edam

ISSN 1303-0485 • **eISSN** 2148-7561
DOI 10.12738/estp.2015.3.2278
Copyright © 2015 EDAM • http://www.estp.com.tr
Educational Sciences: Theory & Practice • 2015 June • 15(3) • 635-645

Received | 19 November 2013
Accepted | 10 March 2015
OnlineFirst | 12 May 2015

Comparing the Effectiveness of SPSS and EduG using Different Designs for Generalizability Theory

Gulsen Tasdelen Teker[a]
Sakarya University

Nese Guler[b]
Sakarya University

Gulden Kaya Uyanik[c]
Sakarya University

Abstract
Generalizability theory (G theory) provides a broad conceptual framework for social sciences such as psychology and education, and a comprehensive construct for numerous measurement events by using analysis of variance, a strong statistical method. G theory, as an extension of both classical test theory and analysis of variance, is a model which can deal with multiple sources of error. In conducting the analysis of the G theory, there are several software programs that can be used such as GENOVA, SPSS, SAS, EduG, and G-String. In this study, the general perspectives of G theory are first explained broadly. Then, the SPSS and EduG software programs are used to conduct generalizability analyses on the data obtained from the answers of 30 students (p) to nine open-ended questions (i) as rated by three raters (r). There are three different designs in the study. Two of them are random effects designs, pxixr and pxi:r, and the last one is pxixr design using a fixed rater . According to the findings from the study, SPSS and EduG give the same results for variance component estimates as well as for G (Generalizability) and D (Decision) studies of all designs, as expected. Besides comparing the program outputs, their weaknesses and strengths were also discussed regarding different designs and data sets in this study.

Keywords: Generalizability Theory • G Study • D Study • SPSS • EduG

a Gulsen Tasdelen Teker (PhD), Department of Educational Sciences, Sakarya University, Sakarya, Turkey
Email: gtasdelen@sakarya.edu.tr

b **Corresponding author**
Assoc. Prof. Nese Guler (PhD), Department of Educational Sciences, Sakarya University, Sakarya, Turkey
Research areas: Measurement and evaluation in education; Generalizability theory; Statistics and research methods in social sciences
Email: gnguler@gmail.com

c Gulden Kaya Uyanik (PhD), Department of Educational Sciences, Sakarya University, Sakarya, Turkey
Email: guldenk@sakarya.edu.tr

G theory has formed a comprehensive structure by employing variance analysis which provides a broad conceptual framework for social sciences such as psychology and education (Brennan, 2000, 2001a; Cronbach, Gleser, Nanda, & Rajaratnam, 1972; Shavelson & Webb, 1991). It is also a powerful statistical tool for situations where there are numerous measurements. The theory, as an extension of classical test theory and variance analysis, stands as a model where multiple sources of error can be handled (Brennan, 2001a; Shavelson & Webb, 1991).

Generalizability (G) Theory

The reliability of measurement results in education and psychology was previously examined using classical test theory (CTT) in general. It is assumed in CTT that the observed score is composed of the actual score with no separable score for error. The restriction of this assumption, especially in performance measurements where the probability of the existence of more than one source of error is high, reveals the importance of G theory in which more than one source of error is handled and can be predicted simultaneously (Brennan, 2000). Another advantage of G theory in using performance assessment is that while there is a restrictive parallel assumption in CTT, randomly parallel assumption is adopted in G theory (Brennan, 2011; Kretchmar, 2006). The main aim of G theory is to generalize the scores of a specific measurement tool from a specific group to the universe of generalization which consist of 1) the universe of admissible observations and generalizability studies (G studies), 2) the universe of G studies and decision studies (D studies). While G studies provide an estimate of the generalizability coefficient of variances from all facets and this coefficient includes the examinee's universe score, D studies enable one to examine the interactions among all applicable facets (tasks, raters, observations, etc.) and the subject of measurement for calculating the dependability coefficient (Brennan, 2000; Crocker & Algina, 1986; Hsu, 2012).

G theory has four main advantages compared to CTT. *1*) It provides simultaneous evaluation of test-retest reliability, internal consistency, inter-rater reliability, and convergent validity. *2*) It enables estimates of both individual measurement facets and interaction effects. *3*) When assessing an examinee's performance, it gives information about the quality of their absolute structural level of knowledge as well as ranking this information in order. *4*) It allows researchers to optimize the reliability of an assessment within the cost constraints of time and money. For example, assessment developers can provide information about how many items, how many raters, and how many occasions are needed to reach a reliable result (Yin & Shavelson, 2008).

When looking at the historical evolution of G theory, its basic principles were first discussed in articles published by Cronbach, Rajaratnam, and Gleser in 1963 and 1965. Indeed, the use of variance analysis in reliability studies started before the work of Cronbach and his colleagues. Burt in 1936, Hoyt in 1941, and Jackson and Ferguson also in 1941 discussed the use of variance analysis in the prediction of reliability. Then the contributions made by Alexander (1947), Ebel (1951), Finlayson (1951), Loveland (1952), and Burt (1955) followed, as cited in Brennan (1992). These were then followed by the book entitled "The Dependability of Behavioral Measurement" by Cronbach, Gleser, Nanda, and Rajaratnam in 1972, which was an extended form of generalizability theory.

