International Journal of Assessment Tools in Education
2019, Vol. 6, No. 5-Special Issue, 57–69
https://dx.doi.org/10.21449/ijate.658747
Published at http://www.ijate.net http://dergipark.gov.tr *Research Article*

# Coping with Unbalanced Designs of Generalizability Theory: G String V

**Gülşen Taşdelen Teker** [1,*]

[1]Hacettepe University, Faculty of Medicine, Department of Medical Education and Informatics, Ankara, Turkey

**ARTICLE HISTORY**

*Received: 28 October 2019*
*Accepted: 15 December 2019*

**KEYWORDS**

Generalizability theory,
G String V,
Unbalanced design

**Abstract:** The aim of this paper is to introduce a software that is appropriate for the generalizability theory for not only balanced but also unbalanced data sets. Because it is possible to have unbalanced data sets while conducting a study, the researchers have devised an easy solution, other than deleting data, to balance the design to cope with this situation. Thus, the software G String V will be introduced. First, the generalizability theory will be reviewed, followed by a description of the unbalanced synthetic data that was used to conduct the analysis using the software. Explanations are provided for installing the software, preparation of the data, and the step-by-step data analysis. Moreover, the interpretation of the data is also explained. Finally, the limitations of the software are shared.

## 1. INTRODUCTION

Generalizability (G) theory, which was developed by Cronbach, Gleser, Nanda and Rajaratnam (1972) as an alternative to the classical test theory (CTT), is a statistical theory for evaluating the dependability or reliability of behavioral measurements (Brennan, 2001a; Shavelson and Webb, 1991). Conceptually, the G theory can be regarded as a multifaceted extension of the CTT and can be seen as a combination of the CTT and variance analysis (Brennan, 2000; Suen & Lei, 2007). Reliability, which is defined in the CTT as the consistency of the scores obtained through measurements, can vary according to the source to which the error is connected. For example, (i) when designing a reliability study to produce two sets of observations, one might give the same test two times, separated by two weeks: test-retest reliability; (ii) designing a reliability study to create two parallel forms of the test, as Form 1 and Form 2, and give the two forms of the test on the same day: parallel forms reliability; or (iii) calculating the reliability of a single form of a test on a single occasion: split-half reliability. Although there are multiple sources of error for these three examples, the CTT takes only one error source as time, forms, and items, respectively. In other words, the errors in the measurement results are considered as the errors coming from only one source of variability, and this emerges as a restriction of the CTT. Because the G theory can consider several sources of error simultaneously, estimations can be made more accurately than the ones in the CTT.

CONTACT: Gülşen TAŞDELEN TEKER ✉ gulsentasdelen@gmail.com Hacettepe University, Faculty of Medicine, Department of Medical Education and Informatics, Ankara, Turkey

ISSN-e: 2148-7456 /© IJATE 2019

G theory is structured around different sources of variation, called facets. There are two main types of facets, which are *facet of differentiation* and *facet of instrumentation*, according to Brennan's (2001) classification. In G theory, a behavioral measurement is conceived of as a sample from a u*niverse of admissible observations*, which consists of all possible observations on an *object of measurement* that a decision maker considers to be acceptable substitutes for the observation in hand. The object of measurement is also referred to as the facet of differentiation. You can easily determine the object of measurement of your research by answering the question: "What are you trying to attach the measurement to?" The answer can be students, items, etc. For instance, if you want your friends (n=10) to rate a number of dark chocolates (n=3) on five three-point scales, the object of measurement is dark chocolate. If you want to get patients' satisfaction with their experiences on a hospital's inpatient ward, the object of measurement is ward. Alternatively, as a more familiar example, if you want to evaluate the students' performances in a classroom activity, then the object of measurement is student. Whatever the object of measurement is, there is only one per analysis.

Although there are some researchers who use Brennan's classification of facets (Cardinet, Johnson & Pini, 2010: Cardinet, Tourneur & Allal, 1981), the facet of instrumentation is also referred to as the facet of generalization in some sources (Bloch & Norman, 2015; Bloch & Norman, 2012; Cardinet, Tourneur & Allal, 1976). This is acceptable and is also used in G String terminology. Every observation of object of measurement is subject to error, derived from various sources (Bloch & Norman, 2015). These sources are also called *facets of instrumentation* and address the following question: "To what extent can I generalize a measurement from one situation to another with a different level of the facet of instrumentation?" There are two types of instrumentation facets, called fixed and random. Typically, a random facet is created by randomly sampling levels of a facet. Meanwhile, when the levels of a facet have not been sampled randomly from the universe of admissible observations, and the intended universe of generalization is infinitely large, the concept of exchangeability may be invoked to consider the facet as random (Shavelson & Webb, 1981). A fixed facet arises in three conditions: (a) purposely selecting certain conditions and not interested in generalizing beyond them, (b) finding it unreasonable to generalize beyond the levels observed, or (c) when the entire universe of levels is small and all levels are included in the measurement design (Shavelson & Webb, 2006). G theory is essentially a random effects theory. Therefore, there should be at least one random facet in the data set.

According to Brennan (2001, p.108) the rules and equations of G theory assume that the objects of measurement are not nested within some other facet. However, G theory can treat such nested, or stratified, objects of measurement, but requires special consideration to do so. Objects of measurement are stratified with respect to some other variable, and an investigator may be interested in variability within levels of the stratification variable, as well as the variability across levels (Brennan, 2001, p.153). For instance, assume that there are 100 people in each of four regions (east, west, south, and north). Here, people are the object of measurement, and they are nested in regions. According to Brennan (2001) it is quite complex to cope such designs. Moreover, it is stated that if the design is unbalanced the procedures discussed do not apply, and appropriate procedures are much more complicated. Bloch and Norman (2018) defined this situation by using another term. According to them, when the facet of differentiation (object of measurement) is nested within another facet, this facet is referred to as a *facet of stratification*. For instance, there are students who are at different educational levels (senior vs. junior students). Commonly the difference between the two groups is viewed as a test of construct validity. However, in terms of reliability, the person variance should be computed within educational levels (we want to see if we can differentiate among individuals at the same level). Here, the educational level is a stratification facet (Bloch and Norman (2018). Although the term "stratification facet" in not in Brennan's (2001a) Generalizability Theory book, which can

be defined as the bible of the G theory, it is important to explain the stratification facet because G String V also uses this terminology.

