International Journal of Assessment Tools in Education
2019, Vol. 6, No. 2, 279–299
https://dx.doi.org/10.21449/ijate.569996

Published at http://www.ijate.net http://dergipark.org.tr/ijate **Research Article**

Thematic Content Analysis of Studies Using Generalizability Theory

Gülşen Taşdelen Teker[1*], **Neşe Güler**[2]

[1] Hacettepe University, Faculty of Medicine, Department of Medical Education and Informatics, Ankara, Turkey
[2] İzmir Demokrasi University, Faculty of Education, Measurement and Evaluation Department, İzmir, Turkey

ARTICLE HISTORY

Received: 11 March 2019
Revised: 14 May 2019
Accepted: 22 May 2019

KEYWORDS

Educational sciences,
Generalizability theory,
Thematic content analysis

Abstract: One of the important theories in education and psychology is Generalizability (G) Theory and various properties distinguish it from the other measurement theories. To better understand methodological trends of G theory, a thematic content analysis was conducted. This study analyzes the studies using generalizability theory in the field of education in Turkey by using the method of thematic content analysis. It reviews 60 studies, including 31 articles and 29 theses published from 2004 to 2017. The selected studies underwent thematic content analysis using parameters including tagged information, aim, G Theory type, number of facets used in the study, Turkish word for "facet," object of measurement, sample size, design type, mixed-design availability, shared results of G and D studies, computer programs, method of calculating negative variance, availability of fixed facets, and design balance. The data were interpreted on the basis of frequencies; both table and figures are included in the study. According to the results, there is an increase in the number of studies conducted by using G theory by years. Of these, many compare theories; most of them applying univariate G Theory and consider two-faceted measurement situations. While a small subset of studies features mixed design, a large group features crossed design, with individuals as the object of measurement. The computer program most commonly used in analyses is EduG. The majority of studies use balanced design. Recommendations are provided accordingly with the results.

1. INTRODUCTION

One of the most important steps taken in any scientific study is the measurement process used to obtain information needed to analyze a particular object or property. However, the data obtained in this process may contain various types of "error." These errors, which differ in accordance with the measurement conditions, have a different meaning from the one that is traditionally assumed. Errors occur naturally in measurement; it is therefore essential to determine how and under what conditions to carry out "ideal" acts of measurement, given this reality. In education and psychology, this issue is discussed as an aspect of "reliability," which

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

may be defined as the extent to which the observed scores are consistent (or inconsistent) (Brennan, 2011).

One of the theories concerning reliability in education and psychology is the Generalizability (G) Theory. This theory enables a researcher to determine the source and number of inconsistencies in the observed scores. Another theory, Classical Test Theory (CTT), consists of observed scores (X), real scores (T), and error scores (E) (X= T + E). Although only one error term appears in this model for CTT, the term contains all probable errors. In this context, one of the most important advantages of G Theory is that it enables the investigation of different sources of error within the model it is based on. For instance, the relation between G Theory and the process of measurement where there are K number of sources of error can be described as follows:

$$X = \mu_s + E_1 + E_2 + \dots + E_K \qquad (1)$$

Here, μ_s in Equation 1 is the universe score, interpreted in a similar way to the real score in CTT. The universe score is defined as the expected value of the observed scores obtained through repetitive measurements (Brennan, 2001). One of the properties that makes G Theory important and different is its conceptual framework. The concepts in this framework are the *universe of admissible observations, Generalizability (G) study*, the *universe of generalization*, and *decision (D) study*. The present study uses a sample situation to ensure that these G Theory concepts are understood better. For example, consider a measurement process in which the mathematical problem-solving skills of students are measured in different tasks (t) and scored by more than one rater (r). This process contains two *facets*, labelled "tasks" and "raters." Facets represent similar situations in measurement. Let us assume that tasks (one facet in this measurement process) contain an infinite number of tasks, while raters (another facet in this measurement process) contain an infinite number of raters. Both facets have been selected from an infinite universe of admissible observations. If each rater scores each task carried out by every student in the sample process, the measurement design is called a crossed design, and the process is represented as *sxtxr*. If, however, each task carried out by all students in the process is scored by different raters, the raters are said to be "nested" in the tasks and the study design is known as a "nested design," represented as *sx(r:t)*. 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-design studies. This situation gives D studies great flexibility.

A careful analysis of the example above makes clear that students also participate in the process, alongside tasks and raters; they too are considered variance sources of the measurement process. Any individuals, students, objects, or situations constituting the subject matter being measured are called the *object of measurement* in G Theory. While the term *universe* is used to denote the facets of measurement in G Theory, *population* is preferred for the object of measurement (Brennan, 1992). Observable scores, obtained by evaluating a task in the population or universe of admissible observations by a rater, are represented in Equation 2:

$$X_{str} = \mu + \upsilon_s + \upsilon_t + \upsilon_r + \upsilon_{st} + \upsilon_{sr} + \upsilon_{tr} + \upsilon_{str} \qquad (2)$$

In Equation 2, μ represents the average within the universe and population, while υ represents each of the seven unrelated components. This is a linear model of *sxtxr* (Brennan, 2011; Güler, Uyanık, & Teker, 2012). A model of this design contains seven sources of variance, known as, "G study variance components." Once these variance components have been estimated, the values can be used in estimates of universe score variance, error variance, various generalizability universe coefficients with similar interpretations, and various D-study designs. Variance components in a G study can be estimated using the expected values of squares average in the variance analysis. As is clear from here, a variance analysis (ANOVA) appears in the statistical structure of Equation 2. However, the F test is not used in G Theory. This case reflects one of the operational differences that distinguish G Theory from traditional variance analysis (Brennan, 1992; Güler et al., 2012).

The variance components obtained through the G study are used to design various D studies. The above-mentioned example can help to explain this situation. Let us assume that there is a process of measurement in which students' mathematical problem-solving skills are evaluated by three raters (r) using five different tasks (t). Each level of the two facets (tasks and raters) is called a *condition*. In this study, there are five conditions for the facet of tasks and three conditions for the facet of raters. Firstly, the variance components are calculated using the data obtained through the G study. After that, various D studies can be set up to decide on designs containing the same or different numbers of conditions of the facets made available by the G study. For instance, in D studies organized on the basis of a G study with five available tasks, designs can be created in which the same number (5), a smaller number (1, 2, 3 or 4), or a larger number (6, 7 etc.) of tasks is available; such designs can also include the same or a smaller or larger number of raters. One point to take careful note of here is that the variance components obtained through the G study are values estimated using a single task and rater (one condition). Thus, estimates made for various numbers of facet and task conditions also constitute D studies.

The universe score variance for a randomly crossed D study with the same structure as the G study in the example above is as follows:

$$\sigma^2\,(\tau) = \sigma^2\,(s) \tag{3}$$

The relative error variance is:

$$\sigma^2(\delta) = \frac{\sigma^2(s\ \ )}{n'_t} + \frac{\sigma^2(s\ \ )}{n'_r} + \frac{\sigma^2(s\ \ \ )}{n'_t n'_r} \tag{4}$$

The absolute error variance is:

$$\sigma^2(\Delta) = \frac{\sigma^2(t)}{n'_t} + \frac{\sigma^2(r)}{n'_r} + \frac{\sigma^2(t\ \ )}{n'_t n'_r} + \frac{\sigma^2(s\ \ )}{n'_t} + \frac{\sigma^2(s\ \ )}{n'_r} + \frac{\sigma^2(s\ \ \ )}{n'_t n'_r} \tag{5}$$

The generalizability coefficient is:

$$E(\rho^2) = \frac{\sigma^2(\tau)}{\sigma^2(\tau) + \sigma^2(\delta)} \tag{6}$$

The dependability coefficient is:

$$\Phi = \frac{\sigma^2(\tau)}{\sigma^2(\tau) + \sigma^2(\Delta)} \tag{7}$$

As equations 3, 4, 5, 6 and 7 make it clear, G Theory is based on average score metrics (Brennan, 2011), unlike CTT, which is based on total score metrics. In addition, the relative error variance and generalizability coefficient are interpreted in a similar way to the error variance and dependability coefficient, which are based on a relative comparison of individuals in CTT. Another point worth noting is that the generalizability coefficient and the dependability coefficient are not equal in G Theory. The error variance and dependability coefficient, based on absolute decisions that can be calculated in G Theory, cannot be calculated in CTT. However, both the absolute error variance and the error variance in CTT are derived from the random sampling assumption; when more than one facet is taken from the generalizability universe, the CTT error variance can be estimated at very low levels (for further details, see Brennan, 1997).