In 1983, Brennan's book "Elements of Generalizability Theory" was published. Crick and Brennan designed a computer program called "A Generalized Analysis of Variance System (GENOVA)" in the same year. However, because the theory and the program prepared for it seemed too complex for users, studies concerning the theory remained limited until 1991. Later in 1991, Shavelson and Webb published their book "Generalizability Theory: A Primer," which made the basics of G theory more understandable and the theory more applicable for relevant research studies. Thus the spread of the theory accelerated. With the book "Generalizability Theory," published in 2001, Brennan discussed univariate as well as multivariate G theory in detail and developed the mGENOVA program for multivariate analyses (Brennan, 2001b) and the urGENOVA computer program for use in the prediction of variance components of random effects in balanced and unbalanced designs (Brennan, 2001c).

Although the programs developed by Brennan are frequently used by those working on G theory, the complexity of the programs caused problems, especially for first-time users. The lack of a user-friendly computer program restricted the widespread use of G theory for a long time (Cardinet, Johnson, & Pini, 2010; Guler, Kaya Uyanik, & Tasdelen Teker, 2012). In search of finding a solution to this problem, Duquesne, a Belgian researcher, developed a program called ETUDGEN in the early 1980's. Even though the program met the basic academic needs of researchers in the field, it remained limited in respect to conducting some applications (Cardinet et al.,

2010). Then, Mushquash and O'Connor presented their syntaxes which they wrote in order to perform generalizability analyses for the SPSS, SAS, and MATLAB programs in their article in 2006. By means of these instructions, generalizability analyses are easily able to be performed with the SPSS, SAS, and MATLAB programs which are frequently used by many researchers. Their syntaxes are accessible for free at https://people.ok.ubc.ca/brioconn/gtheory/gtheory.html. Aside from being able to conduct G and D studies, it is also possible with these programs to produce graphs for absolute error variances, relative error variances, the G coefficient, and the phi coefficient if desired. There are two different syntax groups for generalizability theory analysis through SPSS: G1 and G2. By using G1, it is possible to conduct analysis easily for single and two-facet situations. By using G2, it is possible to analyze designs with more than two-facets, but it is a bit complex. Even though two-facet crossed and nested designs can be easily conducted using the G1 syntax on SPSS, SAS, and MATLAB programs, three or more facet crossed designs can be conducted only on the SPSS and SAS programs by using the more complex syntax, G2 .

The EduG was developed by Cardinet in 2006 as a versatile, user-friendly program for performing generalizability analyses (Cardinet et al., 2010). The program is available at http://www.irdp.ch/edumetrie/englishprogram.htm for free. At the same web address, a user's guide, "Applying Generalizability Theory Using EduG," with explanations and illustrative pictures is also available to users. The number of facets to be included in analyses performed on these programs can be increased as much as desired according to the scope of the research. In other words, there is no limit to the number of facets in performing analyses through this program. However, the lack of graphs in the program output stand as a limitation of the program (Guler et al., 2012).

There have been many studies conducted in the literature that used G theory. When one investigates the details of those studies by way of the software programs used, the variety can easily be seen. On the other hand, the researches which have investigated program features like ease of use, differences, or similarities among them are quite restricted. For instance, Musquash and O'Connor (2006) stated that the reason G theory is infrequently used compared to classical test theory is that software programs like SPSS and SAS are not widely used. Therefore, they conducted a study using SPSS, SAS, and MATLAB programs for generalizability theory analysis. This was not a comparison study. The main aim of the study was to introduce the written forms of syntaxes for popular statistical software to increase the use of the theory. For this purpose, analysis was conducted on the data obtained from the Rosenberg Self-Esteem Scale (N = 329). The program output included variance components, relative and absolute errors, and generalizability coefficients. Aside from these, the obtained graphs for the decision studies were also given. Moreover, the syntaxes were also given as an appendix at the end of the study for further researchers who wanted to use these software programs for generalizability analysis.

Derstine (2007) compared two software programs for G theory studies by means of usability and features of each program as well as the results of G and D studies using different measurement designs to determine which program, if any, was best. One of the programs was G-String, which provides a familiar, Windows-like interface with urGENOVA, and the other one was EduG which is a fully functional stand-alone program for G-studies with a user-friendly interface. As a result of the study, it was concluded that both programs were easy to use. However, although EduG was the recommended program for all balanced designs because of many additional features it offered for manipulating data, it was impossible for it to be used for unbalanced designs. On the other hand, although G-String could be used for all unbalanced designs, there were some limitations about crossed, mixed-model balanced design applications of the program. Aside from these findings, no recommendation was made as to which program should be used for nested, mixed-model designs where the nested facet is fixed.

