Education and Science tedmem

Vol 40 (2015) No 177 285-296

Investigation Of Reliability In Generalizability Theory With Different Designs On Performance-Based Assessment *

Serap Büyükkıdık [1], Duygu Anıl [2]

Abstract

In this research, it has been analyzed how variation in facet number affects reliability with the testing of reliability in generalizability theory by using different designs. The research data have been accessed with the scoring of performances towards non-routine problem solving of 132 6th, 7th and 8th grade students of a primary school in Kütahya in 2011-2012 spring term. In the research, p x t x r and p x t x r x a designs have been used in which (p=person) as a measurement object, and task, rater (t=task, r=rater) and rubric (a=rubric) have been seen as variation sources. The research results show that designs used in generalizability theory affect G and phi coefficients; as the number of source of variability increases, percentage of the description of total variance of the person which is the aim of testing decreases. Also, it has been found that the sort of rubric will affect reliability in testing, scores taken from analytical rubric have more reliability than the ones taken from holistic rubric.

Keywords

Performance-based assessment
Generalizability theory
Interrater realibility
Analytical rubric
Holistic rubric

Article Info

Received: 12.04.2012
Accepted: 10.23.2014
Online Published: 02.15.2015

DOI: 10.15390/EB.2015.2454

Introduction

Whatever the nature of assessment, flexible answering facilities should be aimed rather than a single correct answer; so, it is intented to suggest a shift from machine scored tests to the use of tasks scored by human judge, requiring students to construct response (Linn, & Miller, 2004: p. 6). Multiple choice tests that is frequently used at schools for an objective and a quick scoring and for its reflecting the content better by ranking or giving place to more questions, can be inadequate in evaluating high-level behaviors. Performance based assessment is one of the methods that can be used for resolving this limitation. In performance based assessment, scoring method and reliability of raters are important, because scoring convictions are engaged. For testing the reliability, there are various theories and applications. Differentation of theories and designs in the application of theories results in the differentiation of reliability coefficients and obtaining different information from applications.

* This study is comprised from a part of master thesis titled "Büyükkıdık, S.(2012). Comparison Of Interrater Reliability Based On The Classical Test Theory And Generalizability Theory In Problem Solving Skills Assessment"

[1] Gazi University, Faculty of Education, Educational Sciences, Department of Measurement and Evaluation, Turkey, sbuyukkidik@gmail.com

[2] Hacettepe University, Faculty of Education, Educational Sciences, Department of Measurement and Evaluation, Turkey, aduygu@hacettepe.edu.tr

In this context, reliability of scores obtained from analytical and holistic rubrics used as scoring method in performance based assessment will be compared according to the information obtained by the use of different designs in generalizability theory.

Performance Based Assessment

Performance based assessment is described as a state identification method which students show particular knowledge, concepts and skills with the tasks reflecting 'real-life contents and conditions (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999). The way in which they were described was appealing, in that, performance assessment require students to perform an activity (e.g., build a model) or construct an original response (e.g., explain one's solution to a mathematics problem); assess higher-level thinking and problem solving skills; require students to apply their problem solving in relatively novel real-world situations; afford multiple solutions or strategies; access prior knowledge; and require extended preiods of time, ranging from several minutes to days or more (Aschbacher, 1991; Baron, 1991; Herman, Aschbacher, & Winters, 1992; Madaus & O'Dwyer, 1999; Stiggins, 1987; As cited by: Lane & Stone, 2006: p. 387).

Rubric

The original meaning of rubric had little to do with the scoring of students' work. The Oxford English Dictionary tells us that in the mid-15th century, rubric referred to headings of different sections of a book. This stemmed from the work of Christian monks who painstakingly reproduced sacred literature, invariably initiating each major section of a copied book with a large red letter. Because the Latin word for red is ruber, rubric came to signify the headings for major divisions of a book. A couple of decades ago, rubric began to take on a new meaning among educators (Popham, 1997: p. 72). According to Goodrich (1997: p.14), a rubric is a scoring tool that lists the criteria for a piece of work, or "what counts" (for example, purpose, organization, details, voice, and mechanics are often what count in a piece of writing); it also articulates gradations of quality for each criterion, from excellent to poor. Where and when a scoring rubric is used does not depend on the grade level or subject, but rather on the purpose of the assessment (Moskal, 2000). Rubric is also divided into two parts as analytical rubric that evaluates the performance by separating into parts and holistic rubric that focuses on the whole performance (Mertler, 2001).