There is also specific terminology associated with the design: crossed design and nested design. Assume all students (s) respond to all the items (i) in a test so that students are crossed with items. We denote this design as *s x i*. However, if each student responds to a different set of items then it is expressed as items nested within students. We denote this design as *i : s*. A crossed design is usually preferred in studies conducted using G Theory. The reason for this is that all sources of error, associated with all probable facets and the interactions between those facets, can be estimated in crossed-designed studies.

The importance of G theory lies in its applications to educational measurement. There are two major functions of G theory. One of them is to evaluate the quality of measurement procedures and the other is to make projections about how one can improve the quality of measurement procedures (Chiu, 2001). The former function can be done through the generalizability (G) study and it is possible to attain the second function via a decision (D) study. In other words, to evaluate the dependability of behavioral measurements, a G study is designed to isolate and estimate variation due to the object of measurement while examining as many facets of measurement error as possible. A D study uses the information provided by the G study to design the best possible application of the measurement for a particular purpose (Webb, Shavelson & Haertel, 2006) and answers the question "What if…?" by designing variations in measurement via optimization (Brennan, 2001). While planning the D study, the researcher defines a *universe of generalization*, the set of facets and their levels to which he or she wants to generalize, and specifies the proposed interpretation of the measurement. The decision maker uses the information from the G study to evaluate the effectiveness of alternative designs for minimum error and maximum reliability (Webb, Shavelson & Haertel, 2006).

Although there are wide applications of the theory, it has limitations in its capability of handling unbalanced designs of the data. The number of observations in balanced designs is equal at each level for the source of variability (Brennan, 2001a). By contrast, an unbalanced design has unequal numbers of observations in its sub-classifications. For instance, there can be differing numbers of items nested within testlets, pupils nested within differently sized classrooms, or observers nested within occasions with an unequal number of observers present at each occasion. These three examples are defined as unbalanced because the nested designs may be purposely unbalanced, dictated by the context itself or created by unforeseen circumstances, respectively. One other reason for unbalanced situations can be missing observations from crossed and nested designs (Webb, Shavelson & Haertel, 2006). As sample sizes tend to be small in the G theory analyses (Rios, Li & Faulkner-Bond, 2012; Taşdelen Teker & Güler, 2019), missing data becomes an important topic. Researchers normally prefer listwise deleting, inputing missing observations, or employing unbalanced designs to deal with missing data. Shavelson and Webb (1991) encouraged deleting data to create a balanced design to circumvent estimation challenges. Shavelson, Webb and Rowley (1989) found little effect in the estimated variance components when data was deleted to create balance. However, it can be problematic for very small sample sizes, such as less than 20. Rios, Li and Faulkner-Bond (2012) conducted a systematic review of the most recently published literature to understand the current methodological trends in the G theory better. Unbalanced design was used in 19 of 58 studies reviewed. Taşdelen Teker and Güler (2019) conducted a thematic content analysis of studies using the G theory in the field of education in Turkey and found that 6 of 60 studies were conducted with unbalanced design. According to the results of the above-mentioned review studies of Rios, Li and Faulkner-Bond (2012), and Taşdelen Teker and Güler (2019), the ratio of unbalanced designs is high enough to make coping with them indispensable.

Estimating variance components in unbalanced designs was challenging. Some or all methods had problems of computational complexity, distributional assumptions, and biased estimation, requiring decisions that could not be justified in the context of the G theory, or produced results that were inconclusive (Brennan, 2001). However, now it is possible to run G and D studies to estimate variance components and reliability coefficients by using the software G String V. The main purpose of this study was to introduce a computer program that was appropriate for G theory for balanced, and even more important, unbalanced data sets. According to the manual of G String V (Bloch & Norman, 2018), because the software is based on variance component estimates from urGENOVA, which was written by R. L. Brennan (2001a), the mathematical formulation declared by Brennan (2001a) will be given before introducing the software. Moreover, synthetic data taken from Brennan (2001a, p.224) will be provided to clarify the notations. Lastly, the software will be introduced step by step by using the same synthetic data shown in Table 1.

### 1.1. Synthetic Data and Mathematical Computations

Assume that eight students (s) take an 8-item test that is composed of three testlets (h) containing 2, 4, and 2 items (i), respectively. Because the number of items per testlet is not equal, the design is defined as unbalanced. It is a random facet nested design, symbolized as *sx(i:h)*. The data entry is shown in Table 1.

**Table 1.** *sx(i:h)* Unbalanced Design

| Student | Testlet 1 | | Testlet 2 | | | | Testlet 3 | |
|---|---|---|---|---|---|---|---|---|
| | Item 1 | Item 2 | Item 1 | Item 2 | Item 3 | Item 4 | Item 1 | Item 2 |
| 1 | 4 | 5 | 3 | 3 | 5 | 4 | 5 | 7 |
| 2 | 2 | 1 | 2 | 3 | 1 | 4 | 4 | 6 |
| 3 | 2 | 4 | 4 | 7 | 6 | 5 | 8 | 7 |
| 4 | 1 | 3 | 5 | 4 | 5 | 5 | 4 | 5 |
| 5 | 3 | 3 | 6 | 7 | 5 | 7 | 8 | 9 |
| 6 | 1 | 2 | 5 | 6 | 4 | 4 | 5 | 6 |
| 7 | 3 | 5 | 6 | 8 | 6 | 7 | 7 | 8 |
| 8 | 0 | 1 | 1 | 2 | 0 | 4 | 7 | 8 |

As seen in Table 1, there is no missing data. It is strongly advised to cope with the missing data by using standard statistical approaches before using G String V. However, if the researcher forgets to deal with the missing data, then G String V will replace the missing values by the grand mean and warn the user when this occurs.

The estimation of variance components in terms of mean squares are given in Table 2. The $n_{i+}$ notation, given under the degrees of freedom (*df*) column of Table 2, is the total number of levels of i over all levels of h; that is, $n_{i+} = \sum_h n_{i:h}$. Moreover, $r_i$ and $t_i$, which are used for the estimation of variance components in terms of mean squares, are computed by using the following equations: $r_i = \sum_h \frac{n_{i:h}^2}{n_{i+}}$ and $t_i = \frac{n_{i+}-r_i}{n_h-1}$.