Guler (2009) conducted a study in which generalizability and dependability coefficients for both generalizability study and decision studies were presented by using both GENOVA and SPSS computer packet programs. The outputs of SPSS and GENOVA were almost the same. There were just very small differences in variance components because of the mathematical rounding-off of the programs. As a result of that study it was suggested that while doing analysis for G theory, SPSS could be used because of the practical application of the program in place of the more complex GENOVA.

Yelboga (2011) discussed G theory analysis using GENOVA, SPSS, and SAS programs through an illustrative example of data. The results of variance components, relative and absolute errors, generalizability coefficient, coefficients obtained

from D studies, and graphs of D studies were compared via the program outputs. Finally, it was found that the results of the three statistics programs were quite similar.

Nalbantoglu Yilmaz (2014) compared the outputs of G-String and EduG programs to determine which program was more suitable to use for G theory analysis. The variance components of the main and interaction effects, relative and absolute error variances, generalizability and dependability coefficients obtained from the two programs were the same when using the same design. On the other hand, when the object of measurement was nested within facets, the reliability coefficients calculated by both programs were found to be different due to the handling aspects of the universe score in the calculation formula. Moreover, this difference was found to be more apparent for the phi coefficient. Therefore, the study suggested that when the object of measurement is nested within facets, it should be remembered that the program, G-String, gives lower reliability coefficients for absolute decisions.

Ogretmen and Acar (2014) investigated the estimation of the G coefficient by using LISREL, SPSS, and EduG programs using two different sets of data. There were three variance sources for two data sets within the design of px(i:r). Although there were the same number of items and raters (five and three respectively) for two of the data sets, the number of the third variance source differed from one data set to another: 50 people for the first one and 20 for the second one. For the first data set, although very similar G coefficients were found from LISREL (0.639) and SPSS (0.640), the EduG result (0.740) was a bit higher than the other two programs. On the other hand, all the G coefficients for the second data set were quite similar to each other (0.738 for LISREL, 0.737 for SPSS, and 0.740 for EduG). The obtained G coefficients were also transformed to Fisher's z-statistic and tested with the z-test. As a result, it was found that there was no significant difference between the G coefficients obtained from LISREL, SPSS, and EduG. The limitation of LISREL is that it only produces G-study outputs, without producing an output for D studies. Therefore, it was concluded that the SPSS and EduG programs were more useful.

Problem Statement

Although there are several software programs, some of them are quite complex for performing G theory analyses. As a result, the widespread use of G theory was restricted for many years. This study both highlights the main points of G theory and offers two alternative programs through which generalizability analyses can be conducted. One of the alternatives is a syntax enabling analysis of generalizability using the SPSS program designed by Mushquash and O'Connor (2006), and the other is the EduG program developed by Cardinet in 2006 as a user-friendly and versatile program for performing analyses of G theory (Cardinet et al., 2010). In this research, the two programs were employed and the values obtained from generalizability (G) and decision (D) studies were presented together and compared. Moreover, the weaknesses as well as strengths of the two programs, especially for different data sets and different designs, are discussed.

Method

In order to compare the differences in the programs which performed G analyses, three situations with different designs were handled and analyzed in this study.

The research data was composed of statistics exam scores received from 30 undergraduate students (17 women, 13 men) studying in the department of Psychological Counseling and Guidance in the Educational Faculty of Sakarya University during the 2013-2014 academic year. The statistics mid-term exam was used as a data collection tool and it consisted of nine open-ended questions. Three raters who are experts in statistics and work as academic staff at the university scored the exam. In order to prepare the answer key for the test, the raters answered the items separately and then compared their answers. Consequently, they agreed on the common answers for the answer key. Moreover, in case answers provided by students required comments from the raters, all potential answers were also noted. Thus, an answer key was jointly formed and the raters used this common answer key to independently grade the nine items between 1 to 10 points.

In this case, students were considered as the object of measurement while the items and raters were considered as facets. Change (variance) arising from students taking part in the research is a desired condition because it shows the differences inherent in students. As such, it is not taken as a 'source of error' (facet) in G theory as it is in CTT. Each probable source of error lying outside the object of measurement and having similarities is defined as a facet (Guler et al., 2012).

In the first situation, the scores assigned by the three different raters to the answers of the 30 students for the nine open-ended questions from the statistics mid-term exam were analyzed using G theory. In this situation, because each student (*p*) taking part in the research answered all the items *(i)* and because all student answers were scored by the three raters (*r*) included in the research, the two-facet crossed random design [*p x i x r*] was used.

In the second situation, the data coming from the responses of the same 30 students for the nine items were also used. Different from the first one, the first three items were scored by the first rater, the second three items by the second rater, and the final three items were scored by the third rater. Here, all the students answered all of the items in the examination, but the items scored by the raters differed. Thus, the design handled in the second situation was a two-facet nested random design [*p x (i : r)*].