Holistic Rubric

Holistic word derives from the Greek word 'holos' and means 'whole, entire'. Holistic methods are also known as 'overall impression' or 'all impression' (Priestley, 1982: p. 203). A holistic rubric requires the teacher to score the overall process or product as a whole, without judging the component parts separately (Nitko, 2001; As cited by Mertler, 2001). It requires the product or performance to be evaluated as a whole without judgement of the components of the performance or the resulting product (Moskal, 2000).

Analytic Rubric

Analytical rubrics include the scoring of the parts separately and calculation by adding these person scores (Moskal, 2000).

Generalizability Theory

Generalizability theory is a variance analysis (ANOVA) based statistical theory put forward by Cronbach, Gleser, Nanda and Rajaratnam (1963-1972) on reactions to the limitations of true score model of classical test theory which is still used today and developed by the studies of Shavelson and Webb (1991),Brennan (1992) and lastly Brennan (2001a) which provides the evaluation of reliability in behavioral measures; design and research of reliable observations with G (generalizability) and D (decision) studies and determination of the amount discrepancy resource in observed scores with a single coefficient (Shavelson & Webb, 1991; Brennan, 2001a, 2001b).

Generalizability Study

In G theory, a coefficient is calculated called generalizability coefficient. This coefficient doesn't reinterpret the concept of reliability although similar to the coefficient in classical test theory. G theory also shows how the traditional distinction between reliability and validity can be eliminated by organizing reliable observations. In G theory a universe, its variability sources and conditions of observations are described from the structure explained in the traditional field of validity concept. If G theory provides to show the predictions accurately which are about implicit structure of observations (acceptable observations universe), it defines the observations as reliable (Shavelson & Webb, 1991). The purpose of a G study is to obtain estimates of variance components associated with a universe of admissible observations (Brennan, 2001a: p. 8).

Decision Study

Generalizability theory seperates decision (D) study from generalizability (G) study. D study is regulated to investigate the ways to minimize the errors in measurement made for a specific purpose by using the information obtained from G study (Crocker & AIgina 1986; Shavelson & Webb 1991; Brennan, 2001a).

Perhaps the most important D study consideration is the specification of a universe of generalization, which is the universe to which a decision-maker wants to generalize based on the results of a particular measurement procedure (Brennan, 2001a: p. 9).

In Generalizability theory, there are two kinds of designs: crossed or nested as well as making up designs dependent on the number of variability resource. If all conditions of variability resources in measurement affect all conditions of an another variability resource, it is crossed and is shown by placing the "x" mark between variability resources. If some conditions of a variability resource is observed by some conditions of an another variability resource, it is nested and is shown by placing ":" between the variability resources (Shavelson & Webb, 1991; Brennan, 2001a; Mushquash & O'Connor, 2006). Person is usually not called facet, that is, the possible source of measurement error because the aim of measurement in this study is person.

The aim of research

In this research, it is aimed to examine the effect of different design application on reliability. For this purpose the following questions have been sought.

1. In points obtained from analytical and holistic rubrics, how are the explanation percentages of variance components and total variance of designs p x t x r- in which person measurement object, task and rater variability sources are crossed- and p x t x r x a - in which rubric variability source is handled- estimated in the result of G study?
2. In points obtained from analytical and holistic rubrics, how are G and phi coefficients obtained in the result of decision study made by increasing and decreasing the numbers of rater and task of designs p x t x r and p x t x r x a in which key variability source is handled?
3. Do findings obtained from p x t x r and p x t x r x a designs used in realibility evaluation differ?

The Importance of research

There are many studies on the evaluation of reliability and generalizability of measurements obtained in literature with generalizability theory (Al-Mahroos, 2009; Arce-Ferrer ve Castillo, 2007; Atılgan, 2004; Christ et. al., 2010; Deliceoğlu, 2009; Eser, 2011; Güler, 2008; 2009; 2011; Hoyt ve Melby, 1999; Jarjoura et al, 2004; Kaya, 2011; Kozaki, 2004; Nalbantoğlu, 2009; 2012; Öztürk, 2011; Taşdelen ve diğerleri, 2010; Tindal et. al, 2010; Van Hooft et. al., 2006; Yelboğa, 2007; 2012).