The above equations can also be applied to different generalizability universes and D studies. For instance, let us suppose that the raters assessing student mathematical problem-solving skills are constant. In other words, let us suppose that the purpose is not to generalize the raters in this study into a larger universe (alternatively, assume that all of the raters in the universe are the raters in this study). In this case, the universe score variance is:

$$\sigma^2(\tau) = \sigma^2(s) + \qquad \frac{\sigma^2(s\ \ )}{n'_r} \tag{8}$$

The relative error variance is:

$$\sigma^2(\delta) = \frac{\sigma^2(s\ )}{n'_t} + \frac{\sigma^2(s\ \ \ )}{n'_t n'_r} \tag{9}$$

And the absolute error variance is:

$$\sigma^2(\Delta) = \frac{\sigma^2(t)}{n'_t} + \frac{\sigma^2(t\ \ )}{n'_t n'_r} + \frac{\sigma^2(s\ )}{n'_t} + \frac{\sigma^2(s\ \ \ )}{n'_t n'_r} \tag{10}$$

It is clear that, when the rater facet is constant, estimated error variances decrease and universe score variance increases. This means that dependability coefficients will be estimated at high levels. However, the gain obtained by the increase in dependability values restricts the interpretations that can be made in relation to generalizability. In a similar way, as the sample size increases in D studies (that is to say, as the number of conditions increases) and/or when the study has a nested design, error variance decreases; the increase in the number of nested facets in the design restricts the interpretations that can be made in relation to the generalizability of measurements.

G Theory is basically a theory of measurement based on random facets. Therefore, at least one facet should be taken at random in the measurement process. The measurement models in which constant (as well as random) facets are available are known as mixed models in G Theory (Brennan, 1992).

All the above-mentioned examples relate to univariate G Theory. However, some studies are considered in the context of multivariate G Theory. In the first example above, let us suppose that the students are expected to deal with algebraic and analytic problems. Universes of admissible observations correspond to each context and each universe corresponds to one single constant case. In other words, a univariate mixed model can be formulated with a multivariate model, which requires a constant facet; a more flexible representation of the constant facet is thus assured. At a statistical level, multivariate G Theory analyses involve not only variance components, but also co-variance components (Brennan, 2011). In the case of a simpler explanation, the univariate G Theory may be used to analyze the scores obtained from a single test; multivariate G Theory is used to determine the generalizability of scores obtained from a test composed of different sub-tests (Atılgan, 2004; Brennan, 2001; Deliceoglu, 2009).

Three fundamental theories can be used to determine reliability: CTT, Item Response Theory (IRT), and G Theory. Of these, CTT is generally preferred because its underlying mathematical model is easier to understand and its assumptions are flexible (Hambleton & Jones, 1993). Although IRT contains a more complicated mathematical model and its assumptions are difficult to meet, it takes precedence over individual measurement applications because it can generate independent estimates of item and ability parameters. CTT and G Theory focus on test results, while IRT focuses on responses to items (Brennan, 2011).

Various properties distinguish G Theory from the other two theories. Although the mathematical structures of the basic equations in CTT and G Theory are similar, with unobservable values on the right (X = T + E, and equation, respectively), CTT has only one error term, while G Theory permits the division of error terms to reflect different sources of error. G Theory also has a richer conceptual framework than CTT. Two points are particularly relevant: (1) in G Theory, it is possible to distinguish between constant and random facets; (2) G Theory makes it possible to carry out different types of decision studies (Brennan, 2011).

Cronbach et al., (1972) and Brennan (2001) argue that G Theory removes the difference between reliability and validity. One of the most important differences between G Theory and IRT is that, while G Theory focuses on test scores, IRT focuses on item scores. Although items are a constant facet in IRT, they are almost always considered random in G Theory (Brennan, 2011).

In recent years, Turkey has seen an increase in studies conducted using G Theory. Researchers in education and other fields have shown more interest in G Theory because it differs from CTT and IRT and can be advantageous in a number of situations. Given this context, the present study uses various criterial to analyze G Theory-based research carried out in Turkey. Its main purpose is to determine the general tendency of G Theory-based studies and to provide new resources and information to researchers who may have doubts about using G Theory in their own research. To enhance the quality of future academic work, examination themes are also explained in detail. For researchers hoping to contribute to literature on any topic, general trends in current studies in the relevant field, gaps in the literature, and research characteristics presented are very important.

The literature included a review of studies on the use of G Theory: Rios, Li, and Faulkner-Bond (2012), examined 58 studies published in the field of psychology and education between 1997 and 2012, focusing on sample size, the handling of missing data, the question of balance (or unbalance), multiple group comparisons, analysis trends (e.g., computer programs used, methods of estimating variance components), and reporting results. Other than this study, no published research had explored studies conducted using G Theory, indicating a gap in the literature. The present study sets out to provide detailed information on the theoretical and conceptual bases of G Theory to guide researchers aiming to conduct research on the deficiencies of or mistakes made in published studies. The present study makes an important contribution to the literature by promoting the widespread, correct use of G Theory, which is now widely and increasingly studied, by introducing this theory to researchers in the main branches of science, beyond the educational sciences.

The present study therefore examines research carried out using G Theory in the field of education in Turkey using the thematic content analysis method. The following questions guided the current study:

1. What were the aims of the research studies analyzed?
2. Which types of G Theory are used more often in measurement situations?
3. How many faceted designs are used in the study?
4. How is the term “facet” translated into Turkish?
5. Did the object of measurement specify?
6. What is the sample sizes used in the studies?
7. What types of designs are covered?
8. What types of mixed design exist and how can they be used correctly?
9. What proportion of studies fall into the G and D studies categories?
10. Which computer programs are used most frequently?
11. What is the preferred way of explaining negative variance when analyzing research results?
12. What are the various types of fixed facet and how are they discussed?
13. At what rate do studies use balanced or unbalanced patterns?

2. METHOD

2.1. Research Model

The present study has carried out a thematic content analysis of theses and articles based on G Theory in the field of education in Turkey in 2004–2017; the various themes covered here were selected to reveal their similarities and differences. A thematic content analysis involves the synthesis and interpretation of different research findings on the same subject (Au, 2007, Çalık & Sözbilir, 2014, Finfgeld, 2003, Walsh & Downe, 2005). Studies that conduct a thematic content analysis provide a very rich resource to researchers working in related fields, who cannot access all the work in the field or systematically examine those studies (Çalık, Ayas, & Ebenezer, 2005; Ültay & Çalık, 2012). Compared to meta-analyses and descriptive content analysis studies, relatively few studies offer thematic content analyses (Çalık & Sözbilir, 2014).

2.2. Data Collection

All of the education articles incorporating G Theory published in Turkey between 2004 and 2017 were obtained using the Google Academic search engine and/or which were reached in journals indexed by ULAKBIM (national index) and Social Science Citation Index (SSCI). All of the theses in the Council of Higher Education's National Thesis Centre Database of Turkey were also included in the scope of this study. There are no studies carried out in Turkey used G Theory before 2004. For this reason, the starting point for this study was set as 2004. There were 41 articles from 23 different journals and 29 theses published in six different universities. Ten of the articles analyzed were derived from Master's or Ph.D. theses; they were compared with the original M.A. or Ph.D. theses and found to be no different. The reason for excluding articles derived from Master's or doctoral theses was to avoid duplicating studies. Excluding them made it possible to present a more accurate picture of G Theory studies. The elimination of such studies left a total of 60 studies, 31 articles, 20 Master's theses, and 9 Ph.D. theses for content analysis. The investigated studies are listed in Appendix.