With G studies conducted for both situations as mentioned above, the separate variance values of each source of variation, the variance values of their interactions, and the G and phi coefficients were all calculated both on the SPSS and EduG programs. Additionally, the G and phi coefficients obtained as a result of the D studies were calculated on both programs for both situations and the results were compared.

Additionally, the differences that emerged from performing analyses of G theory through the SPSS and EduG programs in mixed-design situations were also addressed in this study as a third situation. The data from the first situation was used for this. Unlike the first situation, however, the items included in the third situation were regarded as random and the raters as a fixed facet. The purpose in G theory is usually to be able to generalize facets such as items and raters into a population beyond the conditions available in the study. However, this sometimes may not be the case for all facts. For instance, in the process of measuring a performance, there may be raters available only in the study due to financial or logistical reasons, and the purpose may not be to generalize the raters into a bigger population apart from them. In this case, the rater facet is considered as a fixed facet in the study. Yet, G theory is based on considering facets as random. Therefore, at least one facet should be regarded as random. That is to say, it is impossible to conduct a G study in which all facets are fixed. The designs in which both random and fixed facets are available are called mixed designs (Brennan, 1992; Guler et al., 2012).

Finally, the strengths and weaknesses of the programs are also discussed in this research.

Results

This part highlights the results of G and D studies conducted by the SPSS and EduG programs for crossed random *p x i x r* , nested random *p x (i : r)* and mixed *p x i x r* designs, which are dealt with in the scope of the research. Since the aim of the study is just to compare the program features, the explanations of results for all situations are not given. For example, only results from the first situation are explained by means of G study, D study, and obtained graphics.

Situation 1: [*p x i x r*] Results of the Generalizability and Decision Studies on Random Facets

Table 1 shows the SPSS and EduG results concerning the variance values of the sources of variation and the interactions between them for the crossed two-facet random design (*p x i x r*).

An examination of Table 1 shows that there were no differences between the mean square averages

Table 1
*Variance Estimations of Two-Facet Crossed Random Design related to the SPSS and EduG Programs**

Variance source	*df*	Mean Square		Variance		Variance Proportion/Percentage	
		SPSS	EduG	SPSS	EduG	SPSS	EduG
p	29	172.705	172.70481	5.685	5.68549	.401	40.1
i	8	14.730	14.72994	.000	-.06842	.000	0.0
r	2	5.946	5.94568	.000	-.02302	.000	0.0
pi	232	15.952	15.95184	3.991	3.99079	.281	28.1
pr	58	7.224	7.22410	.361	.36052	.025	2.5
ir	16	8.915	8.91512	.165	.16452	.012	1.2
pir	464	3.979	3.97946	3.979	3.97946	.281	28.1

**Note.* The formats used in the values given in the table were left untouched (as they were obtained from the SPSS and EduG programs) to make the differences between program outputs more distinctive.

calculated by SPSS from those calculated by EduG for the *p x i x r* design. It was also observed after examination of the variance values that the variance values from SPSS converted the negative predicted variance values to zeroes whereas EduG left them as they were. On examining the variance values, they were found to be the same as expected because both of the programs were using the same formula. On the other hand, the programs were found handle variance values differently, as a proportion in SPSS and as a percentage in EduG.

As can be seen from Table 1, the estimated variance components and proportion or percentage of total variances was reported. The variance component for students, which is the largest component of all, (σ^2_p = 5.685) was interpreted as the estimated variance of student mean scores, where each mean is the overall items and tasks in this measurement process. The estimated variance of the item mean scores in this situation was 0, which suggests that there was not any difference in difficulty for the items. Similarly, σ^2_r = 0 is the estimated variance of the mean scores from the raters, where each mean is the overall number of students in the population and all items in the process.

Interpretation of the variance components from the interactions are more complex. For example, σ^2_{pi} = 3.991 is considerably greater than zero, which indicates that there was a considerably different rank ordering for the student mean scores for each of the various items. In contrast, σ^2_{ir} = .165 is close to zero, which means that the various raters rank ordered the difficulty of the items similarly. The last variance component σ^2_{pir} = 3.979 is the third largest component, residual error, and includes the interaction of students by items from raters and all other unexplained source of variation. This variance component is also named as a residual.

Table 2 shows the absolute and relative errors obtained by the SPSS and the EduG programs for the *p x i x r* design as well as the phi and G coefficients.

Table 2
*Results of Two-Facet Crossed Random Design related to the SPSS and EduG Programs**

Program	Absolute Error	Relative Error	Phi Coefficient	G-Coefficient
SPSS	.717	.711	.888	.889
EduG	.71707	.71098	.89	.89

**Note.* The formats used in the values given in the table were left untouched (as they were obtained from the SPSS and EduG programs) to make the differences between program outputs more distinctive.