When all these studies are examined, it hasn't been encountered to a similar research in which different rubrics are used and fully crossed designs-rubric is taken as the variability source-are applied in generalizability theory and findings are compared. With this study, it is thought to contribute literature by comparing findings obtained from two faced design in which person, task and rater variability sources are fully crossed and three faced fully crossed design in which rubric is also taken into consideration beside these variability sources.

Method

The Type of Research

It is a descriptive research because it reveals existing condition by analyzing the effect of different design application on reliability in generalizability theory.

Collection of Data

The data of research are collected by scoring the performances of 132 6th, 7th and 8th grade students in two performance tasks prepared for non-routine problem solving skill with recognition response codes and analytical rubric and then with holistic rubric ten days later by four maths teacher raters in a centre elementary school in Kütahya in 2011-2012 education year.

As Polya (1973: 171) states that in general, a problem is a "routine problem" ıf it can be solved either by substituting special data into a formerly solved general problem, or by following step by step, without any trace of originality, some well-worn conspicuous example. Performance tasks used in research include non-routine problems that have not one true solution and response formed by scanning the literature by the researcher. Whether the performance tasks used in research are suitable for student level and the measurement of problem solving skill has been decided in consultation with ten people including elementary math teachers and mathematics education and assessment experts. Likewise, expert view has been taken from these experts for rubrics. To determine the conformity of performance tasks and rubrics to Turkish, a Turkish Language and Literature teacher is consulted.

Raters have examined each paper with content analysis by using behavior recognition codes first after training by the researcher on how scoring should be. Then, raters have carried out scorings in series for each performance with task-specific analytical rubric which includes understanding the problem, identifying solution ways and application, criteria to specify the solution and five levels. After remember time, that is, ten days have passed; same procedure has been carried out with holistic rubric.

Analysis of Data

Reliability analysis has been done by using designs p x t x r and fully crossed p x t x r x a in which kind of key is taken as variability source for analytical and holistic rubrics in generalizability theory. In the fist step, variance values have been estimated in pxtxr design for main and common effects by performing G study for both rubrics in generalizability theory. In the second step, variance values analysis has been done for main and common effects by performing G study in pxtxrxa design within generalizability theory. In the third step, G and phi coefficients have been estimated by performing a decision study for the same design in the event of an(one) increase and decrease of the number of rater and duties. In the last step, it has been examined how values obtained from both designs used affect reliability. EduG 6.0 program is utilized in the analysis of data.

Results

G Study Results

Generalizability study has been done to the scores obtained from scoring of two performances that 132 elementary second grade students showed by four raters with analytical and holistic rubrics with designs p x t x r and p x t x r x a in G theory. Estimated variance components and total variance percentages of each variance source are given in Table 1.

Table 1. Variance Components And Total Variance Percentages Estimated İn The Result Of G Study Belonging To p x t x r and p x t x r x a Designs

design		**pxtxr** *Analytic Rubric*		**pxtxr** *Holistic Rubric*		**pxtxrxa**	
Variance source	**df**	**Variance σ^2**	**%**	**Variance σ^2**	**%**	**Variance σ^2**	**%**
p	131	15.423	80.6	14.836	76.0	14.149	72.5
t	1	0.107	0.6	0.243	1.2	0.084	0.4
r	3	0.078	0.4	0.310	1.6	0.274	1.4
a	1	-	-	-	-	0.034	0.2
pt	131	0.004	0.0	0.000	0.0	0.000	0.0
pr	393	2.344	12.2	2.191	11.2	0.074	0.4
pa	131	-	-	-	-	2.946	15.1
tr	3	0.025	0.1	0.042	0.2	0.000	0.0
ta	1	-	-	-	-	0.016	0.1
ra	3	-	-	-	-	0.000	0.0
ptr	393	1.154	6.1	1.901	9.8	0.017	0.1
pta	131	-	-	-	-	0.388	2.0
pra	393	-	-	-	-	0.416	2.1
tra	3	-	-	-	-	0.042	0.2
ptra,e	393	-	-	-	-	1.072	5.5
Toplam	2111		100		100		100