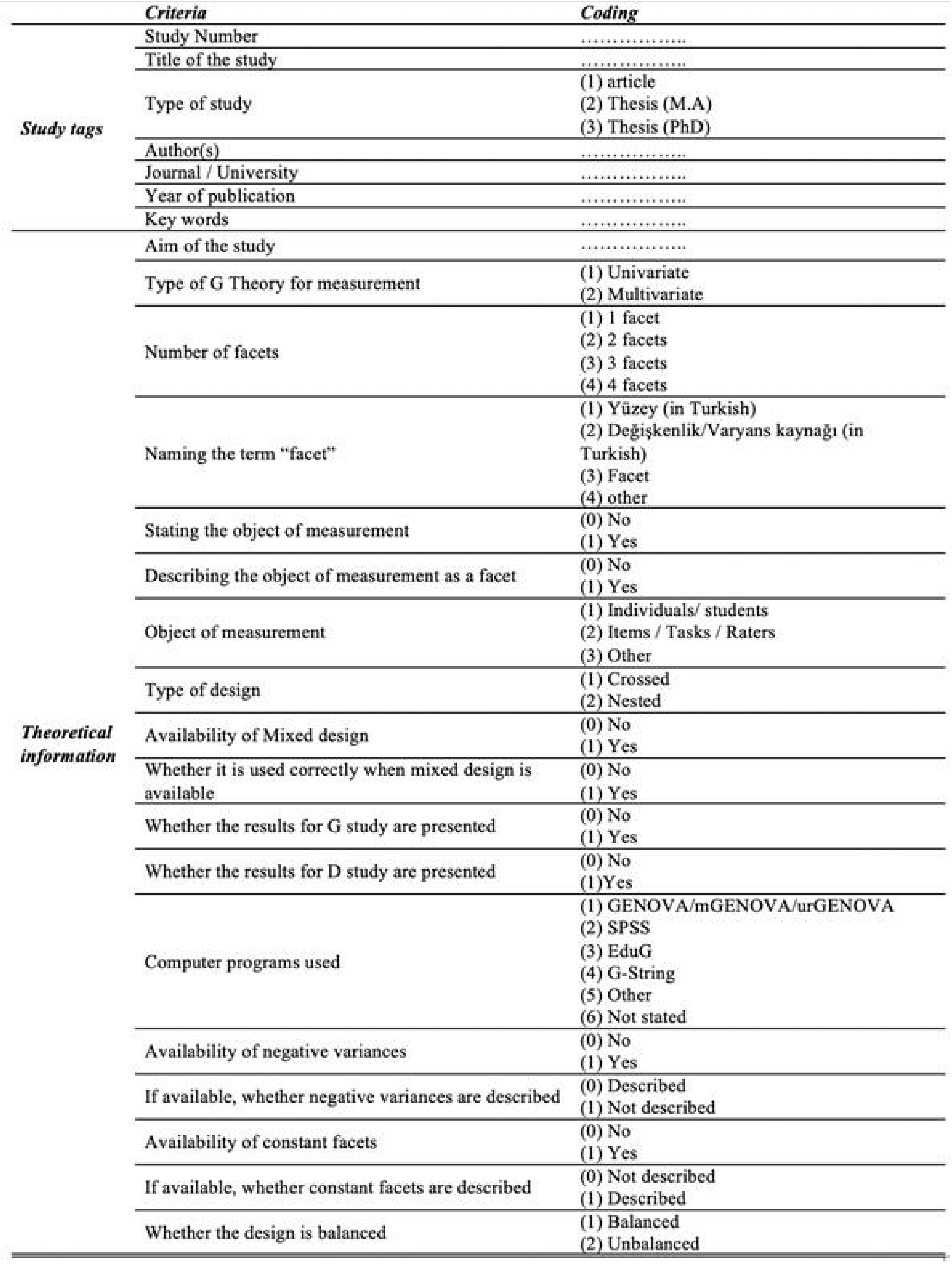
2.3. Data Analysis

Before carrying out the content analysis, the researchers developed a checklist to help them analyze studies incorporating G Theory. The purpose of the checklist was to set standard criteria for analyzing the articles. The checklist had two main parts: "study tag" and "theoretical information." Expert opinions were obtained from three measurement and evaluation specialists, who had carried out studies on G Theory and were able to evaluate the checklist. The specialists recommended including key words and author names in the tags used to describe studies under analysis. The checklist was updated to reflect these views; the version shown in Figure 1 was ultimately used by two researchers in this study.

To ensure consistency across different researchers, five randomly selected studies were examined independently by two researchers. Using the data obtained, Equation 11 (suggested by Miles and Huberman (1994)) was used to calculate consistency between researchers, as follows:

$$\text{Reliability} = \frac{\text{N} \qquad \text{o r}}{\text{Nu} \qquad \text{o r} \qquad +\text{N} \qquad \text{o d}} = .86 \tag{11}$$

The interrater consistency obtained using Equation 11 was calculated as .86. This value should be .80 or above (Miles & Huberman, 1994, Patton, 2002). This result compared to the criterion drawn from the literature, sufficient coherence is obtained. Within the scope of this study, 60 studies were reviewed by researchers, in accordance with the themes in Figure 1, to identify any inconsistencies in the data. Articles or theses with inconsistencies were reviewed by the researchers again independently, to see whether there was any disagreement. For just one study researchers had a disagreement. The researchers came together to discuss the issue and tried to reach agreement, as well as obtaining the opinion of a third independent researcher. As a result of this process, once consent of researchers has been obtained, all the data were combined. Frequency and percentage analyses of codes were carried out for each theme.



	Criteria	***Coding***
Study tags	Study Number	……………..
	Title of the study	……………..
	Type of study	(1) article (2) Thesis (M.A) (3) Thesis (PhD)
	Author(s)	……………..
	Journal / University	……………..
	Year of publication	……………..
	Key words	……………..
Theoretical information	Aim of the study	……………..
	Type of G Theory for measurement	(1) Univariate (2) Multivariate
	Number of facets	(1) 1 facet (2) 2 facets (3) 3 facets (4) 4 facets
	Naming the term "facet"	(1) Yüzey (in Turkish) (2) Değişkenlik/Varyans kaynağı (in Turkish) (3) Facet (4) other
	Stating the object of measurement	(0) No (1) Yes
	Describing the object of measurement as a facet	(0) No (1) Yes
	Object of measurement	(1) Individuals/ students (2) Items / Tasks / Raters (3) Other
	Type of design	(1) Crossed (2) Nested
	Availability of Mixed design	(0) No (1) Yes
	Whether it is used correctly when mixed design is available	(0) No (1) Yes
	Whether the results for G study are presented	(0) No (1) Yes
	Whether the results for D study are presented	(0) No (1)Yes
	Computer programs used	(1) GENOVA/mGENOVA/urGENOVA (2) SPSS (3) EduG (4) G-String (5) Other (6) Not stated
	Availability of negative variances	(0) No (1) Yes
	If available, whether negative variances are described	(0) Described (1) Not described
	Availability of constant facets	(0) No (1) Yes
	If available, whether constant facets are described	(0) Not described (1) Described
	Whether the design is balanced	(1) Balanced (2) Unbalanced

Figure 1. Checklist used in the study

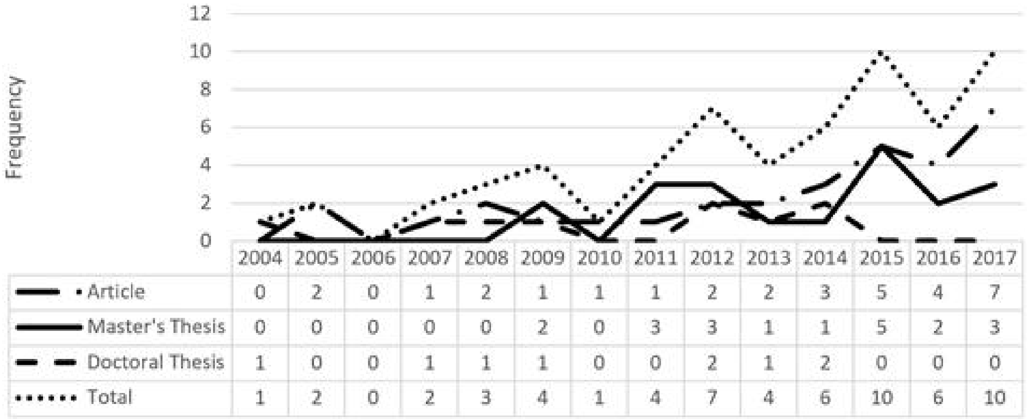
3. RESULT / FINDINGS

The research findings are presented in two parts. The first section headings refer to the tags used to categorize the articles and theses; the second section focuses on theoretical information.

