Journal of Human Sciences
ISSN: 2458-9489
Volume: 13 Issue: 3 Year: 2016

Using generalizability theory to investigate the reliability of peer assessment[1]

Gülşen Taşdelen Teker[2]
Melek Gülşah Şahin[3]
Kemal Baytemir[4]

Abstract

In this study, the effectiveness of peer assessment, which has an important role in measurement and evaluation, was attempted to be defined. For this purpose, performance task, which is one of the alternative assessment techniques, was evaluated with the help of a scoring rubric prepared by the researchers. As a basic research, the working group was 41 sophomore students and their instructor. Three of 41 students were acted as rater and they rated their 38 peers' performances with the instructor. The analysis of the data was carried out by using fully crossed two-facet design (sxtxr) of generalizability theory in three steps: G-studies for peer and peers-instructor ratings and D-study for peer ratings. According to the results of the G studies, the reliability coefficient obtained from the peer ratings and peer-instructor ratings were quite high (0.86 and 0.82 respectively). According to the result of the D study of peer ratings, just two peer raters are enough for getting high reliability coefficient. With the help of the gained results, it is suggested that peer assessment, which is effective on learning and decision making processes of students, should be used more often in education systems.

Keywords: Peer assessment; generalizability theory; reliability; peer rater.

1. Introduction

Rapid changes in science and technology have affected different aspects of societies such as social, political, economic, cultural, and educational structures. For this reason, the present education system must raise manpower that can both adapts to changes in science and technology and also possesses qualifications that the age requires. While cognitive proficiency, which has an important role in education system, was predominant previously in learning-teaching process, the rapid changes in technology also influenced the individuals' status of accessing, using and transferring information. In this process, where traditional methods fall behind, the importance and usage of process-oriented supplementary assessment methods in which higher level skills are at the forefront are gradually increasing. Reynolds, Livingston and Willson (2009) indicated that the most striking difference between traditional and alternative assessment methods is the level of real life situations included.

Many problems experienced today are because of the individuals' fail at looking at objectively to themselves, to others, to events and phenomenon around them. This situation,

[1] Part of this study was presented in Measurement and Evaluation in Education and Psychology Congress in 19-21 September 2012 in Abant İzzet Baysal University, Bolu, Turkey.

[2] Ph.D., Sakarya University, Faculty of Education, gtasdelen@sakarya.edu.tr
[3] Ph.D., Gazi University, Faculty of Education, melekgulsah@gmail.com
[4] Assistant Professor, Amasya University, Faculty of Education, kemalbaytemir@gmail.com

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

arising from individuals' lack of recognition of themselves, may reach to a point that could affect both themselves and others in a negative way (Kutlu, Dogan, & Karakaya, 2009). Because of this, the participation of students in the assessment process has been given much more importance in recent years, for it not only gives them a chance to evaluate others objectively, but also a chance to let them get to know themselves better. Peer assessment has been addressed within the scope of this study by means of its importance on development of students' sense of responsibility, use critical thinking skills and opinion about other peers' learning.

Peer assessment can be defined as an activity done with the purpose of individuals to consider the value, worth, quality or success of learning outcomes of their peers, who are at the similar status/degree (Topping 1998; Topping, Smith, Swanson and Eliot, 2000; Topping, 2009). In other words, peer assessment is a technique in which individuals assess each other according to some specified criteria and which requires students to use their knowledge and skills to review, clarify and correct their peers' works (Topping, 1998; McDowell, 1995). Such an approach would seem to offer large both staff teaching classes and their students' significant time and learning benefits-marking time decreases together with a decrease in feedback time and an increase in the quality and quantity of comments (Falchikov, 1998; Topping et al., 2000).

Peer assessment also promotes the acquisition of life-long learning skills due to the active involvement of students in the assessment experience (Ballantyne, Hughes & Mylonas, 2002). Moreover, it is often claimed that peer assessment encourages students to become critical, independent learners as they become more familiar with the application of assessment criteria and develop a clearer concept of the topic being reviewed (Falchikov, 1995; Searby & Ewers, 1997). Peer assessment itself has additional benefits. Falchikov (1986) reports increased student responsibility and autonomy as a result of the scheme of peer assessment; and her students found it challenging, helpful and beneficial, making them think more, learn more, and become more critical and structured.