p: person, t: task, r: rater, a: rubric

When variance and total variance percentages- estimated in the result of G study of p x t x r design- in data obtained by using analytical rubric in Table 1 are examined, the biggest variance component is seen as person main effect that differs in problem solving skills with σ_p^2 (15.423) variance component value and % 80.6 total variance percentage. This can be the indication to the reflection of skill differentations caused by person which is aimed at performance based assessment and to a heterogeneous group in terms of problem solving skill. Person x task common interaction is a variability source as a result of effect of the task unwanted to be measured on person effect wanted to be measured. That person didn't differ in one task to another is seen with σ_{pt}^2(0.004) variance rate and % 0.0 total variance percentage. That raters aren't generous relative to each other while evaluating the performance of person is seen with σ_r^2(0.078) variance component value and % 0.4 total variance percentage. It is seen that different raters make similar scorings. Person x task x rater interaction (bgp) $\sigma_{ptr,e}^2$ is a residual variability source resulting from measurement error. When Table 1 is analyzed, it is understood with $\sigma_{ptr,e}^2$(1.154) variance component value and % 6.1 total variance percentage that after person (object of measurement) and p x t interaction, this variability source also called random error is the biggest variability source.

When variance and total variance percentages are analyzed which are estimated in the result of G study of pxtxr design in data obtained with holistic rubric; it is seen that variance component of person main effect explains % 76.0 of total variance mostly and person x task common interaction doesn't contribute to total variance with % 0.0 value. It is seen with σ_t^2(0.243) variance component value and % 1.2 total variance percentage that both tasks are in the same difficulty. It is seen that raters aren't more generous relative to each other while evaluating the performance of person with σ_r^2(0.310) variance component value and % 1.6 total variance percentage. It is understood that different raters do similar scorings. It is seen with σ_{pr}^2(2.191) variance component value and % 11.2 total variance that conditions of person differ partly from one rater to the other. It is understood with $\sigma_{ptr,e}^2$(1.901) variance component value and % 9.8 total variance explanation percentage that after the object of measurement the person and pxr interaction, the third biggest variability source is the variability source called random error.

When variance and total variance percentages are analyzed estimated in the result of G study of b x g x p x a design in Table 1, it is seen that variance component of person main effect explains % 72.5 of total variance mostly whereas person x task, task x rater, rater x rubric common interaction doesn't contribute to the total variance with %0.0 value at the very least. When the variance component estimated in G theory is of negative value, Cronbach and others suggest that negative variance component should be taken as zero to calculate variance components (Brennan, 2001a). In researchs performed in Turkey, zero value is also written instead of negative variance component (Atılgan, 2004; Taşdelen et al., 2010). So; while person x task common effect variance component is σ_{pt}^2(-0.136) , it is accepted as σ_{pt}^2(0.00) in Table 1. Likewise; while the variance components of task x rater common effect and rater x rubric common effect are σ_{tr}^2(-0.001) and σ_{ra}^2(-0.004) respectively (negative), they are accepted as σ_{tr}^2(0.00) and σ_{ra}^2(0.00) in Table 1.

D Study Results

Generalizability coefficient, phi coefficient, relative error variance and absolute error variance are given in Table 2 estimated for scenarios done by determination of person as object of measurement and increase/decrease of task and rater numbers in designs p x t x r in which all variances are crossed and pxtxrxa in which kind of key is taken as variance source in scores obtained from analytical and holistic rubrics.

Table 2. Decision Study Results Done With An Increase/Decrease Of Raters and Task Numbers In p x t x r and p x t x r x a Designs

		Analytic Rubric		*Holistic Rubric*			
		p x t x r		**p x t x r**		**p x t x r x a**	
nt	**nr**	**G**	**Phi**	**G**	**Phi**	**G**	**Phi**
1	2	0.897	0.889	0.878	0.857	0.871	0.857
1	4	0.946	0.938	0.935	0.916	0.882	0.872
1	6	0.963	0.955	0.955	0.937	0.886	0.877
2	2	0.913	0.908	0.904	0.888	0.884	0.873
2	3	0.940	0.935	0.934	0.920	0.889	0.880
2	**4**	**0.954**	**0.950**	**0.949**	**0.937**	**0.892**	**0.884**
2	5	0.965	0.962	0.959	0.947	0.893	0.886
2	6	0.969	0.965	0.965	0.954	0.894	0.888
3	2	0.918	0.914	0.913	0.899	0.888	0.878
3	4	0.957	0.954	0.954	0.944	0.895	0.888
3	6	0.971	0.968	0.969	0.960	0.897	0.892