3.1. Findings Related to Tagged Information in the Studies Analyzed

3.1.1. *Year of publication*

Figure 2 shows the distribution of articles and theses by publication year. Although the increase has not been steady, there has clearly been an increase in the number of articles and theses written using G Theory since 2004. While the increase in the number of articles has reached a peak in recent years, the number of Master's theses reached its highest value in 2015; although G Theory continues to be used regularly, frequency has decreased in the last few years. Among doctoral theses, there was an increase between 2012 and 2014; after that date, there is no doctoral thesis conducted on G Theory.



	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017
Article	0	2	0	1	2	1	1	1	2	2	3	5	4	7
Master's Thesis	0	0	0	0	0	2	0	3	3	1	1	5	2	3
Doctoral Thesis	1	0	0	1	1	1	0	0	2	1	2	0	0	0
Total	1	2	0	2	3	4	1	4	7	4	6	10	6	10

Figure 2. Distribution of articles and theses by year

As is clear from Figure 2, the first Turkish doctoral thesis to use G Theory was completed in 2004. It was followed by one doctoral thesis per year in 2007, 2008, 2009, and 2013, and two doctoral theses in 2012 and in 2014. The first Master's thesis was written in 2009 (f=2). While no Master's theses used G Theory in 2010, it was used in 1–5 theses every year after 2011.

3.1.2. *Keywords*

The keywords in the 31 articles and 29 theses were reviewed to determine their frequencies. 109 different key words were used in the studies. The most frequently used word was *Generalizability Theory* (f=56) – as expected. *Reliability* (f=23), *Classical Test Theory* (f=13), *interrater reliability* (f=8) and *decision study* (f=8) were the most frequently used key words. These were followed by *generalizability coefficient* (f=6), generalizability (G) study (f=5) and *Phi coefficient* (f=5), as the basic concepts in G Theory. In addition, 75 key words were used just one time each. "Rating/rater/scoring" concepts were included among the 39 key words; of these concepts, "scoring key/rubric" appears in 12 of them. Computer programs used in G Theory analysis, including EduG, GENOVA, mGENOVA, SAS, and SPSS were used 13 times.

3.2. Findings Involving Theoretical Information

Aim of the study: G Theory can be used in a range of different academic fields. According to the topics, theory comparison was the most commonly studied subject throughout the studies. 14 studies compared G Theory to CTT and Many-Facet Rasch Model (MFRM). In addition, six studies compared performance assessment tools (checklists and analytical and holistic rubrics). In nine studies set out to establish the most appropriate number of raters, the quality of raters involved in evaluation, and interrater reliability. They examined the reliability of scores obtained from measurement tools including the Vee diagram, concept map, multiple-choice tests, structured grids, and performance tasks (f=10). Other studies investigated instructor, peer, and self-evaluation results; standard setting methods; computer software used in analyses (SPSS, GENOVA, EduG, and SAS), and the results obtained from the crossed and nested designs. Various reliability methods were also compared.

Type of G Theory used for measurement: Generalizability Theory can be univariate or multivariate, depending on the measurement situation involved. Most of the studies examined (f=54) included univariate G Theory analyses. Because of its complexity, multivariate G Theory was used in just three Ph.D. theses and three articles written by the same authors.

The number of facets used in the study: Measurement situations with two facets appeared in 45 studies; nine studies included measurement situations in which one and two facets were considered together. It has been observed that one facet in three studies, three facets in two studies, and four facets in one study.

Translating the term "Facet": Since the term was translated into Turkish in different ways in G Theory studies in Turkey, the naming of this concept was also considered. It was most frequently used as "the source of variability (değişkenlik kaynağı)" in the studies analyzed in this study (f= 18). It was used as "surface (yüzey)" in seven studies, as "the source of variance (varyans kaynağı)" and "variable (değişken)" in four studies, and "component (bileşen)" and "variance component (varyans bileşeni)" in one study each. It was called "facet" in one study written in Turkish without finding Turkish concept for it. Only one of the 13 works written in English used "variance source" instead of "facet". In 11 studies, however, it was observed that the term was not considered although the G Theory was used.

Presenting and describing the object of measurement as a facet: Only two studies considered the object of measurement as a facet. While 44 studies clearly defined the object of measurement, 14 gave no explanation. When observations related to the state of the measurement object, only 25 studies clearly defined the object of measurement. Of the remaining 35 studies, 23 provided no information and 10 provided the correct usage. Two studies failed to mention the measurement object and used the concept in the wrong way.

Sample size: Table 1 shows the sample sizes used in these studies.

Of the studies investigated, 16 had sample sizes below 30, while 25 had sample sizes between 30 and 100. Only 19 studies had samples larger than 100.

Table 1. Object of Measurement Sample Size

Object of measurement	Sample size	Frequency	Total
Person	<30	11	55
	30–100	25	
	102–187	10	
	203–249	5	
	309	1	
	689	1	
	1000	1	
	1500	1	
Item	16	1	3
	18	1	
	20	1	
Occasion	7	1	1
Task	6	1	1

Type of design: The most frequently used design was a crossed design (f=48). Nested designs were used in seven studies, while five studies used both crossed and nested designs.

The availability and correct use of mixed designs: Only six of the 60 studies used mixed designs. One study that claimed to have a mixed design actually had a random design.

The presentation of G and D study results: G study results were presented as expected in most studies except one. In addition, D study results were not given in only three studies. It was observed that the D study was not performed because the purpose of these studies was not to estimate the reliability for different measurement situations.

Computer programs used: An evaluation of the computer programs used in the analyses, revealed that the most frequently used program was EduG (f=32). The second most frequently used program was SPSS (f=16). GENOVA was used in nine studies, mGENOVA in five studies, and G-String, R, and SAS were used in two studies each. In five studies, the computer program was not specified. Since some studies used more than one program (e.g., SPSS-EduG and SPSS-GENOVA), the sum of the frequency values above exceeds the number of studies examined.

The availability and description of negative variance: In 21 studies, a negative variance was observed. Adopting the approach of Cronbach et al., the negative variance was treated as zero in 11 studies; zero was also used to estimate other variance components. In 10 studies, adopting Brennan's approach, the negative variance was regarded as zero and used as it was in estimates of other variance components.

The availability and description of fixed facets: Six of the studies analyzed had constant facets. Only two explained these constant designs to readers.

Design balance: Only six of the 60 Turkish studies had an unbalanced design. Of these, two were Ph.D. theses and one was an article written by one of the Ph.D. authors.

4. DISCUSSION, CONCLUSION and SUGGESTIONS

The present study analyzed 60 Turkish studies in two stages, using tag information and their theoretical foundations. An examination of the years in which G Theory studies were published revealed an overall increase in publications, despite occasional decreases. The theoretical structure of G Theory is complex and difficult to analyze at elementary levels; for this reason, it is used primarily in doctoral theses. However, beginning with the year 2009, it has also appeared in Master's theses. One of the reason for this may be that G Theory Master's level analyses are now being conducted by means of user-friendly computer programs, such as EduG, rather than the more advanced GENOVA. Another reason may be the increasing number of workshops are held at congresses and courses are taught in Master's and doctoral programs. User friendly computer programs will make it possible to carry out more analyses based on G theory for various studies. Another explanation for this result is the increase in the number of researchers working in the area of measurement and evaluation. In particular, applications for research-assistant posts in the field of measurement and evaluation have been increased since 2002, resulting in a larger number of researchers working in the field and therefore more studies based on G Theory.

The keywords presented in the study tags were also analyzed, revealing that G Theory was used most often in studies involving interrater dependability and standard settings. The fact that 88 words appeared only once or twice appears to show that a range of studies on diverse topics have been carried out using G Theory. Among studies that feature rating and rater keywords (f = 39), G Theory is frequently discussed in relation to rater reliability, consistency, the rater effect, the number of raters, the reliability of ratings, and rating methods. Although CTT is used more frequently than G Theory in the literature, it cannot be used to determine the number of raters needed to obtain more reliable results or the number of criteria to include in scoring keys. The information available to G Theory through D studies may make it a better choice than CTT, especially in such studies.