**Table 2.** *sx(i:h)* Unbalanced Design

| Source of Variance | df | Mean Squares | Estimators of the variance components in terms of mean squares |
|---|---|---|---|
| *s* | $n_s$-1 | $MS_s$ | $\sigma^2_s = [MS_{(s)} - r_i MS_{(sh)}/t_i + (r_i - t_i)MS_{(si:h)}/t_i]/n_{i+}$ |
| *h* | $n_h$-1 | $MS_h$ | $\sigma^2_h = [MS_{(h)} - MS_{(i:h)} - MS_{(sh)} + MS_{(si:h)}]/n_s t_i$ |
| *i:h* | $n_{i+} - n_h$ | $MS_{i:h}$ | $\sigma^2_{i:h} = [MS_{(i:h)} - MS_{(si:h)}]/n_s$ |
| *sh* | $(n_s-1)(n_h-1)$ | $MS_{sh}$ | $\sigma^2_{sh} = [MS_{(sh)} - MS_{(si:h)}]/t_i$ |
| *si:h* | $(n_s-1)(n_{i+}-n_h)$ | $MS_{si:h,e}$ | $\sigma^2_{si:h,e} = MS_{(si:h,e)}$ |

By using the variance components obtained from the G study, it is possible to estimate relative and absolute error variances. After that, the error variances are used to estimate the generalizability and dependability coefficients. The equations used for the estimations of relative/absolute error variances and generalizability/dependability coefficients are given below. Equation 1 is used to estimate the relative error variance ($\sigma^2(\delta)$) and Equations 2 and 3 are used to compute generalizability coefficients ($E\rho^2$) for unbalanced sx(i:h) design. Equation 4 is used for the estimation of absolute error variance. Then Equations 2 and 5 are used for the estimation of index of dependability ($\phi$). The $\breve{n}_h$ term used for the estimation of relative and absolute error variances is equal to $\breve{n}_h = \frac{n_{i+}^2}{\sum_h n_{i:h}^2}$.

$$\sigma^2(\delta) = \frac{\sigma_{sh}^2}{\breve{n}_h} + \frac{\sigma_{si:h}^2}{n_{i+}} \quad [1]$$

$$\sigma^2(\tau) = \sigma^2(s) \quad [2]$$

$$E\rho^2 = \frac{\sigma^2(\tau)}{\sigma^2(\tau) + \sigma^2(\delta)} \quad [3]$$

$$\sigma^2(\Delta) = \frac{\sigma_h^2}{\breve{n}_h} + \frac{\sigma_{i:h}^2}{n_{i+}} + \frac{\sigma_{sh}^2}{\breve{n}_h} + \frac{\sigma_{si:h}^2}{n_{i+}} \quad [4]$$

$$\phi = \frac{\sigma^2(\tau)}{\sigma^2(\tau) + \sigma^2(\Delta)} \quad [5]$$

## 2. THE SOFTWARE: G STRING V

G String V (Bloch & Norman, 2018) is a software that functions on the basis of urGENOVA (Brennan, 2001b) and is used in G theory analyses. Because urGENOVA is a traditional command line program that does not have a graphical user interface, users must specify their parameters, which makes it difficult to work with. Moreover, although urGENOVA provides the variance components for the individual effects, it does not calculate variance coefficients under different conditions. However, G String V does this as well (Bloch & Norman, 2018). G String V was designed and coded by Ralph Bloch as part of a project commissioned by The Medical Council of Canada and was subsequently further developed. It is written in Java on the Linux platform. The most recent version of the program runs under the Windows operating system (Bloch & Norman, 2015) and Macintosh and Linux operating systems (Bloch & Norman, 2018). The G-String V has a more user-friendly interface and therefore, is much easier to use compared to urGENOVA.

### 2.1. Installing the Software

G String V can be downloaded for free from the Web. Researchers may install the latest version of the G String V software, released in July 2018 from https://healthsci.mcmaster.ca/merit/research/g_string_v. The program is contained in a software package called "G_String_V.jar".

Before downloading the software, install Java Runtime JRE 8 on your computer if it is not already installed. Then create a new folder called "G_String_V" in a suitable location of your file system. After selecting your computer's operating system (Windows, Mac-OS or Linux), download the software package and copy it from the Downloads folder into the G_String_V folder. Next, create a new sub-folder within the G_String_V folder called "work." Then double click on G_String_V. jar. As shown in Figure 1, set the "work" sub-folder as your working directory by clicking the Setup and Set Working Directory buttons.

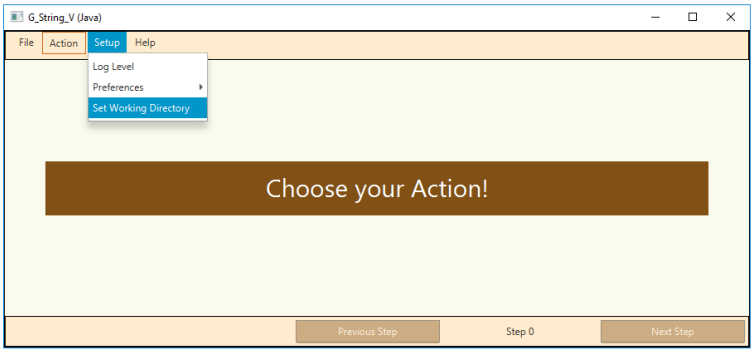



**Figure 1.** G String V Software Working Directory Setup Screen

## 2.2. Software Interface

The main interface of the G String V software is shown in Figure 2. This interface consists of the main menu that includes File, Action, Setup, and Help. To start the analysis, click on Action. There are three sub-menus under Action as seen in Figure 2. To start a new analysis click on Action→ Start New and create a new G String V run. If you want to do multiple runs on a pre-used data base then click on Use Existing. For the G String V to automatically count the number of levels of each facet, which can be helpful for unbalanced nested designs, select Auto Index.

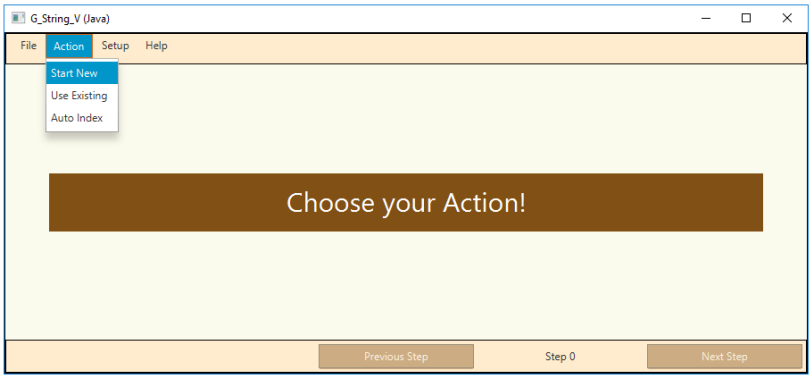



**Figure 2.** G String V Software Main Screen

## 2.3. Data Analysis using G String V

If starting with a new data set click on Start New as seen in Figure 2. From this point on, analyzing with the G String V will be explained step by step. As seen in Table 1, Brennan's example (2001, p. 224) will be used to better illustrate the steps. There are three testlets (h) containing 2, 4, and 2 items (i), respectively; and answers to all of the questions for the eight students (s). Because the items are nested within testlets and students answer all of the items, the design can be symbolized as *sx(i:h)*. Moreover, because the number of items per testlet is not the same, the design is unbalanced, as previously stated.