On examining Table 2, no difference is seen between the absolute and relative errors obtained from SPSS and EduG, nor was there any difference for the phi and G coefficients for the *p x i x r* design. While absolute error is defined as the difference between a student's observed score and universe score, relative error indicates the difference between a student's observed deviation score and their deviation from their universe score. Additionally, the phi coefficient is equal to the ratio of variance in universe score (variance of students in here) to itself plus absolute error variance. Similarly, the generalizability coefficient, which is analogous to a reliability coefficient in CTT, is the ratio of universe score variance to itself plus relative error variance (Brennan, 2000).

The SPSS and EduG results concerning the D studies performed for the crossed two-facet random design are shown in Table 3. In Table 3, both the phi and G coefficients that were obtained by changing the number of raters while the number of items was nine are presented, and the phi and G coefficients that were obtained by changing the number of items while the number of raters was three are compared.

On examining Table 3 it is found that there was no difference between the phi and G coefficients that were obtained by "changing the number of raters while the number of items was nine or by changing the number of items while the number of raters was three."

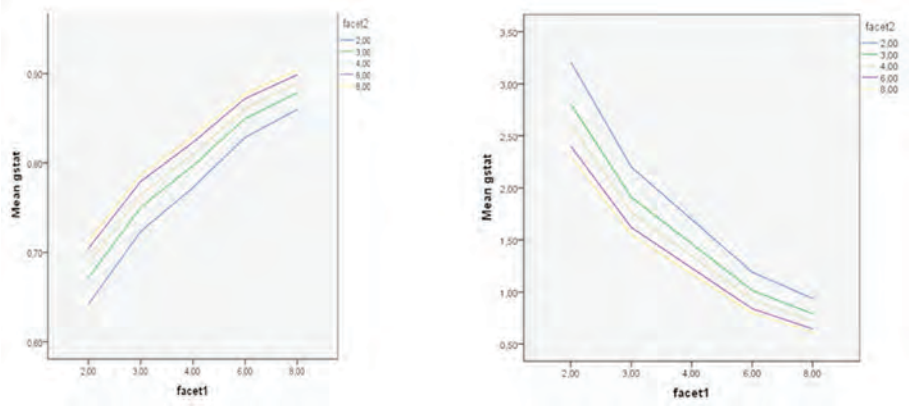
In table 3, different D study scenarios are illustrated. In the first one, while the number of items remained fixed at nine, the number of raters was changed. For example, according to two raters and the nine items, the phi and G coefficients were estimated at .869 and .871 using SPSS, and .87 and .87 using EduG, respectively. In the second one, while the number of raters remained at three, the items differed, ranging from two to eight. So it can easily be seen that the phi and G coefficients were calculated at .888 and .889 using SPSS, and .89 and .89 using EduG when n_i = 4 and n_r = 3, respectively. As a consequence of the D study results, it can easily be decided which condition would be the most effective for future measurement procedures.

Another comparison of the two programs used for generalizability theory is apparent in the graphs that the SPSS program yields unlike EduG. In SPSS, the values of the D studies conducted in relation to the G and phi coefficients give graphs of absolute and relative error variances separately. Graphs of the G coefficient and relative error variance are represented as an example in Figure 1. These graphs were obtained from the results concerning the decision study for the *p x i x r* design.

Table 3
*D Study Results of Situation 1**

Program	Number of Raters (all in $n_i = 9$)									
	2		3		4		6		8	
	Phi	G	Phi	G	Phi	G	Phi	G	Phi	G
SPSS	.869	.871	.888	.889	.898	.898	.907	.908	.912	.913
EduG	.87	.87	.89	.89	.90	.90	.91	.91	.91	.91
Program	Number of Items (all in $n_r = 3$)									
	2		3		4		6		8	
	Phi	G	Phi	G	Phi	G	Phi	G	Phi	G
SPSS	.748	.750	.848	.850	.888	.889	.909	.910	.922	.923
EduG	.75	.75	.85	.85	.89	.89	.91	.91	.92	.92

**Note.* The formats used in the values given in the table were left untouched (as they were obtained from the SPSS and EduG programs) to make the differences between program outputs more distinctive.



G Coefficients Relative Error Variances

Figure 1: Graphics of crossed design D studies of situation 1.

As seen in Figure 1, it can be said that the increase in the number of items can be more effective than the increase in the number of raters if this is practical for future studies.

Situation 2: [*p x (i : r)*] Results of the Generalizability and Decision Studies on Random Facets

Table 4 shows the SPSS and EduG results concerning the variance values of each source of variability and the interactions between them for the two facet random design [*p x (i : r)*], where the items are nested on raters.

A close examination of Table 4 makes it clear that there is no difference between the mean square averages calculated for the *p x (i : r)* design through the SPSS and EduG programs.

Table 5 shows the absolute and relative errors obtained from the SPSS and EduG programs for the two-facet random design as well as the phi and G coefficients.

