tezbaşaran (2005), Gao and Brennan (2001) and Kamış and Doğan (2017). In the case where the object of measurement was a person, the G coefficient obtained in the actual case for four raters was 0.61 while the G coefficient estimated for four raters in D studies using two and six raters was smaller than 0.60. Likewise, in the case where object of measurement was an item, the Phi coefficient obtained in actual cases for two raters was 0.87 while the Phi coefficient estimated for two raters in D studies using four raters was smaller than 0.80. These findings may show that the reliability coefficients obtained for different rater numbers in actual cases were not similar to those estimated in the D studies in terms of the reliability level. In other words, the levels of reliability coefficients estimated in D studies were lower in some cases than their actual levels. Differences arising in cases at levels considered to be boundary values for reliability (0.60 and 0.80) can become significant. However, the G and Phi coefficients estimated in D studies may not always reflect actual cases. These differences can become problematic when decisions are critical. Accurate estimation of G and Phi coefficients is crucial for correct interpretation of D studies. In addition, while D studies where reliability coefficients were estimated to be lower than actual values may lead to resource waste, D studies, which are estimated to be higher than actual values may result in insufficient reliability levels.

It is predicted that the error due to the facet will be reduced by increasing the number of conditions in that facet in GT, and the theoretically higher reliability coefficients can be reached in this manner (Shavelson and Webb, 1991). However, according to the findings of this research, smaller G coefficients were obtained when the number of raters was increased by selecting randomly from the relevant universe in the case where the object of measurement was determined to be a person. While the G coefficient was obtained as 0.61 in the case where the rater number is actually four, the G coefficient was calculated to be 0.50 in the case where the rater number is actually six. This might be due to the fact that the raters were selected randomly from the universe. The fact that each new rater group with characteristics different from the previous rater group was formed at each selection may have caused differences in the variance values related to different sources of variability. In addition, raters' past training experiences may have been effective in their scoring.

According to the findings of this research, a pattern between the coefficients estimated in D studies and the coefficients obtained in actual cases was found in the case where the raters were selected randomly from the relevant universe, and item and person-item were selected as the object of measurement. While G and Phi coefficients were estimated to be larger than their actual values in estimations for a bigger number of raters than the number of actual raters; G and Phi coefficients were estimated to be smaller than their actual values in estimations made for fewer number of raters than the actual number of raters. However, a similar pattern was not obtained when a person was selected as object of measurement. Accordingly, it can be stated that the values of the G and Phi coefficients in actual cases could not be predicted systematically from D studies. However, Atılgan and Tezbaşaran (2005) concluded that the G and Phi coefficients, which are estimated by increasing the number of raters in the study they performed, were larger than their actual values and the reliability coefficients estimated by decreasing the number of raters were smaller than their values in actual cases. Arsan (2012) concluded that the coefficients estimated by increasing or decreasing the number of referees in D studies were similar to coefficients obtained from original referee numbers. While findings of the current research did not coincide with the findings of Arsan (2012), the pattern findings of the current research coincided with the study of Atılgan and Tezbaşaran (2005). In addition, the findings of the research are consistent with the findings of Gao and Brennan (2001) and Kamış and Doğan (2017).

Consistency in findings may also be related to the characteristics of selected raters. As a matter of fact, in Arsan's (2012) work, the referees were selected according to their expertise. In other words, judges constituting the judge's universe were qualified persons in terms of refereeing criteria. According to the findings of Arsan (2012),

the percentage of the variance values related to the referees in the evaluations made in different years varied between 0.00 and 1.80. There is no clear information about the rater selection in the research made by Atılgan and Tezbaşaran (2005) and Gao and Brennan (2001). However, raters were not selected randomly from the universe in the research performed by Kamış and Doğan (2017) due to practical conditions. In the current study, a universe was defined for raters, and the raters were selected randomly from this universe. The process and characteristics of determining the raters may have influenced the consistency of research findings.

The difference between coefficients estimated in different D studies and obtained in actual cases for the same number of raters may also be generated from the variance estimation method used in studies. In the current study, variance analysis (ANOVA) was used as the variance estimation method. In the study conducted by Özberk and Gelbal (2014), ANOVA and bootstrap methods were compared and it was found that the bootstrap methods were more suitable than the ANOVA method for estimating the variance components of data with non-normal distribution.