### 2.3.1. *Preparing Data for Analysis*

While urGENOVA requires the data to be in ASCII text files (.dat or .txt), G String V is set up to handle tab-delimited or fixed format text files. ASCII files can be easily generated from a spreadsheet, such as Excel, by simply saving as a "Text - tab delimited (*.txt)" file (Bloch & Norman, 2018). Like all previous versions of G String V and other software used for G theory analysis (SPSS and EduG), G String V requires that the data be ordered, so that all records related to a particular level of a facet are together. After entering the data in an Excel spreadsheet, and saving as Text - tab delimited (*.txt) it can be used to run G theory analysis via G String V.

### *Steps 1 and 2: Title and details of the conducted research*

As seen in Figure 3, you can provide a unique name for your research in Step 1 and add more comments to describe the details of the analysis in Step 2. Comments provide an explanation of the study to the reader of the output. The information written here does not affect the computations of the software, so if you do not want to enter any information omit this step by clicking the Next Step button. To exemplify how it would appear in the output file, a title and brief description of the data was entered.

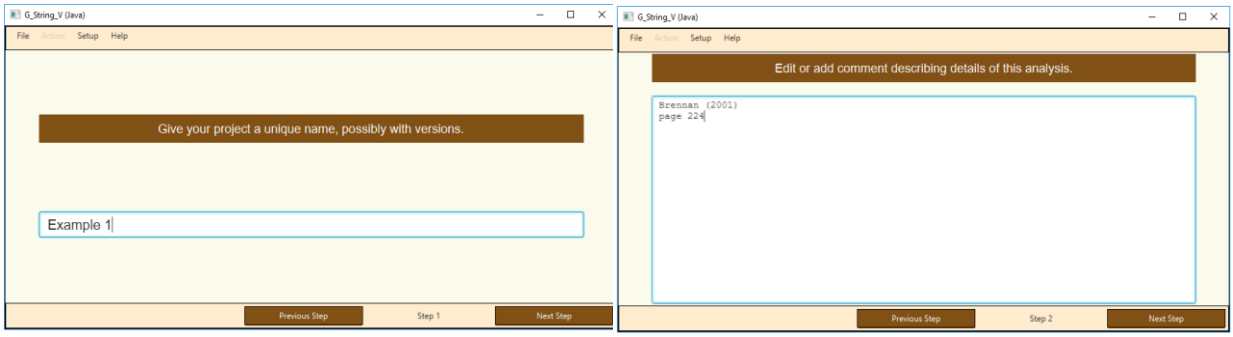



**Figure 3.** G String V Software: (a) Step 1 and (b) Step 2

### *Steps 3 and 4: Defining object of measurement and facets*

As seen in Figure 4(a), the object of measurement is specified in Step 3 by providing a descriptive name and a corresponding one-character lowercase abbreviation for the object of measurement. Although the object of measurement is usually crossed with other facets, it may also be nested within another facet. For instance, as given in Brennan's example (2001, p.154), people can be nested within regions. Because of this, the nesting situation of the object of measurement also should be specified. For the example discussed here, "crossed" should be selected. Then specify the number of facets. For the example used here, there are two, testlet and item. Each facet should be given a descriptive name and a one-character abbreviation in Step 4, as seen in Figure 4(b). Moreover, the nested facets are specified by changing the default "crossed" to "nested." In our example, because the items are nested within testlets, the nesting conditions of items have "nested" selected. The nesting condition of testlets remains "crossed" by default.

The order of the facets is also important. They must be listed in the order they are encountered in the data file, from slowest-moving to fastest-moving (Bloch & Norman, 2018). In other words, the first facet to be declared is the one whose levels change least rapidly and the last facet to be declared would be the one whose levels change the fastest (Cardinet, Johnson & Pini, 2010). In our example, the first facet is testlet as it changes more slowly than items. More clearly, if the data have one record per student, with all data for each testlet, then the responses on each item of testlet, the order of facets would be: Testlet, Item.

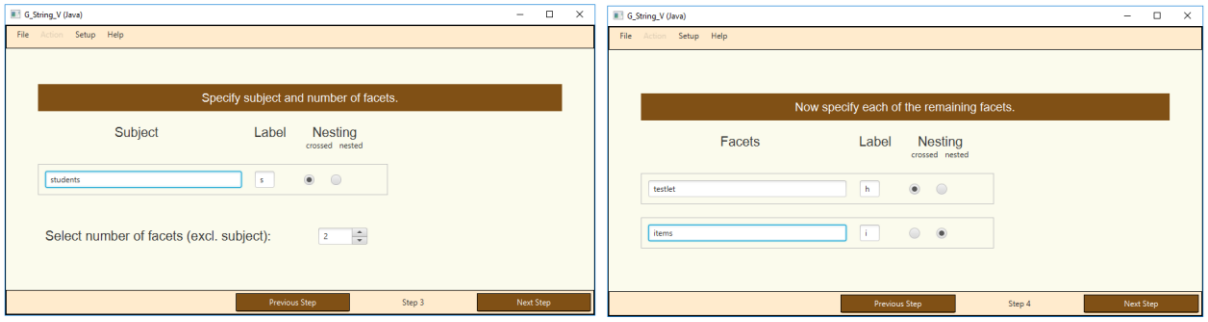



**Figure 4.** G String V Software: (a) Step 3 and (b) Step 4

### *Steps 5 and 6: Setting order of facets and arrange nesting of the facets*

In Step 5, the order of facets should be specified as they appear in the data set. For instance, because each student's answers to eight items are listed in one row, the asterisk is beside student as seen in Figure 5(a). In Step 6, drag and drop the nested facets from the left side to the right side. By doing so, the nested facets are located under the facet in which they are nested nested as seen in Figure 5(b). Because items are nested within testlets in our example, the item facet is dragged and dropped under the testlet facet and appears as i:h.