Table 4*
Variance Estimations of Two-Facet Nested Random Design related to SPSS and EduG Programs

Variance sources	*df*	Mean Square		Variance		Variance Proportion/Percentage	
		SPSS	EduG	SPSS	EduG	SPSS	EduG
p	29	57.029	57.02937	5.003	5.00345	.401	40.1
r	2	34.493	34.49259	.254	0.25411	.020	2.0
i:r	6	4.437	4.43704	.000	-0.01252	.000	0.0
p:r	58	11.998	11.99834	2.395	2.39527	.192	19.2
pi:r	174	4.813	4.81252	4.813	4.81252	.386	38.6

**Note.* The formats used in the values given in the table were left untouched (as they were obtained on the SPSS and EduG programs) to make the differences between program outputs more distinctive.

Table 5*
Results of Two-Facet Nested Random Design related to SPSS and EduG Programs

Program	Absolute Error	Relative Error	Phi Coefficient	G-Coefficient
SPSS	1.418	1.333	.779	.790
EduG	1.41785	1.33315	0.78	0.79

**Note.* The formats used in the values given in the table were left untouched (as they were obtained on the SPSS and EduG programs) to make the differences between program outputs more distinctive.

On examining Table 5, no difference was found between the absolute and relative errors, or the Phi and G coefficients which were obtained from the SPSS and the EduG programs for the *p x (i : r)* design.

The SPSS and EduG results for the D studies conducted in relation to the two-facet random design where raters are nested are shown in Table 6. The phi and the G coefficients obtained by changing the number of raters while the number of items stayed at three are compared, and the phi and G coefficients obtained by changing the number of items while the number of raters remained at three are also compared in Table 6.

An examination of Table 6 demonstrates that there are no differences between the phi and G coefficients obtained in the scenarios of "changing the number of raters while the number of items stays at three, and changing the number of items while the number of raters stays at three."

The graphs of the G coefficient and relative error variance obtained from the results concerning the decision study for the *p x (i : r)* design are shown in Figure 2.

Situation 3: The [*p x i x r*] Results of the Generalizability Study on Fixed Facet

Because the researchers had aimed to generalize all of the items and the raters into a larger population rather than the population in which the generalizability study was performed in both situations, it was assumed that all the sources of variability were random. However, in some cases it is impossible to generalize a facet into the external conditions available in the study, or such an aim might not even be held. In such cases, the facet is considered to be fixed and the models with at least one fixed facet are defined as *mixed models*.

Generalizability Study on Fixed Facet through SPSS

The analysis of generalizability studies through SPSS is performed in three basic steps. Firstly, variance analysis is conducted by considering all sources of variability as random, thus predicting the variance components. These variance values are shown in Table 1 which was derived from situation 1. Next, one determines the common variance components to be calculated with the random part of the mixed design. In order to derive the variance components in our example, error (*pi,e*) calculations were performed with students (*p*) and items (*i*), which are outside the fixed facet of rater (*r*) in this example, as well as on the interaction between them. For the purposes of distinguishing these variance values from the ones in situation 1, they will be represented as $\sigma^2_{p^*}$, $\sigma^2_{i^*}$, σ^2_{pi,e^*}. Finally, it is necessary to calculate these variance values. The values can be predicted with the help of the equations below.

$$\sigma^2_{p^*} = \sigma^2_p + \frac{1}{n_r}\sigma^2_{pr} = 5{,}685 + \frac{1}{3}0{,}361 = 5{,}805$$

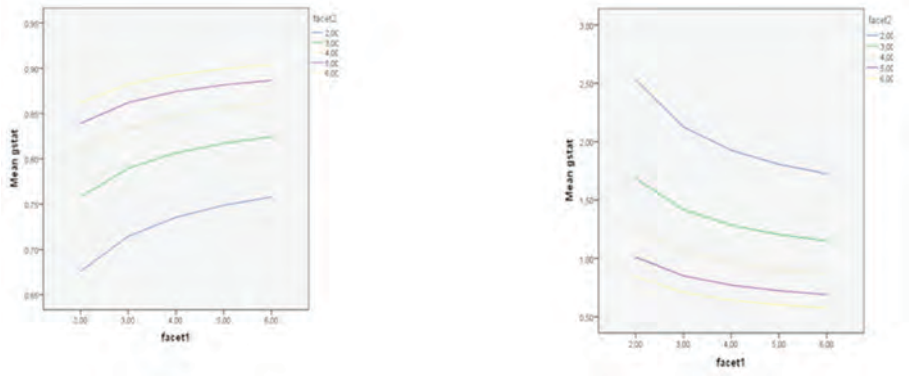
$$\sigma^2_{i^*} = \sigma^2_i + \frac{1}{n_r}\sigma^2_{ir} = 0 + \frac{1}{3}0{,}165 = 0{,}055$$

$$\sigma^2_{pi,e^*} = \sigma^2_{pi} + \frac{1}{n_r}\sigma^2_{pir,e} = 3{,}991 + \frac{1}{3}3{,}979 = 5{,}317$$