nt: number of tasks, nr: number of raters

It is estimated that G coefficient is 0.954 and phi coefficient is 0.950 of the scores obtained with analytical rubric in the direction of two tasks, each 132 people, by four raters in pxtxr design used in the research. When Table 2 is analyzed, it is seen that two increase/decrease in the number of rater has more effect than one increase/decrease in the number of task on G and phi coefficients. Also, it is seen that increasing the number of task and rater increases G and phi coefficients and vice versa. As seen in Table 2, when the number of task and rater is the least (n_r = 2, n_t = 1); G coefficient is 0.897, phi coefficient is 0.889 and it takes the lowest reliability values in decision study. When the number of task and rater is the most (n_r = 6, n_t = 3); G coefficient is 0.971, phi coefficient is 0.968 and it takes the highest reliability values in decision study.

In points obtained from holistic rubric in p x t x r design, it is estimated that G coefficient is 0.949 and phi coefficient is 0.937. When Table 2 is analyzed, it is seen that two increase/decrease in the number of rater has more effect than one increase/decrease in the number of task on G and phi coefficients. Also it is seen that increasing the number of task and rater increases G and phi coefficients and vice versa. When the number of task and rater is the least (n_r = 2, n_t = 1); G coefficient is 0.878, phi coefficient is 0.857 and it takes the lowest reliability values in decision study. When the number of task and rater is the most (n_r = 6, n_t = 3); G coefficient is 0.969 and phi coefficient is 0.960 and it takes the highest reliability values in decision study.

It is estimated that G coefficient is 0.892 and phi coefficient is 0.884 of the scores obtained with analytical and holistic rubrics in the direction of two tasks, each 132 persons, by four raters in pxtxrxa design in the research. When Table 2 is analyzed, it is again seen that increasing the number of task and rater increases G and phi coefficients and vice versa. When the number of task and rater is the least (n_r = 2, n_t = 1); G coefficient is 0.871 and phi coefficient is 0.857 and it takes the lowest reliability values in decision study. When the number of task and rater is the most (n_r = 6, n_t = 3); G coefficient is 0.897 and phi coefficient is 0.871 and it takes the highest reliability values in decision study.

Comparison of Findings Obtained from Each Two Design

When total variance percentages are analyzed in result of G study of p x t x r and bxgxpxa designs; it is seen that person main effect has the most variance percentage with % 80.6 total variance percentage for the data obtained from analytical rubric, % 76 total variance percentage for the data obtained from holistic rubric,% 72.5 total variance percentage when the key is the variability source. It is seen that person-object of measurement- total variance percentage decreases as the number of variability source increases. Also, it is seen that the usage of analytical rubric increases total variance percentage of measurement object in comparison with holistic rubric.

When person x task x rater common effect is analyzed in every two design used; it is seen that residual variability source dependent on measurement error has % 6.1 total variance percentage in data obtained with analytical rubric, % 9.8 total variance percentage in data obtained with holistic rubric, % 0.1 total variance percentage in p x t x r x a design when the key is variability source. In obtaining these findings; it is thought to be effective that kind of key is the variability source in p x t x r x a design and other variability sources share the effect on total variance. It is seen with % 15.1 total variance percentage that the conditions of persons differs partly from one key to the other.

When the decision study findings are analyzed; G and phi coefficients in data obtained with analytical rubric are estimated as G coefficient is 0.954, phi coefficient is 0.950, G coefficient is 0.949, phi coefficient is 0.937 respectively.

Discussion, Conclusion and Suggestions

When G study is done with fully crossed p x t x r design on scores obtained with analytic rubric, more reliable results have been achieved in comparison with scores obtained with holistic rubric. While residual effect also called random error is lower in scores obtained with analytical rubric, the effect caused by the difference in problem solving skills of people (aim of measurement) is higher. These results overlap with many studies in which scores obtained from analytical rubric show relatively higher reliability in comparison with scores obtained from holistic rubric in classical test theory (Bauer, 1981; Follman & Anderson, 1967; Jonsson & Svingby, 2007) and overlap partly with findings of studies that reach to the conclusion that scores obtained from analytical rubric are substantially more reliable in comparison with scores obtained from holistic rubric (Boring, 2002; Klein et. al, 1998).