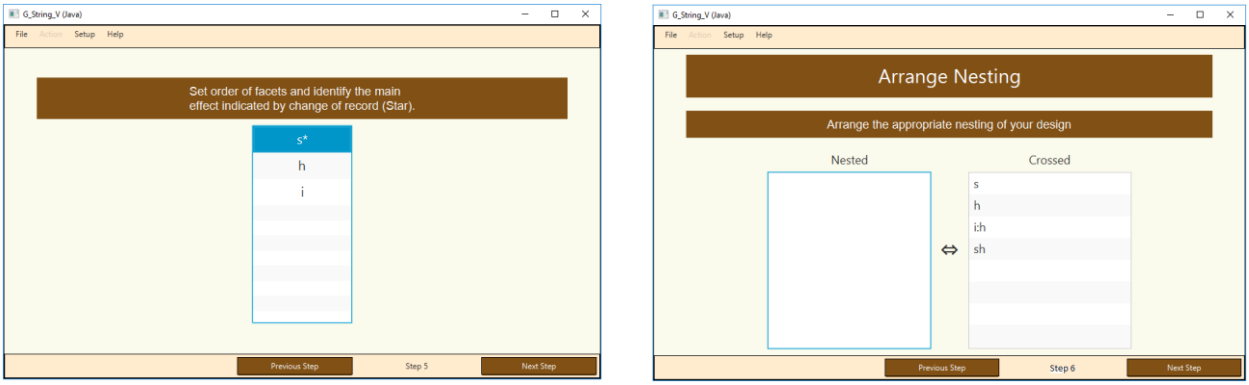



**Figure 5.** G String V Software: (a) Step 5 and (b) Step 6

### *Step 7: Locating data file*

As seen in Figure 6(a), the exact location of the data file is selected. The data must be in an ASCII text file. To do so, you can enter your data in Excel and save it as "Text (Tab delimited) (*.txt)." After selecting the location of the data, you will see it on the screen. As seen in Figure 6(b), there are nine columns in the data file. The first column contains the student ID, which means the actual data begins in the second column. To indicate how many columns are to be skipped, enter the information in the "Skip" field. As you can see from the second screenshot of Figure 6, "1" is in the "Skip" area, so the first column of the data becomes colorless, which means it will not be analyzed. You can only skip fields at the beginning of the data, never in the middle.

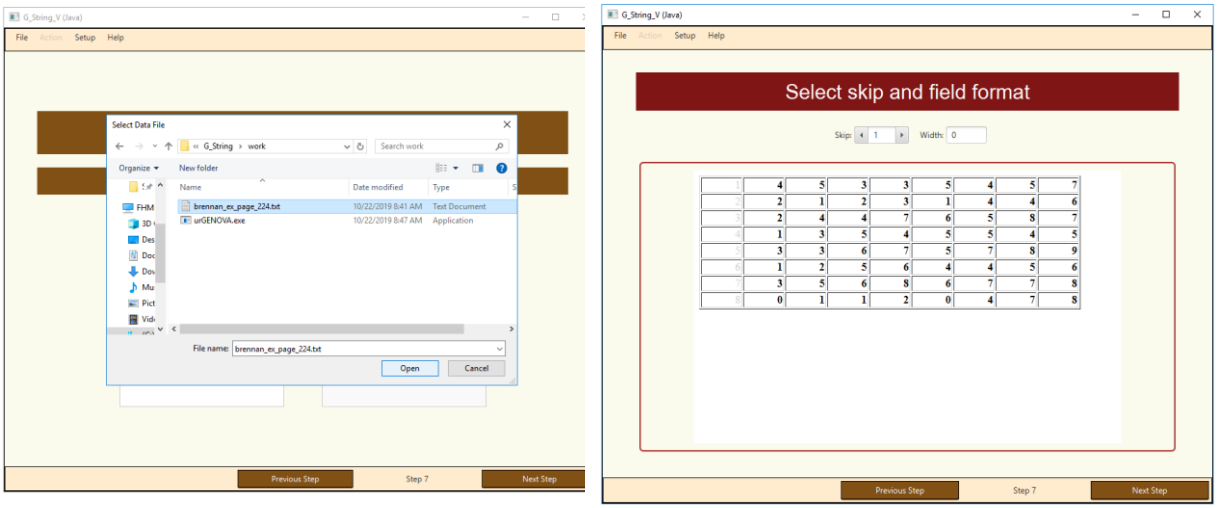



**Figure 6.** G String V Software Step 7: (a) Data location and (b) Data view

### *Step 8: Specifying the sample sizes of object of measurement and facets*

In Step 8, G String V asks for the sample size of the object of measurement and then the facets' sample sizes. For nested variables, specify the number of levels at each level of the nesting variable. For the object of measurement, this will be 8; for the testlet facet, this will be 3; and for the item facet, this will be the number of items per testlet; 2, 4, and 2, as seen in Figure 7.

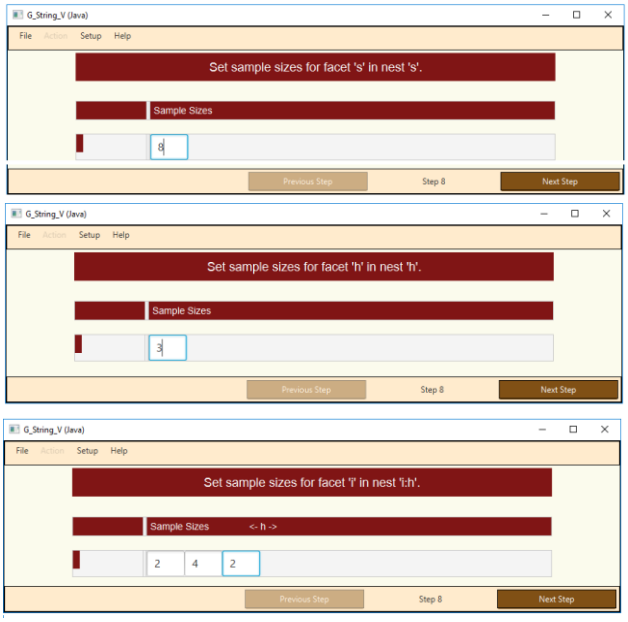



**Figure 7.** G String V Software Step 8

### *Step 9: Saving the Control File and obtaining variance components*

After completing the specification by running the previous eight steps, a control file called "gControl.txt" was also generated. It was stored in the working directory by default. In Step 9, you can change both its name and folder as seen in Figure 8(a). After saving the proper control file path, urGENOVA is executed automatically to calculate the variance components and the coefficients (Bloch & Norman, 2018) as seen in Figure 8(b). Step 9 shows the variance components as part of a G study of G theory.