The difference between coefficients estimated in D studies and obtained in actual cases might also be due to sample size. Although there are research in the literature related to number of individuals necessary to correctly predict actual cases in D studies, there are no research on the numbers of raters. It was determined in a research performed on the number of persons (Atılgan, 2013) that a sample formed of at least 50 people from the 480691-person universe should be selected in order to accurately predict actual cases. In this research, inconsistency between reliability coefficients obtained in actual cases and predicted in D studies might also be due to the fact that the number of raters was limited to a certain number. In cases consisting of a large number of raters (10, 15 and 20 instead of 2, 4 and 6 raters), there may be an overlapping between the reliability coefficients estimated in the D studies and obtained in actual cases.

In this current research, reliability coefficients obtained in actual cases and estimated in D studies were descriptively compared and differences between the coefficients were not statistically tested. When the literature was examined, no statistically agreed method was found for comparing the G and Phi coefficients. In addition, G and Phi coefficients were descriptively compared in studies conducted by well known researchers (Atılgan and Tezbaşaran, 2005; Gao and Brennan, 2001) working in the field.

Items and raters can be used as a source of variability in many measurement applications such as selection and placement, feedback and grading. G and D studies in GT can be used in such measurement applications when determining the reliability for different objects of measurement. When a measurement application is to be redesigned or to be implemented in the future, the number of raters is increased or decreased according to the results of the D study. According to the findings of this research, it was concluded that in such cases, the actual cases can not be estimated precisely by results of D studies. The reliability coefficients estimated in the D studies were lower than their actual values in some cases, but in some cases they were higher than their true values. While underestimated reliability coefficients may lead to resource waste, overestimated reliability coefficients may result in insufficient reliability levels.

However, in the case where the measurement object is selected as an person, the reliability coefficients estimated in D studies performed by decreasing or increasing the number of the raters were found to be larger or smaller than their actual values in some cases. While in the case where the measurement object was person-item and an item, the reliability coefficients estimated in the D studies by increasing the number of raters were found to be larger than their actual values; the reliability coefficients estimated in the D studies by decreasing the number of raters were found to be smaller than their actual values. According to this, it was cocluded that a pattern may not always exist between the reliability coefficients estimated in D studies and the reliability coefficients obtained in actual cases.

The practitioners who will conduct further studies taking into account the results of the D study in a given research area are recommended to be aware of the fact that D studies do not always accurately represent actual cases and that they should pay attention to this fact in their practice. It is suggested that researchers emphasize the limitations of D study results while reporting their research and D studies should be interpreted within these limitations.

References

Anıl, D., & Büyükkıdık, S. (2012). Genellenebilirlik kuramında dört facetli karışık desen kullanımı için örnek bir uygulama [A sample application for using four facet mixed designs in generalizability theory]. *Journal of Measurement and Evaluation in Education and Psychology*, *3*(2), 291-296.

Arsan, N. (2012). *Buz pateninde hakem değerlendirmelerinin genellenebilirlik kuramı ve rasch modeli ile*

incelenmesi [Investigation of the raters' assessment in ice skating with generalizability theory and rasch measurement] (Ph.D. Dissertation). Retrieved from https://tez.yok.gov.tr/UlusalTezMerkezi/giris.jsp

Atılgan, H. (2013). Genellenebilirlik kuramında G ve Phi katsayılarının kestirilmesi için örneklem büyüklüğü [Sample size for estimation of G and Phi coefficients in generalizability theory]. *Eurasian Journal of Educational Research*, (51), 215-228.

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.

Bailey, K. D. (1994). *Methods of social research*. New York: The Free Press.

Brennnan, R. L. (2001). *Statistics for social science and public policy generalizability theory*. Iowa: Springer.

Büyükkıdık, S., & Anıl, D. (2015). Performansa dayalı durum belirlemede güvenirliğin genellenebilirlik kuramında farklı desenlerle incelenmesi [Investigation of reliability in generalizability theory with different designs on performance-based assessment]. *Education and Science*, *40*(177), 285-296.

Can Aran, Ö., Güler, N., & Senemoğlu, N. (2014). Öğrencilerin disiplinli zihin özelliklerini belirlemede kullanılan dereceli puanlama anahtarının genellenebilirlik kuramı açısından değerlendirilmesi [An evaluation of the rubric used in determining students' levels of disciplined mind in terms of generalizability theory]. *Dumlupınar University Journal of Social Sciences*, (42), 165-172.

Cohen, R. J., & Swerdlik, M. E. (2009). *Psychological testing and assessment: an introduction to tests and measurement*. United States: McGraw−Hill Companies.