Table 6
*D Study Results of Situation 2**

Program	Number of Raters (all in $n_i = 3$)									
	2		3		4		5		6	
	Phi	G	Phi	G	Phi	G	Phi	G	Phi	G
SPSS	.702	.714	.779	.790	.825	.833	.855	.862	.876	.882
EduG	.70	.71	.78	.79	.82	.83	.85	.86	.88	.88
Program	Number of Items(all in $n_r = 3$)									
	2		3		4		5		6	
	Phi	G	Phi	G	Phi	G	Phi	G	Phi	G
SPSS	.748	.758	.779	.790	.796	.807	.806	.817	.813	.824
EduG	.75	.76	.78	.79	.80	.81	.81	.82	.81	.82

**Note.* The format of the values given in the table were left untouched (as they were obtained on the SPSS and EduG programs) to make the differences between program outputs more distinctive.



G Coefficients Relative Error Variances

Figure 2: Graphics of nested design D studies of the situation.

Generalizability Study on Fixed Facet through EduG

So as to perform the analyses in mixed measurement designs having one fixed facet through the EduG program, there is no need to do manual calculations as in SPSS. The fixed and random facets are determined while describing the levels of the facets in the program. As the program screen in Figure 3 shows, the populations of the random facets were described with the letters "INF" (infinite) whereas the population of the fixed facet was described as "3" for the number of raters in our example. Thus, while the item was described as random, the rater facet was described as fixed.

The output obtained after performing analyses through EduG is shown in Table 7. On examining the values it was found that the variance values for students affected by the fixed facet and for student-item interaction did not differ in both program outputs. Yet, the variance values for the item facet differed. The difference stems from the way the programs handle the variance value, which is a negative number.

Because it is impossible for variance to be negative, Cronbach et al. (1972) recommend that zero should be taken instead of a negative value in such a case. The SPSS program performs analyses based on this view. Brennan (1992) also suggests, as in the previous approach, that zero should be taken instead of a negative variance. Different from the previous approaches, however, the author recommends that operations should be done using the negative variance value for calculating all other variance components (Guler et al., 2012). In other words, it is suggested that after calculating all the variance components using the negative values, the negative values should then be replaced by zero. The EduG program performs analyses using this view.

Table 7*
Variance Estimations of Two-Facet Mixed Design related to the SPSS and EduG Programs

Variance Source	SPSS	EduG
p	5.805	5.80567
i	.055	-.01358
r	.000	-.02302
pi	5.317	5.31728
pr	.361	.36052
ir	.165	.16452
pir	3.979	3.97946

**Note.* The format of the values given in the table were left untouched (as they were obtained from the SPSS and EduG programs) to make the differences between program outputs more distinctive.

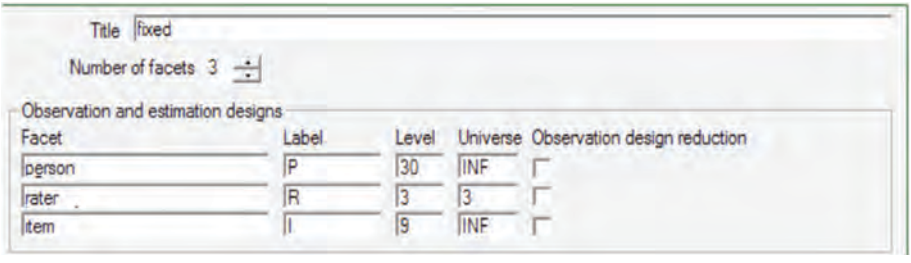


Figure 3: EduG Window.

The results of mixed measurement designs having one fixed facet through SPSS and EduG is given in Table 7.

Since the programs used in this study utilize different approaches while calculating variances, the obtained results given in Table 7 are different for EduG and SPSS.

Discussion

The generalizability theory analyses in this research were performed using the SPSS and the EduG programs, two different and user-friendly programs. The same results were obtained in the variance values of the G and phi coefficients predicted through G study, and in the G and phi coefficients obtained through different scenarios in the D study. Since both programs were based on the same statistical model, obtaining the same results was not entirely unexpected. According to a few different studies conducted about program comparisons of G theory, there was no difference in the results of variance values (Derstine, 2007; Guler, 2009; Nalbantoglu Yilmaz, 2014; Yelboga, 2011). Although, the same results of G and Phi coefficients were obtained for many designs, Nalbantoglu Yilmaz (2014) found some differences for the designs where the object of measurement was nested within the facets. Since there was no difference for most of the designs, it would be convenient to choose a program on the basis of its strengths and weaknesses.