An investigation of the aims of the studies in question found that they made a great number of theoretical comparisons. By comparison, the most frequently used theory is CTT, which has been used for many years in the literature and is better known than G Theory in the theoretical literature. Another theory often used in theory comparisons is MFRM. It is thought that this model, which is covered in IRT, tends to be preferred because it allows analyses to be carried out using fewer parameters than other IRT models. Like G theory, MFRM is frequently used to determine the reliability of a rater; it can consider more than one error source at the same time. These can also be cited as reasons for comparing G Theory to MFRM. As well as being used in theory comparisons, G Theory was also preferred when researchers wished to determine the reliability of scores obtained using various measurement tools. G Theory has the advantage of being able to simultaneously handle many sources of error in a measurement process at the same time. While this topic is not new in the literature, many studies have investigated the reliability of self and peer assessments, which have been discussed more frequently in recent years. G Theory makes it possible to evaluate the rater as a facet, while also evaluating the points of self, peers, and instructors as conditions of this facet. Since G Theory makes it possible to estimate the magnitude of the variance between evaluations of different raters and the reliability coefficient, it may be preferred in such studies. A final category of studies compared the results obtained using different types of computer software able to carry out G Theory analyses. Because of free and user-friendly software and their manuals, the use of G theory potentially increases.

Most of the studies analyzed in the course of this research were produced using univariate G Theory. Only five studies used multivariate G Theory, potentially reflecting the following two factors: (1) the situations considered by researchers were better suited to the use of univariate G Theory, and (2) researchers preferred not to use multivariate G Theory because it was relatively difficult and complicated to analyze.

It was found that the most of the studies investigated used two-faceted measurement designs. Frequency measurement situations with two surfaces may reflect standard educational practice (items and raters as facets). Very few of these studies had three- or four-faceted designs. As the number of facets increases in G Theory, the number of estimated variance values for each facet and the interactions between facets increase. Interactions are therefore difficult to interpret. For example, where a two-faceted crossed design consists of seven components of variance, in a three-faceted design, this number increases to 14. Researchers tend to avoid highly faceted designs because it is difficult to interpret the large number of variance components that result from the increased number of facets.

The concept of “source of variability” was used in almost half of the studies analyzed instead of the English term “facet” available in the G Theory. The use of agreed on words can be supported by reaching an agreement on Turkish equivalents to the English terms and compiling them in a glossary, and thus comprehensibility of the G Theory studies can be increased. Indeed, there is such a glossary study conducted in Turkey by the Association of Measurement and Evaluation in Education and Psychology, and accordingly, it is recommended that the words “yüzey (facet)” or “değişkenlik kaynağı (source of variability)” be used as corresponding to English word “facet”. Yet, it was observed that using “source of variability” as Turkish equivalence to the English word “facet” could cause confusions in studies conducted in Turkish. The sentence “A one-facet design has four sources of variability” from an important resource book, “Generalizability Theory: A Primer” by Shavelson and Webb (1991, p.4) would exemplify our claim because the translation of the sentence into Turkish would also cause confusion. Considering this situation, it is thought by researchers that using “yüzey” rather than “değişkenlik kaynağı” in Turkish as equivalence to English “facet” would be more appropriate.

In G Theory, the objects to be measured such as students, individuals, methods etc. and decisions will be made on it known as the object of measurement. Differences in the object of measurement are defined as “the sources of error” in CTT, since variance that depends on the object of measurement is a desired situation. These differences are not considered to be facets in G Theory either. Two of the studies examined presented this idea inaccurately. First, the difference between the concepts of facet and object of measurement is difficult to understand. Second, the fact that “facet” is used to express the source of variability in the metric target state, a distinction that does not exist in MFRM (a theory that G Theory is often compared to) may add to this confusion. Researchers should therefore be encouraged to clarify which sources of variability discussed in their studies are facets or objects of measurement. In 54 of the 60 studies examined, the object of measurement was the individual. Items, tasks, situations, and raters can all be objects of measurement, depending on the type of measurement. It is very important for researchers to clearly define the object of measurement and to accurately define it as a measured object to ensure an accurate interpretation of the findings.

Sample size is quite important in most statistical methods, as it influences the accuracy of estimates and can increase or reduce errors. Of the studies examined here, 44 had a sample size of 30 or more. This ratio was considerably higher than the value obtained when Rios, Li and Faulkner-Bond (2012) conducted 58 studies using G Theory between 1997 and 2012. They found that the mod of the sample size was 20. Atılgan (2013) examined the effect of sample size on the G and Phi coefficients and found that it was impossible to make stable predictions if the sample size was 30 or below. Results could be considered sufficiently unbiased if the sample size was 50, 100, 200, or 300. With a sample size of 400, the result can be considered definite. In the context of educational sciences studies undertaken using G Theory in Turkey, the majority of results are based on an adequate sample size. However, the samples are small in some of these studies. Due to logistical, economic, and time constraints related to data collection, some studies based on G Theory have been carried out using smaller samples (Rios, Li, & Faulkner-Bond, 2012). At this point, a balance must be established between the need to increase sample sizes to achieve a correct estimate of variance components and the pressures of staff, time, and cost limitations. While a G Theory sample consisting of 20 individuals is considered the lower limit (Webb, Rowley, & Shavelson, 1988), more accurate estimates can be obtained from larger samples. In future studies, it may be advisable to use the largest sample size possible. In particular, researchers struggling with unbalanced, complex designs may end up deleting data and changing to a balanced format. Such situations reduce the size of study samples. In recent years, it has become possible to overcome these difficulties via user-friendly software, capable of carrying out unbalanced pattern analyses.

Many studies in the literature have been conducted using G Theory; the majority of studies examined here (f = 47) have adopted a crossed design, possibly because all possible sources of variance can be estimated in fully crossed designs. Only in fully crossed designs can researchers access the variance values of each source of variability, as well as their interactions. In nested designs, it is not possible to estimate the variances of nested facets alone. For this reason, crossed designs may be preferred to nested designs. In the studies investigated here, the nested facet in nested designs was generally raters. Particularly in a performance assessment, it may not be possible for each individual to be evaluated by all of the raters. Such measurement situations should be monitored in relation to variables including cost and the effective use of the resources. If it is not possible to use a crossed design, the study should be carried out using nested designs.

G Theory is basically a theory of measurement with random facets; for this reason, at least one facet should be random (Güler et al., 2012). If at least one facet in a design is constant and the other facets are random, the design is said to be mixed. The present investigation of studies conducted using G Theory in Turkey found that only six of the 60 studies examined used the term "mixed design". In only one of these six studies, the mixed design was defined as "the combined use of crossed and nested facets". On the other hand, in basic sources of G theory in the literature mixed models are defined as the the measurement models in which constant (as well as random) facets are available are known as mixed models in G Theory (Brennan, 1992; Shavelson & Webb, 1991), and the other five studies used the mixed model in accordance with this definition. Actually, those five studies used multivariate G Theory. While multivariate G Theory is used to generalize scores obtained from a test containing different sub-tests, the sub-tests are regarded as constant facets (Brennan, 2001). For this reason, multivariate G Theory studies are essentially mixed design studies. Although one study analyzed here had a random design, it was presented as mixed, possibly because the crossing and nesting of facets in the same design was wrongly perceived as indicating a mixed design. In describing a design as mixed, it is important not to use crossed and nested designs together; at least one of the facets must be fixed. Taking this into consideration will enable the terminology of G Theory to be used more accurately.

In G Theory, it is possible to investigate dependability in two stages, via a Generalizability (G) study and a Decision (D) study (Brennan, 2001; Goodwin, 2001; Shavelson & Webb, 1991). Except in one of the studies examined, G study results were shared. It is appropriate to share the results of analyses obtained to serve the purpose of a study. If the variance components estimated as a result of the G study are not evaluated, it may be sufficient to share only the predicted G and Phi coefficients. The present study also considered the D studies carried out within G Theory by 56 of the studies examined. In education, researchers frequently investigate ways to reduce error and improve the reliability of measurements designed for particular purposes. The fact that most of the examined studies carried out D studies, which serve this purpose within the framework of G Theory, may reflect a general effort to increase credibility.