Crocker, L., & Algina, J. (2008). *Introduction to classical and modern test theory*. Ohio: Cengage Learning.

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.

Diederich, P. B. (1973). *Short-cut statistics for teacher-made tests*. New Jersey: Educational Testing Service.

Erkuş, A. (2014). *Psikolojide ölçme ve ölçek geliştirme-I temel kavramlar ve işlemler [Measurement and scale development in Psychology -I basic concepts and procedures]*. Ankara: Pegem Akademi.

Gao, X., & Brennan, R. L. (2001). Variability of estimated variance components and related statistics in a performance assessment. *Applied Measurement in Education*, *14*(2), 191-203. https://doi.org/10.1207/S15324818AME1402_5

Güler, N. (2009). Genellenebilirlik kuramı ve spss ile genova programlarıyla hesaplanan g ve k çalışmalarına ilişkin sonuçların karşılaştırılması [Generalizability theory and comparison of the results of g and d studies computed by spss and genova packet programs]. *Education and Science*, *34*(154), 93-103.

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

Haladyna, T. M. (1997). *Writing test items to evaluate higher order thinking*. Massachusetts: Allyn and Bacon.

Han, C. (2016). Investigating score dependability in english/chinese interpreter certification performance testing: a generalizability theory approach. *Language Assessment Quarterly*, *13*(3), 186-201. https://doi.org/10.1080/15434303.2016.1211132

Hoyt, W. T., & Melby, J. N. (1999). Dependability of measurement in counseling psychology: an introduction to generalizability theory. *The Counseling Psychologist*, *27*(3), 325-352. https://doi.org/10.1177/0011000099273003

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.

Karasar, N. (2006). *Bilimsel araştırma yöntemi [Scientific research method]*. Ankara: Nobel Akademik.

Kutlu, Ö., Doğan, C. D., & Karakaya, İ. (2014). *Ölçme ve değerlendirme performansa ve portfolyoya dayalı durum belirleme [Measurement and evaluation in education]*. Ankara: Pegem Akademi.

Lin, C.-K., & Zhang, J. (2014). Investigating correspondence between language proficiency standards and academic content standards: a generalizability theory study. *Language Testing*, *31*(4), 413-431. https://doi.org/10.1177/0265532213520304

Nalbantoğlu Yılmaz, 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.

Özberk, E. H., & Gelbal, S. (2014). Genellenebilirlik kuramı karar çalışmalarında kullanılan farklı varyans bileşenleri kestirim yöntemlerinin karşılaştırılması [Comparing different variance component estimation methods used in generalizability theory decision studies]. *Journal of Measurement and Evaluation in Education and Psychology*, *5*(2), 91-103.

Özgüven, İ. E. (2012). *Psikolojik testler [Psychological tests]*. Ankara: Nobel Akademik.

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

Suen, H. K. (1990). *Principle of test theories*. New York: Lawrence Erlbaum Associates.

Urbina, S. (2004). *Essentials of psychological testing*. New Jersey: John Wiley & Sons.

Wu, J., Hu, L., Zhang, G., Liang, Q., Meng, Q., & Wan, C. (2016). Development and validation of the nasopharyngeal cancer scale among the system of quality of life instruments for cancer patients (QLICP-NA V2.0): combined classical test theory and generalizability theory. *Quality of Life Research*, *25*(8), 2087-2100. https://doi.org/10.1007/s11136-016-1251-4

Yelboğa, A. (2008). Güvenirliğin değerlendirilmesinde genellenebilirlik kuramının kullanılması: endüstri ve örgüt psikolojisinde bir uygulama [The assessment of reliability with geııeralizabiiity theory: an application in industrial and organizational psychology]. *İstanbul University Journal of Psychology Study*, *28*(1), 35-54.

Yelboğa, A. (2012). Genellenebilirlik kuramı'na göre iş performansı ölçeklerinde güvenilirlik [Dependability of job performance ratings according to generalizability theory]. *Education and Science*, *37*(163), 157-164.

Yelboğa, A., & Tavşancıl, E. (2010). Klasik test ve genellenebilirlik Kuramına göre güvenirligin bir iş performansi ölçegi üzerinde incelenmesi [The examination of reliability according to classical test and generalizability on a job performance scale]. *Educational Sciences: Theory & Practice*, *10*(3), 1825-1854.

Copyrights

Copyright for this article is retained by the author(s), with first publication rights granted to the journal.

This is an open-access article distributed under the terms and conditions of the Creative Commons Attribution license (http://creativecommons.org/licenses/by/4.0/).