When students doing peer assessment have insufficient information about the study process they may do superficial or inadequate evaluations (Kutlu, Dogan, & Karakaya, 2009). Due to this complication, scoring rubrics are used in peer assessment that were prepared in accordance with some specific standards and those let students to see what their peers did, and how much and where the mistakes and deficiencies were. While Popham (1997) stated it as a scoring tool that lists the criteria for student work and shows what can be done in that work, Goodrich Andrade (2001) defined it as a scoring guide which is used to determine and monitor the situation of students. Scoring rubrics describe the various aspects of a task, inform students about the degree of mastery required for each level of the task, and highlight the criteria upon which they will be graded on (Reed & Burton, 1985; Luft, 1997; Popham, 1997; Hafner & Hafner, 2007).

There are many studies evaluate the reliability of peer assessment by using different statistical techniques. For instance Hughes and Large (1993a, 1993b) found that peer marks or grades reported acceptably high reliability, often expressed in correlation coefficients, percentage agreement, or measures of central tendency and variance, sometimes with indication of statistical significance (Topping, 1998). Han, Mun and Ahn (2009) and Strang (2013) used Kappa coefficient and Alfallay (2003) used Pearson correlation coefficient to investigate interrater reliability of peers acted as raters while using scoring rubrics. Sadde and Good (2006) used Kappa, percentage of agreement and correlation coefficient together. Moreover there are many studies used many facet Rasch model to investigate reliability of peer assessment (Baştürk, 2008; Semerci 2011a, 2011b; Karakaya, 2015; Yüzüak, Yüzüak and Kaptan, 2015; Esfandiari, 2015; Aryadoust, 2016; Şahin, Taşdelen Teker, Güler, 2016). Besides these techniques there is another effective way of investigating not only the reliability of peer assessment but also the appropriate number of rater by means of peer: Generalizability theory.

Generalizability (G) theory is a framework for analyzing how well observed scores allows users to make generalizations about a person's behavior (Shavelson & Webb, 1991). Instead of partitioning an observed score into just two parts as true score and error score without

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

differentiating the various sources that contribute to the error score is a limitation of classical test theory (CTT) (Güler, 2009; Baykul, 2000), G theory partitions the error variance into multiple components representing several different sources of error simultaneously and shows the influence of each. Therefore, G theory can be viewed as an extension of CTT in that G theory involves separating out various sources of error (Brennan, 1992; Brennan, 2001; Shavelson & Webb, 1991; Cronbach, Gleser, Nanda, & Rajaratnam, 1972).

Another advantage of using G theory is that it can estimate the reliability of the mean rating for each examinee, while simultaneously accounting for both interrater and intrarater inconsistencies as well as discrepancies due to various possible interactions which are impossible in CTT. Since, classical reliability procedures do not allow the researcher to simultaneously estimate the amount of measurement error from multiple sources (Brennan, 2001).

Each source of variation such as the items, raters, or different measurement situations available in the measurement process in the G theory is called a *facet*. Facet can be interpreted as the measurement situations having similarities (Brennan, 2001). Each level on the facets is referred to as a condition. For instance, in the process of a 10-item test, items constitute a facet, and each item is one condition of the facet (Brennan, 2001; Shavelson & Webb, 1991). The source revealing the variability of concern (students, items etc.) is called the object of measurement constituting the real, systematic variability, rather than being called the source of variation (Musquash and O'Connor, 2006). In this study, the object of measurement is students (*s*) and the two facets are tasks (*t*) and raters (*r*).

There are two different studies in G theory: Generalizability (G)-study and Decision (D)-study. A G-study is done to determine how well the scores can be used for multiple situations. Therefore, the concern in a G study is the generalizability of the obtained results. A G-study involves estimating variance components that might in turn be used in a D-study for computing generalizability coefficients. On the other hand, D-study is conducted for the purpose of determining the most efficient measurement procedure for a given situation. Although there are only relative decisions made in CTT, there are two different types of decisions as relative and absolute since there are two different types of error variance as relative and absolute in G theory (Yin and Shavelson, 2008; Brennan, 2001; Brennan, 1992; Shavelson & Webb, 1991). The relative error variance of G-theory, which is used in relative decisions, can be thought of as an analog to the error variance of CTT (Lee & Frisbie, 1999).

Since there are two types of error variances, there are also two coefficients of reliability as generalizability (G) and dependability (Phi). The two have a similar structure that is analogous to the structure of the reliability coefficient in classical test theory (Crocker & Algina, 1986). The difference between the two coefficients is based on the definition of what constitutes error for the type of decision to be made.