An evaluation of computer programs revealed that a majority of studies used EduG. This may reflect the fact that EduG is free and relatively user-friendly. The common use of G Theory depends not only on the need for available reference materials that clearly and comprehensibly explain its theoretical foundations, but also on the availability of user-friendly computer programs to perform analyses. The first computer program developed to carry out G Theory analyses was GENOVA, developed by Brennan in 1983. Brennan wrote a detailed explanation of both univariate and multivariate G Theory in “Generalizability Theory,” released in 2001; the author developed mGENOVA for multivariate analyses, and urGENOVA to estimate balanced and unbalanced designs and random effect variance components. The syntax in which G Theory analyses could be performed on the SPSS, SAS, and MATLAB programs was released in 2006. The fact that this syntax has been organized for use with relatively common programs is a positive step toward expanding the use of G Theory. In 2006, Jean Cardinet released EduG Program. Finally, in 2011, the G-String Program—which can also be used for unbalanced designs—was produced by Bloch and Norman. In addition to these software tools, it is possible to carry out a G Theory analysis via the “gtheory” package using R, which is a free software. It is possible to see the impact of software by examining the yearly distribution of G Theory studies (see Figure 2). Improved software compatible with G Theory analyses and the publication of user guides will increase its use among researchers from various disciplines. In this context, computer programs are clearly important.

As the literature indicates, variance estimates can sometimes be negative. Negative estimates are caused by erroneous measurement models or sampling errors (Güler et al., 2012). Since a negative variance indicates the wrong choice of models or samples, precautions should be taken in cases where the variance is negative. 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 (1983, 2001), 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). When the variance is negative, the value is either replaced by zero or used as it is. The decision-makers are not researchers, but computer programs. None of the analyzed studies explained this situation. Depending on the software that are used, researchers can access the analysis results or carry out estimates using both approaches.

Another point of importance in G Theory is whether or not the design is balanced. In a balanced design, the number of observations is equal at every level of the source of variability. However, observations per variable are not equal in an unbalanced design, due to lost data or differences between the number of observations and the levels of variables (Brennan, 2001). Let us consider, for instance, a measurement situation in which individuals respond to two different testlets, and in which the items are nested within the testlets and the individuals are crossed with them. If the testlets have an equal number of items, the design is balanced. If each testlet has a different number of items, then the design is unbalanced. In other words, if there are three items in each testlet, the design is balanced—but if there are two items in one testlet and four in the other, the design is unbalanced. The fact that unbalanced designs have been used in studies based on G Theory in recent years indicates that researchers are considering using the theory in different measurement situations. The very small number of studies using unbalanced designs (f=3) may reflect the fact that designs involving unbalanced data are relatively complex. Another explanation may be that the researchers have removed some data or filled in the missing data to change their unbalanced designs into balanced ones. One final explanation may be that, previously, G Theory analyses of unbalanced data could only be carried out using the urGENOVA program—which was very complex and avoided by researchers. The G String program produced in 2011 by Bloch and Norman (used in three unbalanced-design studies) is an easy-to-use and useful program. This program may become widespread and commonly used with the researches conducted with unbalanced datasets.

Since the study has offered both detailed conceptual and theoretical explanations, as well as information on general biases, it can serve as a resource for prospective researchers. One limitation of the present research is the fact that it focused on studies in the field of educational sciences. Despite this, it provides information that could be of substantial value to the researchers in other fields, helping to promote the widespread and accurate use of G Theory in many other scientific fields.

Initiatives designed to increase the use of G Theory, such as organizing seminars and workshops, supervising post-graduate theses, and writing books and articles to inform researchers, will raise awareness among scientists, encourage them to use G Theory and increase its use. G Theory can also be introduced to all departments in educational faculties and medical schools. Many departments within faculties of education, health sciences (a field in which G Theory studies are relatively common), and educational sciences fields that carry out measurement and evaluation research will also benefit from using G Theory. Since there can be multi faceted measurement designs in the above mentioned fields, carrying out the researches by using G Theory can improve the qualifications of those studies. For instance, there are studies related to special education (Pekin, Çetin, & Güler, 2018), science education

(Shavelson, Baxter, & Pine, 2009; Yin & Shavelson, 2008), mathematics education (Kersting, 2008; Lane, Liu, Ankenmann, & Stone, 1996), medical education (Lafave & Butterwick, 2014; Turner, Lozano-Nieto, & Bouffard, 2006) and dentistry education (Taşdelen Teker & Odabaşı, 2019; Gadbury-Amyot et al., 2014).

ORCID

Gülşen Taşdelen Teker https://orcid.org/0000-0003-3434-4373
Neşe Güler https://orcid.org/0000-0002-2836-3132

5. REFERENCES

Arık, R. S. & Türkmen, M. (2009). *Examination of the articles in the scientific journals published in the field of educational sciences.* Paper presented at I. International Congress of Educational Research, Çanakkale, Turkey.

Atılgan, H. (2004). *Genellenebilirlik kuramı ve çok 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.

Atılgan, H. (2013). Sample size estimation of G and Phi coefficients in generalizability theory. *Eurasian Journal of Educational Research*, *51*, 215–228.

Au, W. (2007). High-stakes testing and curricular control: A qualitative metasynthesis. *Educational Researcher*, *36*, 258–267.

Bektaş, M., Dündar, H. & Ceylan, A. (2013). Investigation of several variables papers national classroom teacher education symposium. *Uşak University Journal of Social Sciences*, *6*(2), 201–226. DOI: http://dx.doi.org/10.12780/UUSBD167

Bloch, R. & Norman, G. (2011). *G String 4 user manual* (Version 6.1.1). Hamilton, Ontario, Canada. Retrieved from http://fhsperd.mcmaster.ca/g_string/download/g_string_4_manual_611.pdf

Brennan, R. L. (2011). Generalizability Theory and Classical Test Theory. *Applied Measurement in Education, 24*, 1–21. doi:10.1080/08957347.532417.

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

Brennan, R. L. (1997). A perspective on the history of Generalizability Theory. *Educational Measurement: Issues and Practice*, *16*(4), 14–20. https://doi.org/10.1111/j.1745-3992.1997.tb00604.x

Brennan, R. L. (1992). Elements of Generalizability Theory. NY: Springer-Verlag.

Çalık, M. & Sözbilir, M. (2014). Parameters of content analysis. *Education and Science*, *39*(174), 33–38. doi:10.15390/EB.2014.3412

Çiltaş, A. (2012). Content analysis of the graduate thesis and dissertations in mathematics education in Turkey between 2005-2010. *The Journal of Academic Social Science Studies*, *5*(7), 211–228.

Çiltaş, A., Güler, G. & Sözbilir, M. (2012). Mathematics education research in Turkey: A content analysis study. *Educational Sciences: Theory & Practice*, *12*(1), 565–580.

Deliceoğlu, G. (2009). *The Comparison of the reliabilities of the soccer abilities'rating scale based on the Classical Test Theory and Generalizability*. (Doctoral Dissertation). Ankara University, Ankara.

Doğru, M., Gençosman, T., Ataalkın, A. N. & Şeker, F. (2012). Fen bilimleri eğitiminde çalışılan yüksek lisans ve doktora tezlerinin analizi [Analysis of master's and doctoral theses in science education]. *Journal of Turkish Science Education*, *9*(1), 49–64.

Finfgeld, D. L. (2003). Metasynthesis: The state of the art-so far. *Qualitative Health Research*, *13*(7), 893–904. DOI: 10.1177/1049732303253462

Gadbury-Amyot, C. C., Kim, J., Palm, R. L., Mills, G. E., Noble, E. & Overman, P. R. (2003). Validity and reliability of portfolio assessment of competency in a baccalaureate dental hygiene program. *Journal of Dental Education, 67*(9), 991-1002.

Göktaş, Y., Küçük, S., Aydemir, M., Telli, E., Arpacık, Ö., Yıldırım, G. & Reisoğlu, İ. (2012). Educational technology research trends in Turkey: A content analysis of the 2000–2009 decade. *Educational Sciences: Theory & Practice*, *12*(1), 191–196.