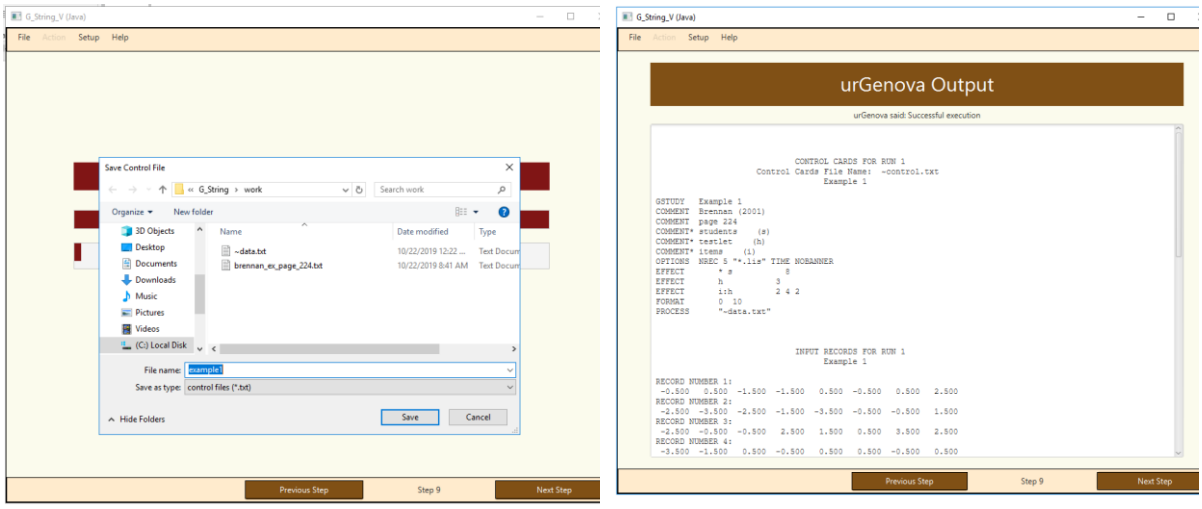



**Figure 8.** G String V Software Step 9: (a) Saving control card and (b) urGENOVA output

### *Step 10: Calculating G Coefficients and Running D Studies*

In Step 10, G String V calculates the G coefficients as seen in Figure 9(a). The first coefficient is called the G coefficient and is symbolized as $E\rho^2$, which is used for relative decisions. The second coefficient is called the index of dependability (Brennan & Kane, 1977) and symbolized as $\phi$, which is used for absolute decisions. Furthermore, by changing levels and types of facets, you can calculate different coefficients to answer the question "What if…?" as part of a D study of the G theory. As seen in Figure 9(b), the level of testlet changed from 3 to 4, but the type of facets was left as random. After this change, the $E\rho^2$ and $\phi$ coefficients were also changed from .73 to .79 and .45 to .53, respectively. After completing all the intended D studies, by changing the levels of facets and clicking Next Step to obtain the results, close the software by clicking File→Close from the menu bar.

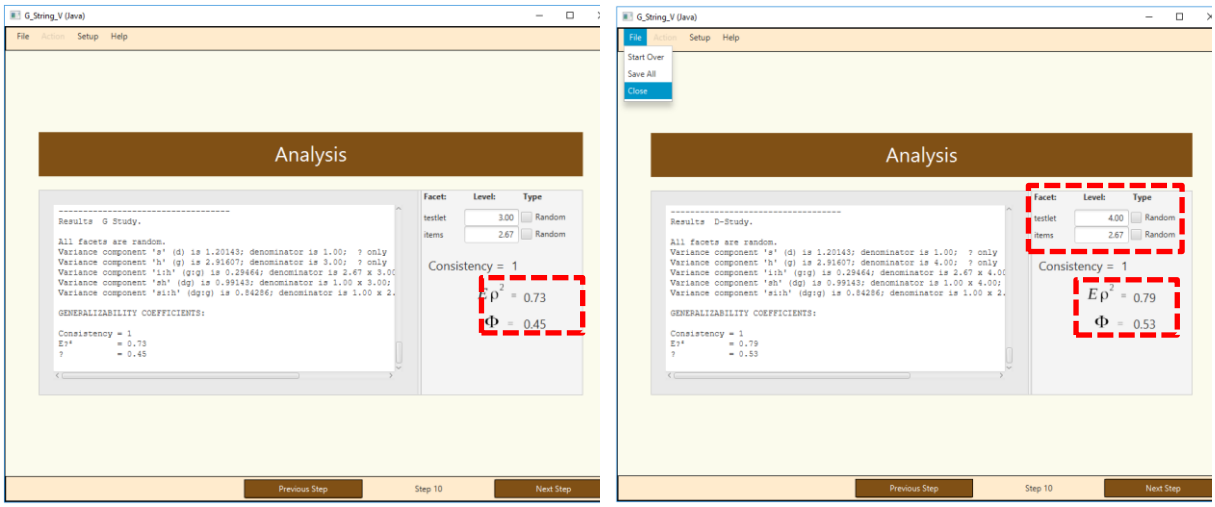
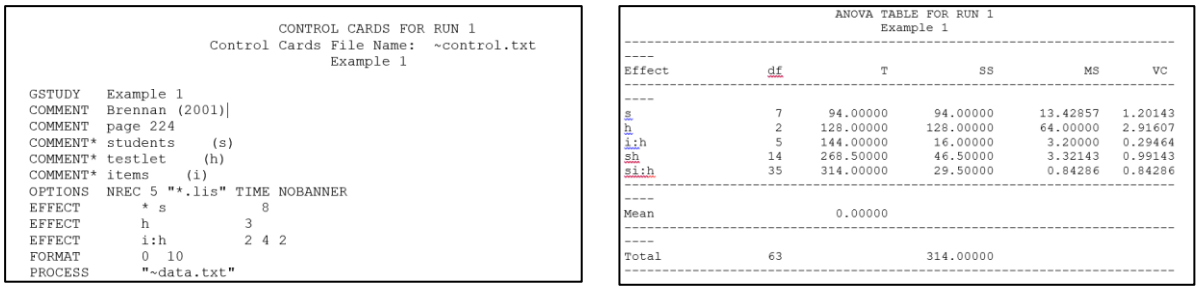