When K study is done in the design used (p x t x r), it is again seen that scores obtained from analytical rubric have higher G and phi coefficients that the ones obtained fron holistic rubric.

It is seen that G and phi coefficients also increase partly when the number of task and rater is increased in both rubrics, but increasing the number of rater has slightly more effect than increasing the number of task in increasing coefficients.

As in the other design used (p x t x r), when the number of task and rater is also increased in bxgxpxa design; G and phi coefficients increase partly, but increasing the number of task has slightly more effect than increasing the number of rater in increasing coefficients by contrast with findings obtained from decision study in pxtxr design. Also, it is understood from decision study in p x t x r and p x t x r x a designs that increasing the number of variability source in the design used lowers G and phi coefficients relatively. In this instance, it is thought that it is because total variance is shared between more variability sources.

Consequently, practitioners can use behavior recognition codes and task- specific analytical rubric by taking the aim of measurement and thinking styles of students into account to increase rater reliability in performance based assessment. Researchers can carry out reliability and generalizability study with designs in which different variability sources are handled in generalizability theory by studying with a bigger research group, rater and task. In performance-based assessment, when interrater consistency is low, reliability studies can be carried out together with qualitative studies to question the reason of this situation.

References

Al-Mahroos, F. (2009). Construct validity and generalizability of pediatrics clerkship evaluation at a problem-based medical school, Bahrain. *Evaluation & the Health Professions, 32*(2), 165-183.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

Arce-Ferrer, A. J., & Castillo, I. B. (2007). Investigating postgraduate college admission interviews: generalizability theory reliability and incremental predictive validity. *Journal of Hispanic Higher Education, 6(2)*, 118-134.

Atılgan, H. (2004). *Genellenebilirlik kuramı ve çok değişkenlik kaynaklı rasch modelinin karşılaştırılmasına ilişkin bir araştırma.* Unpublished Phd thesis, Hacettepe Üniversitesi, Ankara.

Bauer, B. A. (1981). *A study of the reliabilities and cost-efficiencies of three methods of assessment for writing ability. (*ERIC Document Reproduction Service No. ED 216357).

Boring, R. L. (2002). *Human and computerized essay assessment: a comparative analysis of holistic, analytic and latent semantic methods*. Unpublished thesis, Department of Psychology, New Mexico State University, Las Cruses, New Mexico.

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

Brennan, R. L. (2001b). An essay on the history and future of reliability from the perspective of replications. *Journal of Educational Measurement, 38(*4), 295-317.

Crocker, L. M., & Algina, L. (1986). *Introduction to classical and modern test theory.* New York: Holt, Rinehart and Winston.

Christ, T. J., Tillman C., Chafouleas, S. M., & Boice C. H. (2010). Direct behavior rating (DBR): generalizability and dependability across raters and observations. *Educational and Psychological Measurement, 70*(5), 825-843.

Çakıcı, D. (2011). *Genellenebilirlik kuramı ve lojistik regresyona dayalı hesaplanan puanlayıcılar arası tutarlığın karşılaştırılması.* Unpublished master thesis, Hacettepe Üniversitesi, Ankara.

Deliceoğlu, G. (2009). *Futbol yetilerine ilişkin dereceleme ölçeğinin genellenebilirlik ve klasik test kuramına dayalı güvenirliklerinin karşılaştırılması.* Unpublished Phd thesis, Ankara Üniversitesi, Ankara.

Follman, J. C., & Anderson, J. A. (1967). An ınvestigation of reliability of five procedures for grading english themes. *Research in the Teaching of English, 1(2)*, 190-200.

Goodrich, H. (1997). Understanding rubrics. *Educational Leadership, 54*(4), 14-17.

Güler, N. (2008). *Klasik test kuramı genellenebilirlik kuramı ve rasch modeli üzerine bir araştırma.* Unpublished Phd thesis, Hacettepe Üniversitesi, Ankara.

Güler, N. (2009). Genellenebilirlik kuramı ve SPSS ile GENOVA programlarıyla hesaplanan G ve K çalışmalarına ilişkin sonuçların karşılaştırılması. *Eğitim ve Bilim, 34*(154), 93-103.

Güler, N. (2011). Rastgele veriler üzerinde genellenebilirlik kuramı ve klasik test kuramı'na göre güvenirliğin incelenmesi. *Eğitim ve Bilim, 36*(162), 225-234.