Gülbahar, Y. & Alper, A. (2009). Trends and issues in educational technologies: A review of recent research in TOJET. *The Turkish Online Journal of Educational Technology – TOJET*, *8*(2), 124-135.

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

Güler, N. (2008). *A research on classical test theory, generalizability theory and Rasch model* (Doctoral Dissertation). Hacettepe University, Ankara.

Günay, R. & Aydın, H. (2015). Inclinations in studies into multicultural education in Turkey: A content analysis study. *Education and Science*, *40*(178), 1–22. DOI: http://dx.doi.org/10.15390/EB.2015.3294

Hambleton, R. K. & Jones, R. W. (1993). Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practice*, *12*(3), 3847. http://dx.doi.org/10.1111/j.1745-3992.1993.tb00543.x

Kaleli Yılmaz, G. (2015). Analysis of technological pedagogical content knowledge studies in Turkey: A meta-synthesis study. *Education and Science*, *40*(178), 103–122. DOI: http://dx.doi.org/10.15390/EB.2015.4087

Karadağ, E. (2009). Eğitim bilimleri alanında yapılmış doktora tezlerinin incelenmesi [A Thematic Analysis on Doctoral Dissertations Made In the Area of Education Sciences],. *Ahi Evran Üniversitesi Eğitim Fakültesi Dergisi 10*(3), 75–87.

Kersting, N. (2008). Using Video Clips of Mathematics Classroom Instruction as Item Prompts to Measure Teachers' Knowledge of Teaching Mathematics. *Educational and Psychological Measurement*, *68*(5), 845-861. DOI:10.1177/0013164407313369

Kılıç Çakmak, E., Çebi, A., Mihçi, P., Günbatar, M. S. & Akçayir, M. (2013). *A content analysis of educational technology research in 2011*. 4th International Conference on New Horizons in Education. INTE 2013 Proceedings Book, 397–409.

Lafave, M. R. & Butterwick, D. J. (2014). A generalizability theory study of athletic taping using the Technical Skill Assessment Instrument. *Journal of Athletic Training, 49*(3), 368-372. doi: 10.4085/1062-6050-49.2.22

Lane, S., Liu, M., Ankenmann, R. D. & Stone, C. A. (1996). Generalizability and Validity of a Mathematics Performance Assessment. *Journal of Educational Measurement*, *33*(1), 71-92. https://doi.org/10.1111/j.1745-3984.1996.tb00480.x

Miles, M, B., & Huberman, A. M. (1994). *Qualitative data analysis: An expanded Sourcebook*. (2nd ed). Thousand Oaks, CA: Sage.

Patton, M.Q. (2002). *Qualitative research and evaluation methods* (3rd Ed.). London: Sage Publications, Inc.

Pekin Z., Çetin S. & Güler N. (2018). Comparison of Interrater Reliability Based on Different Theories for Autism Social Skills Profile. *Journal of Measurement and Evaluation in Education and Psychology, 9*(2), 202-215. https://doi.org/10.21031/epod.388590

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.

Saban, A. (2009). Content analysis of Turkish studies about the multiple intelligences theory. *Educational Sciences: Theory & Practice, 9*(2), 833–876.

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

Shavelson, R. J., Baxter, G. P. & Pine, J. (1991). Performance Assessment in Science. *Applied Measurement in Education*, *4*(4), 347-362. DOI: 10.1207/s15324818ame0404_7

Taşdelen Teker, G. & Odabaşı, O. (2018). Reliability of scores obtained from standardized patient and instructor assessments. *European Journal of Dental Education*, *23*, 88-94. DOI: 10.1111/eje.12406

Turner, A. A., Lozano-Nieto, A. & Bouffard, M. (2006). Generalizability of extracellular-to-intracellular fluid ratio using bio-impedance spectroscopy. *Physiological Measurement*, *27*(4), 385-397. DOI: 10.1088/0967-3334/27/4/005

Yalçın, S., Yavuz, H. Ç. & İlgün Dibek, M. (2015). Content analysis of papers published in educational journals with high ımpact factors. *Education and Science*, *40* (182), 1–28. DOI:10.15390/EB.2015.4868

Yin, Y. & Shavelson, R. J. (2008). Application of Generalizability Theory to Concept Map Assessment Research. *Applied Measurement in Education*, *21*(3), 273-291. DOI: 10.1080/08957340802161840

Walsh, D. & Downe, S. (2005). Meta-synthesis method for qualitative research: A literature review. *Journal of Advanced Nursing*, *50*(2), 204–211.

Webb, N. M., Rowley, G. L., & Shavelson, R. J. (1988). Using Generalizability Theory in counseling and development. *Measurement and Evaluation in Counseling and Development*, *21*, 81–90.

6. APPENDIX: List of studies included in research

Atılgan, H. (2004). *A Reseach on comparisons of Generalizability Theory and many facets Rasch measurement* (Unpublished doctoral dissertation). Hacettepe University, Ankara, Turkey.

Aktaş, M. (2013). *An investigation of the reliability of the scores obtained through rating the same performance task with three different techniques by different numbers of raters according to Generalizability Theory* (Unpublished master's thesis). Mersin University, Mersin, Turkey.

Alkan, M. (2013). *Comparison of different designs ın scoring of PISA 2009 reading open ended ıtems according to Generalizability Theory* (Unpublished doctoral dissertation). Hacettepe University, Ankara, Turkey.

Anadol, H. Ö. (2017). *The examination of realiability of scoring rubrics regarding raters with different experience years* (Unpublished master's thesis). Ankara University, Ankara, Turkey.

Arsan, N. (2012). *Investigation of the raters' assessment in ice skating with Generalizability Theory and Rasch measurement* (Unpublished doctoral dissertation). Hacettepe University, Ankara, Turkey.

Bağcı, V. (2015). *The comparison of different designs in generalizability theory with Classical Test Theory in the measurement of mathematical reasoning ability* (Unpublished master's thesis). Gazi University, Ankara, Turkey.

Başat, B. (2015). *A generalizability analysis of the reliability of measurements, applied in the sixth grade science course "Let's circuit electric" unit* (Unpublished master's thesis). Gaziosmanpaşa University, Tokat, Turkey.

Büyükkıdık, S. (2012). *Comparison of interrater reliability based on the Classical Test Theory and Generalizability Theory in problem solving skills assessment* (Unpublished master's thesis). Hacettepe University, Ankara, Turkey.

Çakıcı Eser, D. (2011). *Comparison of interrater agreement calculated with Generalizability Theory and logistic regression* (Unpublished master's thesis). Hacettepe University, Ankara, Turkey.

Deliceoğlu, G. (2009). *The comparison of the reliabilities of the soccer abilities' rating scale based on the Classical Test Theory and Generalizability* (Unpublished doctoral dissertation). Ankara University, Ankara, Turkey.

Güler, N. (2008). *A research on Classical Test Theory, Generalizability Theory and Rasch Model* (Unpublished doctoral dissertation). Hacettepe University, Ankara, Turkey.

Gülle, T. (2015). *Development of a speaking test for second language learners of Turkish* (Unpublished master's thesis). Boğaziçi University, İstanbul, Turkey.

Gündeğer, C. (2012). *A comparison of Angoff, Yes/No and Ebel standard setting methods* (Unpublished master's thesis). Hacettepe University, Ankara, Turkey.

Kaya Uyanık, G. (2014). *Investigation of two facets design with generalizability in item response modeling* (Unpublished doctoral dissertation). Hacettepe University, Ankara, Turkey.

Kaya, G. (2011). *Application of Generalizability Theory to fill-in concept map assessment* (Unpublished master's thesis). Hacettepe University, Ankara, Turkey.

Kızıltoprak, F. (2016). *The comparision of reliability of PISA mathematical literacy items' competency needs scores obtained with comptency scheme based on Generalizability Theory and Classical Test Theory* (Unpublished master's thesis). Gazi University, Ankara, Turkey.

Küçük, F. (2017). *Assessing academic writing skills in Turkish as a foreign language* (Unpublished master's thesis). Boğaziçi University, İstanbul, Turkey.

Nalbantoğlu Yılmaz, F. (2012). *Comparison of balanced and unbalanced designs in Generalizability Theory* (Unpublished doctoral dissertation). Ankara University, Ankara, Turkey.