Firstly, SPSS is not a free program. Therefore, it is necessary to have a licensed program to conduct analyses with it. However, since SPSS as one of the most widely-used statistical package programs is available to many researchers, performing analyses through this program will not cause an extra burden on researchers. Moreover, it is also possible to download the syntax of G theory written by Mushquash and O'Connor (2006) for free. EduG, on the other hand, is freeware. Therefore it is quite easy to reach this program.

The most remarkable advantage of performing G theory analyses with the SPSS program is the graphs it yields in contrast to EduG. In SPSS, the values from the D studies conducted separately for relative and absolute error variances, and the G and phi coefficients are also presented in graphs. The only limitation of SPSS about graphs is that it only works for random designs, not mixed ones.

One restriction of SPSS is the maximum number of two facets using the simple syntax (G1). Because of the likelihood of working on situations having more than two facets for analyses in social sciences, such a limitation makes conducting studies difficult for higher numbers of facets, which can only be done using G2, a complex syntax. In conducting G theory analyses through the EduG program, there are no restrictions on the number of facets required to be considered simple. This property may be thought of as the most important advantage of this program.

Whereas the negative variance values obtained in the EduG program outputs are presented as program output, the negative variance values are reduced to zeroes in SPSS outputs.

Even though there are no differences between the two programs in relation to data input, there are differences in the order of facets to be considered. The object of measurement considered in generalizability studies is not described as a facet, and the SPSS program also works in this way. In a sample situation with two facets, the facet which changes the most among all facets having sources of variability falling outside the object of measurement is entered as the first facet while the one that changes the least is entered as the last facet in the program. That is to say, the facets are entered into the program in the order of most changing to least changing. In conducting analyses through the EduG program, however, two facets are described for a sample situation with one facet, and three facets are described for a situation with two facets. Thus, the object of measurement is also entered in the program as a facet. Unlike the SPSS, the facets are entered in the program in the order of least changing to most changing.

Finally, in fixed facet designs, manual calculations need to be done in addition to the analyses performed through the SPSS program, but this is not the case with EduG as the program yields the desired results.

As a suggestion for researchers who are interested in G theory, for unbalanced designs where the number of levels of a nested facet varies for each level producing an unequal number of levels, SPSS and EduG are not recommended for use. Since both of them require manual calculations, for unbalanced designs, urGENOVA or a more user-friendly program like G-String could be used instead of SPSS and EduG.

References

Brennan, R. L. (1992). *Elements of generalizability theory* (rev. ed.). Iowa City, IA: American College Testing.

Brennan, R. L. (2000). Performance assessment from the perspective of generalizability theory. *Applied Psychological Measurement, 24*(4), 339-353.

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

Brennan, R. L. (2001b). *Manual for mGENOVA Version 2.1.* Iowa Testing Programs Occasional Papers Number 50. Iowa City, IA: Iowa Testing Programs, University of Iowa.

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

Brennan, R. L. (2011). Generalizability theory and classical test theory. *Applied Measurement in Education,. 24*, 1-21.

Cardinet, J., Johnson, S., & Pini, G. (2010). *Applying generalizability theory using EduG*. New York, NY: Routledge - Taylor & Francis Group.

Crick J. E., & Brennan R. L. (1983). Manual for GENOVA: A generalized analysis of variance system. Iowa City, IA: The American College Testing Program.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. New York, NY: Harcourt Brace Javanovich College Publishers.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. New York, NY: Wiley.

Derstine, P. (2007, August). *Which software program for generalizability studies is best? Comparing G_String II and EduG.* Paper presented at the eighth Annual Master of Health Professions Education Summer Conference, College of Medicine West, USA.

Guler, N. (2009). 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.

Guler, N., Kaya Uyanik, G., & Tasdelen Teker, G. (2012). *Generalizability theory*. Ankara: PegemAkademi Publishing.

Hsu, L. (2012). Applications of generalizability theory to estimate the reliability of EFL learners' performance-based assessment: A preliminary study. *Educational Research, 3*(2), 145-154.

Kretchmar, J. (2006). Assessing the reliability of ratings used in undergraduate admission decision. *Journal of College Admission, 192*, 10-15.

Mushquash, C., & O'Connor, B. P. (2006). SPSS and SAS programs for generalizability theory analysis. *Behavior Research Methods, 38*(3), 542-547.

Nalbantoglu Yilmaz, F. (2014, June). *Which program suitable to use G Theory analysis?* Paper presented at the fourth annual Conference of Measurement and Evaluation in Education, Hacettepe University, Turkey.

Ogretmen, T., & Acar, T. (2014). Estimation of generalizability coefficients: An application of structural equation modeling. *Journal of Education and Practice*, *5*(14), 113-119.

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

Yelboga, A. (2011, July). *Investigation of generalizability theory analysis results with different statistical programs.* Poster presented at the XII. European Congress of Psychology, Istanbul, Turkey.

Yin, Y., & Shavelson, J. (2008). Application of generalizability theory to concept map assessment research. *Applied Measurement in Education, 21*, 273-291.