**Figure 9.** G String V Software Step 10: (a) G study results and (b) D study results

## 2.4. Evaluation of the Results

After closing the software, the output file will be in your working directory, named as "example1.txt.lis". This file can be opened by Word. As seen in Figure 10(a), there is a control card at the beginning of the output file. It contains the information entered in Step 1 (Example 1) and Step 2 (Brennan, 2001a, p.224). The names and levels of facets and the design of the study are also on the control card. Figure 10(b) shows the ANOVA table created by urGENOVA. The variance components used in the calculations of the $E\rho^2$ and $\phi$ coefficients are shown to the right of the table. It is possible to estimate variance components as negative because of erroneous measurement models or sampling errors (Güler, Kaya Uyanık & Taşdelen Teker, 2012). There are two different approaches to handle this situation. Cronbach et al., (1972) initially said that the negative variance should be replaced with zero, and that zero should be used to calculate other variance components. Brennan (2001a), however, argued that this suggestion could cause biased calculations of variance components. Cronbach responded by saying that although the negative variance should be replaced by zero, the negative value itself should be used to calculate other variance components (Atılgan, 2004). Negative variances are set to zero when computing coefficients by using G String V.

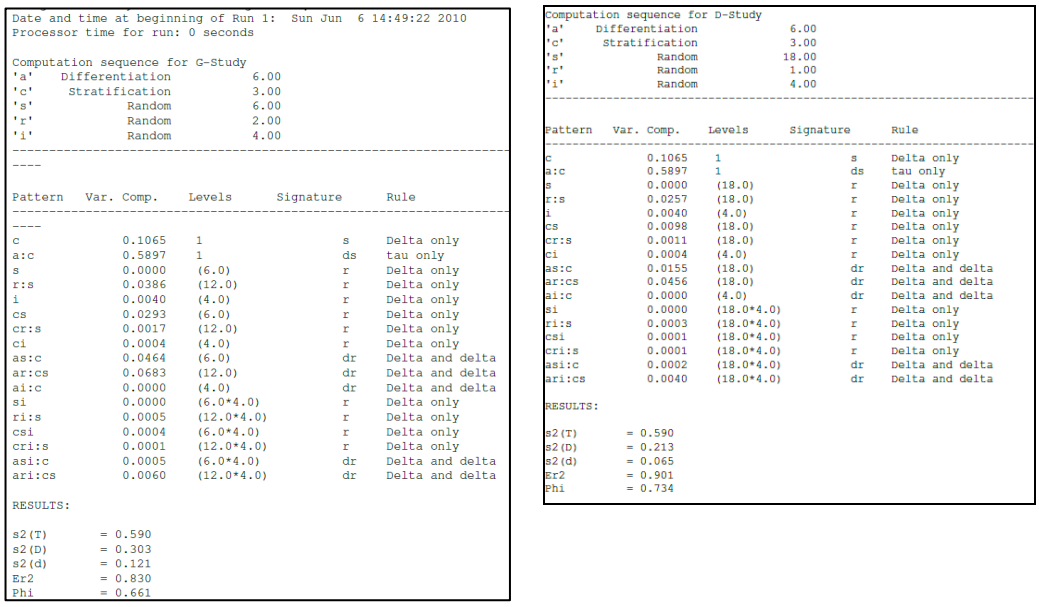


```
                    CONTROL CARDS FOR RUN 1
              Control Cards File Name:  ~control.txt
                        Example 1

GSTUDY    Example 1
COMMENT   Brennan (2001)|
COMMENT   page 224
COMMENT*  students     (s)
COMMENT*  testlet      (h)
COMMENT*  items      (i)
OPTIONS   NREC 5 "*.lis" TIME NOBANNER
EFFECT        * s            8
EFFECT        h              3
EFFECT        i:h            2 4 2
FORMAT        0  10
PROCESS       "~data.txt"
```

```
                    ANOVA TABLE FOR RUN 1
                          Example 1

----------------------------------------------------------------------------
----
Effect              df           T            SS           MS          VC
----------------------------------------------------------------------------
----
s                    7     94.00000     94.00000     13.42857     1.20143
h                    2    128.00000    128.00000     64.00000     2.91607
i:h                  5    144.00000     16.00000      3.20000     0.29464
sh                  14    268.50000     46.50000      3.32143     0.99143
si:h                35    314.00000     29.50000      0.84286     0.84286
----------------------------------------------------------------------------
----
Mean                      0.00000
----------------------------------------------------------------------------
----
Total               63                314.00000
----------------------------------------------------------------------------
```

**Figure 10.** Output file: (a) Control card and (b) ANOVA Table

Calculations of $E\rho^2$and ϕ coefficients are shown in the output below the ANOVA table. In Figure 11(a), the estimated $E\rho^2$and ϕ coefficients are seen at the bottom of the figure and the results of the D studies are shown in Figure 11(b). When the level of testlet changed from 3 to 4, there is an increase in $E\rho^2$ and ϕ coefficients. More clearly, if the researcher increases the number of testlets, for instance to cover content area better, the results will be more reliable. Moreover, only the level of facets was changed and there was no change on the type of the facet from random to fixed. The reason of remaining the testlet facet as random was that since the entire universe of testlet levels was quite large and all levels were impossible to be included, the researcher was interested in generalizing beyond 3 or 4 testlets.



```
Date and time at beginning of Run 1:  Sun Jun  6 14:49:22 2010
Processor time for run: 0 seconds

Computation sequence for G-Study
'a'    Differentiation          6.00
'c'     Stratification          3.00
's'             Random          6.00
'r'             Random          2.00
'i'             Random          4.00
-------------------------------------------------------------------------
----

Pattern   Var. Comp.    Levels      Signature      Rule
-------------------------------------------------------------------------
----
c            0.1065    1                  s     Delta only
a:c          0.5897    1                  ds    tau only
s            0.0000    (6.0)              r     Delta only
r:s          0.0386    (12.0)             r     Delta only
i            0.0040    (4.0)              r     Delta only
cs           0.0293    (6.0)              r     Delta only
cr:s         0.0017    (12.0)             r     Delta only
ci           0.0004    (4.0)              r     Delta only
as:c         0.0464    (6.0)              dr    Delta and delta
ar:cs        0.0683    (12.0)             dr    Delta and delta
ai:c         0.0000    (4.0)              dr    Delta and delta
si           0.0000    (6.0*4.0)          r     Delta only
ri:s         0.0005    (12.0*4.0)         r     Delta only
csi          0.0004    (6.0*4.0)          r     Delta only
cri:s        0.0001    (12.0*4.0)         r     Delta only
asi:c        0.0005    (6.0*4.0)          dr    Delta and delta
ari:cs       0.0060    (12.0*4.0)         dr    Delta and delta

RESULTS:

s2(T)       = 0.590
s2(D)       = 0.303
s2(d)       = 0.121
Er2         = 0.830
Phi         = 0.661
```

```
Computation sequence for D-Study
'a'    Differentiation          6.00
'c'     Stratification          3.00
's'             Random         18.00
'r'             Random          1.00
'i'             Random          4.00

-------------------------------------------------------------------------

Pattern   Var. Comp.    Levels      Signature      Rule
-------------------------------------------------------------------------
c            0.1065    1                  s     Delta only
a:c          0.5897    1                  ds    tau only
s            0.0000    (18.0)             r     Delta only
r:s          0.0257    (18.0)             r     Delta only
i            0.0040    (4.0)              r     Delta only
cs           0.0098    (18.0)             r     Delta only
cr:s         0.0011    (18.0)             r     Delta only
ci           0.0004    (4.0)              r     Delta only
as:c         0.0155    (18.0)             dr    Delta and delta
ar:cs        0.0456    (18.0)             dr    Delta and delta
ai:c         0.0000    (4.0)              dr    Delta and delta
si           0.0000    (18.0*4.0)         r     Delta only
ri:s         0.0003    (18.0*4.0)         r     Delta only
csi          0.0001    (18.0*4.0)         r     Delta only
cri:s        0.0001    (18.0*4.0)         r     Delta only
asi:c        0.0002    (18.0*4.0)         dr    Delta and delta
ari:cs       0.0040    (18.0*4.0)         dr    Delta and delta

RESULTS:

s2(T)       = 0.590
s2(D)       = 0.213
s2(d)       = 0.065
Er2         = 0.901
Phi         = 0.734
```