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

Jarjoura, D., Early, L., & Androulakakis, V. (2004). A multivariate generalizability model for clinical skills assessments. *Educational and Psychological Measurement, 64(1),* 22-39.

Jonsson, A., & Svingby, G. (2007). The use of scoring rubrics: reliability, validity and educational consequences. *Educational Research Review. 2(2)*, 130-144.

Kaya, G. (2011). *Genellenebilirlik kuramının doldurma kavram haritası değerlendirme çalışmasına uygulanması.* Yayımlanmamış yüksek lisans tezi, Hacettepe Üniversitesi, Ankara.

Kozaki, Y. (2004). Using GENOVA and SPSS to set multiple standards on from Japanese into English performance assessment for certification in medical translation. *Language Testing, 21*(1), 1-27.

Klein, S. P., Stecher, B. M., Shavelson, R. J., McCaffrey, D., Ormseth, T., Bell, R. M., Comfort, K., & Othman, A. R. (1998). Analytic versus holistic scoring of science performance tasks. *Applied Measurement in Education, 11(2),* 121-137.

Lane, S., & Stone, C. A. (2006). *Performance assessment.* Brennan, R.L. (Ed.) Educational Measurement (4th ed.). 387-431. Westport, CT: Praeger Puublishers.

Linn, R. L., & Miller D. M. (2004). *Measurement and assesment in teaching.* (9th edition). Upper Saddle River: Printice-Hall Inc.

Martin J. Bergee. (2007). Performer, rater, occasion, and sequence as sources of variability in music performance assessment. *Journal of Research in Music Education, 55*(4), 344-358.

Mertler, C. A. (2001). Designing scoring rubrics for your classroom. *Practical Assessment, Research & Evaluation, 7(*25). Retrieved 1.11.2011 http://PAREonline.net/getvn.asp?v=7&n=25

Moskal, B. M. (2000). Scoring rubrics: what, when and how?. *Practical Assessment, Research & Evaluation, 7(*3). Retrieved 30.10.2011 http://PAREonline.net/getvn.asp?v=7&n=3

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

Nalbantoğlu, F. (2009). *Performans ölçümlerinde genellenebilirlik kuramıyla farklı desenlerin karşılaştırılması.* Unpublished master thesis, Hacettepe Üniversitesi, Ankara.

Nalbantoğlu, F. (2012*). Genellenebilirlik Kuramında Dengelenmiş Ve Dengelenmemiş Desenlerin Karşılaştırılması -Intramuskuler Enjeksiyon Yapma İstasyon Verileri Üzerine Bir Uygulama-.* Unpublished Phd thesis, Ankara Üniversitesi, Ankara.

Öztürk, M. E. (2011). *Voleybol becerileri gözlem formu ile elde edilen puanların genellenebilirlik ve klasik test kuramı'na göre karşılaştırılması.* Unpublished master thesis, Hacettepe Üniversitesi, Ankara.

Polya, G. (1973). *How to sove it? A new aspect of mathematical method.* Second edition. Princeton University press. New Jersey.

Popham, J. W. (1997). What's wrong and what's right with rubric. *Educational Leadership, 55*(2), 72-75.

Priestley, M. (1982). *Performance assessment in education and training: alternative techniques.* First edition. Educational Technology Publications. New Jersey.

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

Taşdelen, G., Kelecioğlu, H., & Güler, N. (2010). Nedelsky ve angoff standart belirleme yöntemleri ile elde edilen kesme puanlarının genellenebilirlik kuramı ile karşılaştırılması. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi. 1*(1), 22-28.

Tindal, G., Yovanoff, P., & Geller, J.P. (2010). Generalizability theory applied to reading assessments for students with significant cognitive disabilities. *The Journal of Special Education, 44*(1), 3-17.

Van Hooft, E. J. A., Born, M., Taris, T. W., & Van Der Flier, H. (2006). The cross-cultural generalizability of the theory of planned behavior: a study on job seeking in the Netherlands. *Journal of Cross-Cultural Psychology, 37*(2), 127-135.

Yelboğa, A. (2007). *Klasik test kuramı ve genellenebilirlik kuramına göre güvenirliğin bir iş performansı ölçeği üzerinde incelenmesi.* Unpublished Phd thesis, Ankara Üniversitesi, Ankara.

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