Nalbantoğlu, F. (2009). *Comparison of different designs in accordance with the Generalizability Theory in performance measurements* (Unpublished master's thesis). Hacettepe University, Ankara, Turkey.

Özberk, E. H. (2012). *Comparing different coefficients in Generalizability Theory decision studies* (Unpublished master's thesis). Hacettepe University, Ankara, Turkey.

Öztürk, M. E. (2011). *The comparison of reliability of the "volleyball abilities observation form" (vaof) points to the Generalizability and the Classical Test Theory* (Unpublished master's thesis). Hacettepe University, Ankara, Turkey.

Pekin, Z. (2015). *Comparison of interrater reliability based on Classical Test Theory and Generalizability Theory for autism social skills profile* (Unpublished master's thesis). Hacettepe University, Ankara, Turkey.

Şalgam, A. (2016). *The comparison of reliability of the Generalizability Theory and the test-retest technique for the short answered maths exam* (Unpublished master's thesis). Gazi University, Ankara, Turkey.

Taşdelen Teker, G. (2014). *The effect of testlets on reliability and differential item functioning* (Unpublished doctoral dissertation). Hacettepe University, Ankara, Turkey.

Taşdelen, G. (2009). *A comparison of angoff and nedelsky cutting score procedures using generalizability theory* (Unpublished master's thesis). Hacettepe University, Ankara, Turkey.

Taştan, Z. (2017). *Investigation of multi-surface patterns in generalizability* (Unpublished master's thesis). Mersin University, Mersin, Turkey.

Yelboğa, A. (2007). *The examination of reliability according to classical test and generalizability theory on a job performance scale* (Doctoral Dissertation). Hacettepe University, Ankara, Turkey.

Yıldıztekin, B. (2014). *The comparison of interrater reliability by using estimating tecniques in classical test theory and generalizability theory* (Unpublished master's thesis). Hacettepe University, Ankara, Turkey.

Yüksel, M. (2015). *Comparison of scores obtained from different measurement tools used in the determination of student achievement* (Unpublished master's thesis). Gaziosmanpaşa University, Tokat, Turkey.

Acar Güvendir, M. & Güvendir, E. (2017). The determination of an english speaking exam's data reliability using Generalizability Theory. *Trakya University Journal of Education*, *7*(1), 1-9.

Anıl, D. & Büyükkıdık, S. (2012). An example application for the use of four facet mixed design in Generalizability Theory. *Journal of Measurement and Evaluation in Education and Psychology*, *3*(2), 291-296.

Atılgan, H. (2005). Generalizability Theory and a sample application for inter-rater reliability. *Educational Sciences and Practice*, *4*(7), 95-108.

Atılgan, H. (2008) Using Generalizability Theory to assess the score reliability of the Special Ability Selection Examinations for music education programmes in higher education. *International Journal of Research & Method in Education*, *31*(1), 63-76, DOI:10.1080/17437270801919925

Atılgan, H. (2013). 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). An investigation on consistency of G and Phi coefficients obtained by Generalizability Theory alternative decisions. *Eurasian Journal of Educational Research*, *5*(18), 28-40.

Can Aran, Ö., Güler, N. & Senemoğlu, N. (2014). 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.

Çetin, B., Güler, N. & Sarıca, R. (2016). Using Generalizability Theory to examine different concept map scoring methods. *Eurasian Journal of Educational Research*, *66*, 212-228. http://dx.doi.org/10.14689/ejer.2016.66.12

Doğan, C. D. & Anadol, H. Ö. (2017). Comparing fully crossed and nested designs where ıtems nested in raters in Generalizability Theory. *Kastamonu Education Journal*, *25*(1), 361-372.

Doğan, C. D. & Uluman, M. (2017). A comparison of rubrics and graded category rating scales with various methods regarding raters' reliability. *Educational Sciences: Theory & Practice*, *17*, 631–651. http://dx.doi.org/10.12738/estp.2017.2.0321

Gözen, G. & Deniz, K. Z. (2016). Comparison of instructor and self-assessments on prospective teachers' concept mapping performances through Generalizability Theory. *International Journal on New Trends in Education and Their Implications*, *7*(1), 28-40.

Güler, 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.

Güler, N. (2011). The comparison of reliability according to Generalizability Theory and Classical Test Theory on random data. *Education and Science*, *36*(162), 225-234.

Güler, N. & Gelbal, S. (2010). Studying reliability of open ended mathematics ıtems according to the Classical Test Theory and Generalizability Theory. *Educational Sciences: Theory & Practice*, *10*(2), 989-1019.

Güler, N. & Taşdelen Teker, G. (2015). The evaluation of rater reliability of open ended items obtained from different approaches. *Journal of Measurement and Evaluation in Education and Psychology*, *6*(1), 12-24.

Güler, N., Eroğlu, Y. & Akbaba, S. (2014). Reliability of criterion-dependent measurement tools according to Generalizability Theory: Application in the case of eating skills. *Abant İzzet Baysal University Journal of Education*, *14*(2), 217-232.

Han, T. (2017). scores assigned by ınexpert efl raters to different quality EFL compositions, and the raters' decision-making behaviors. *International Journal of Progressive Education*, *13*(1), 136-152.

Han, T. & Ege, İ. (2013). Using Generalizability Theory to examine classroom ınstructors' analytic evaluation of EFL writing. *International Journal of Education*, *5*(3), 20-35.

İlhan, M. & Gezer, M. (2017). A comparison of the reliability of the Solo- and revised Bloom's Taxonomy-based classifications in the analysis of the cognitive levels of assessment questions. *Pegem Eğitim ve Öğretim Dergisi*, *7*(4), 637-662, http://dx.doi.org/10.14527/pegegog.2017.023

Kamış, Ö. & Doğan, C. D. (2017). How consistent are decision studies in G Theory? *Gazi University Journal of Gazi Educational Faculty*, *37*(2), 591-610.

Kan, A. (2007). Effects of using a scoring guide on essay scores: Generalizability Theory. *Perceptual and Motor skills, 105,* 891-905.

Kara, Y. & Kelecioğlu, H. (2015). Investigation the effects of the raters' qualifications on determining cutoff scores with Generalizability Theory. *Journal of Measurement and Evaluation in Education and Psychology*, *6*(1), 58-71.

Nalbantoğlu Yılmaz, F. (2017). Reliability of scores obtained from self-, peer-, and teacher-assessments on teaching materials prepared by teacher candidates. *Educational Sciences: Theory & Practice*, *17*, 395–409. http://dx.doi.org/10.12738/estp.2017.2.0098

Nalbantoğlu Yılmaz, F. & Başusta, B. (2015). Using Generalizability Theory to assess reliability of suturing and remove stitches skills station. *Journal of Measurement and Evaluation in Education and Psychology, 6*(1), 107-116.

Öğretmen, T. & Acar, T. (2014). Estimation of generalizability coefficients: An application of structural equation modeling. *Journal of Education and Practice*, *5*(14), 113-118.

Polat Demir, B. (2016). The examination of reliability of vee diagrams according to Classical Test Theory and Generalizability Theory. *Journal of Measurement and Evaluation in Education and Psychology*, *7*(2), 419-431.

Taşdelen Teker, G., Güler, N. & Kaya Uyanık, G. (2015). Comparing the effectiveness of SPSS and EduG using different designs for Generalizability Theory. *Educational Sciences: Theory & Practice*, *15*(3), 635-645. DOI: 10.12738/estp.2015.3.2278

Taşdelen Teker, G., Şahin, M. G. & Baytemir, K. (2016). Using Generalizability Theory to investigate the reliability of peer assessment. *Journal of Human Sciences*, *13*(3), 5574-5586. doi:10.14687/jhs.v13i3.4155

Yelboğa, A. (2008). The assessment of reliability with Generalizability Theory: An application in ındustrial and organizational psychology. *Studies in Psychology, 28*, 35-54.

Yelboğa, A. (2012). Dependability of job performance ratings according to Generalizability Theory. *Education and Science, 37*(163), 157-164.

Yelboğa, A. (2015). Estimation of generalizability coefficient: An application with different programs. *Archives of Current Research International*, *2*(1), 46-53.