**Figure 11.** Coefficient estimation based on (a) G study and (b) D Studies

## 3. LIMITATIONS OF G STRING V

The first limitation of G String V is related to the sample size of the data. The maximum number of the facet of differentiation, which is the object of measurement, is 1500. If your sample size is above this limit, it is stated on the G String V manual that you can write to the developers of the software and they can furnish a modified version of it. The other limitation of the software is related to the number of stratification facets of the study. For practical reasons, it cannot handle more than four stratification facets. According to the results of two review studies conducted by Rios, Li and Faulkner-Bond (2012) and Taşdelen Teker and Güler (2019), there is no study that has more than one stratification facet. Meanwhile, when a researcher has more than four stratification facets it has been suggested to collapse the facets that are unlikely to contribute to error variance.

**Acknowledgements**

The author wish to thank the Hacettepe Üniversitesi Teknokent Teknoloji Transfer Merkezi for proofreading process.

**ORCID**

Gülşen TAŞDELEN TEKER https://orcid.org/0000-0003-3434-4373

## 4. REFERENCES

Atılgan, H. (2004). *Genellenebilirlik kuramı ve cok değişkenlik kaynaklı Rasch modelinin karşılaştırılmasına ilişkin bir araştırma [A research on the comparison of the generalizability theory and many facet Rasch model]* (Doctoral Dissertation). Hacettepe University, Ankara.

Bloch, R. & Norman, G. (2018). *G String V User Manual*. Hamilton, Ontario, Canada.

Bloch, R. & Norman, G. (2015). *G String IV* (Version 6.1.1) User Manual. Hamilton, Ontario, Canada.

Bloch, R. & Norman, G. (2012). Generalizability theory for the perplexed: A practical introduction and guide: AMEE Guide No. 68. *Medical Teacher, 34* (11), 960-992. DOI: 10.3109/0142159X.2012.703791

Brennan, R. L. (2001a). *Generalizability Theory.* New York: Springer.

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

Brennan, R. L. (2000). Performance Assessments from the Perspective of Generalizability Theory. *Applied Psychological Measurement, 24*(4), 339-353.

Brennan. R. L., & Kane, M. T. (1977). An index of dependability for mastery tests. *Journal of Educational Measurement, 14*, 277-289.

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

Cardinet, J., Tourneur, Y. & Allal, L. (1981). Extension of Generalizability Theory and Its Applications in Educational Measurement. *Journal of Educational Measurement, 18* (4), 183-204.

Cardinet, J., Tourneur, Y. & Allal, L. (1976). The Symmetry of Generalizability Theory: Applications to Educational Measurement. *Journal of Educational Measurement, 13* (2), 119-135.

Chiu, C. W. T. (2001). *Scoring performance assessments based on judgments: Generalizability theory.* Boston, MA: Kluwer Academic.

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: Wiley.

Furr, R. M. (2011). *Scale construction and psychometrics for social and personality psychology.* Thousand Oaks, CA: Sage Publications Ltd.

Güler, N., Kaya Uyanık, G. & Taşdelen Teker, G. (2012). *Genellenebilirlik Kuramı [Generalizability Theory].* Ankara: PegemA Yayıncılık.

Rios, J.A., Li, X., & Faulkner-Bond, M. (2012, October). *A review of methodological trends in generalizability theory.* Paper presented at the annual conference of the Northeastern Educational Research Association, Rocky Hill, CT.

Shavelson, J. R. & Webb, N. M. (2006). Generalizability theory. In: Green, J.L., Camill, G., Elmore, P.B., editors. *Handbook of complementary methods in education research*. Mahwah: Lawrence Erlbaum Associates Publishers, p. 309–322.

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

Shavelson, R.J., Webb, N.M., & Rowley, G.L. (1989). Generalizability theory. *American Psychologist*, *44*(6), 922-932.

Shavelson, R. J., & Webb, N. M. (1981). Generalizability theory: 1973–1980. *British Journal of Mathematical and Statistical Psychology*, *34*, 133–166.

Suen, H. K. & Lei, P.W. (2007). Classical Versus Generalizability Theory of Measurement. *Educational Measurement*, 4, 1-13.

Taşdelen Teker, G. & Güler, N. (2019). Thematic Content Analysis of Studies Using Generalizability Theory. *International Journal of Assessment Tools in Education*, 6(2), 279–299. https://dx.doi.org/10.21449/ijate.569996

Webb, N. M., Shavelson, R. J. & Haertel, E. H. (2006). Reliability Coefficients and Generalizability Theory. *Handbook of Statistics*, *26*, 81-124. DOI: 10.1016/S0169-7161(06)26004