When investigated the studies in the literature based on the reliability of peer assessment via G theory, there are some studies. For instance, Donnon, McIlwrick and Woloschuk (2013) investigated the reliability and validity of self and peer assessment. They used G theory to determine the optimal number of peer assessors required to obtain a generalizability coefficient of greater than 0.70. In other words they conducted a D-study for single facet nested design. Gugiu and Gugiu (2012), employed G theory to estimate reliability of peer assessment of undergraduate research papers. The results of their study showed peer assessment was reliable. A new method for computing the minimum acceptable reliability was also introduced in this study.

Sung et al. (2010) conducted a study to determine the rating behaviors of teenagers in self- and peer assessments, and how the number of raters influences the reliability and validity of self- and peer assessments. G theory and criterion-related validity were used to obtain the reliability and validity coefficients of the self- and peer ratings. Analyses of variance were used to compare differences in self- and peer ratings between low- and high-achieving students. The coefficients of reliability and validity increased with the number of raters, reaching the acceptable levels of 0.80 and 0.70, respectively, with 3 or 4 raters. Furthermore, it was found that low- and high-achieving

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

students tended to over- and underestimate the quality of their work in self-assessment, respectively.

Marty et al. (2010) investigated the accuracy and reliability of peer assessment of athletic training students' psychomotor skills. Participants of their study evaluated ten videos of a peer performing three psychomotor skills on two separate occasions using a valid assessment tool. Accuracy of each peer assessment score was examined through percentage correct scores and they used a G study to determine how reliable athletic training students were in assessing a peer performing the aforementioned skills. Decision studies using G theory demonstrated how the peer-assessment scores were affected by the number of participants and number of occasions. As a result of the study, participants had a high percentage of correct scores. Reliability was affected by the variance of the videos. If videos were created with more variance in the displayed skill and the participants' accuracy remained high, they would expect the reliability of these assessments to increase. Moreover, according to the results of D studies, it was concluded that peer assessments must be based on multiple measurement opportunities (multiple participants, multiple occasions, or both) if the stability of the result is important.

Hafner and Hafner (2007) focused on the validity and reliability of the rubric as an assessment tool for student peer-group assessment in an effort to further explore the use and effectiveness of the rubric. A total of 1577 peer-group ratings using a rubric for an oral presentation was used in this 3-year study involving 107 college biology students. A quantitative analysis of the rubric used in this study showed that it was used consistently by both students and the instructor across the study years. Moreover, the rubric appears to be 'gender neutral' and the students' academic strength had no significant bearing on the way that they employed the rubric. A significant, one-to-one relationship between the instructor's assessment and the students' rating is seen across all years using the rubric. Moreover, the results of generalizability study yielded estimates of inter-rater reliability of moderate values across all years and allowed for the estimation of variance components.

Marcoulides and Simkin (1992) described an experiment in which term projects, a preprinted evaluation form, and generalizability theory were used to judge the reliability of student grading. The results suggested that students could be both consistent and fair in their assessments. These findings, along with mostly favorable student reactions and the fact that employee valuation was an important management skill, create a strong case for peer review when evaluating student papers.

The present study contributes to body of knowledge by providing a plausible explanation for some of the contradictory results. Namely, most of the previous reliability studies employed analytical techniques such as kappa, correlations, Cronbach's alpha. On the other hand, G theory is one of the most sophisticated method(s) for estimating reliability. Therefore, this study is important by means of contribution to the reliability of peer assessment studies in terms of usage of G theory which is practical and effective to peer assessment studies. Particularly the two-facet model (in this study, task and rater) is quite suitable for investigation of reliability of peer assessment. Since, the reliability coefficient reported herein is a cleaner measure of reliability because of controlling for two sources of measurement error simultaneously (tasks and raters). The next advantage of this study is investigation of consistency between peers and instructor assessments. By this way, the peer assessment's quality is also demonstrated by means of reliability coefficients calculated between student raters and the instructor. The last aim is to determine the most suitable number of peer as rater. The number of peers attends in peer assessment and their qualifications about evaluation process were important issues in peer assessment. Since G theory (Brennan, 2001; Cronbach, Gleser, Nanda, & Rajaratnam, 1972; Marcoulides, 1989; Shavelson & Webb, 1991; Nunally & Bernstein, 1994) can assist in the design of cost-efficient systems that produce reliable scores, it is a practical and effective way of determining the issues mentioned above.

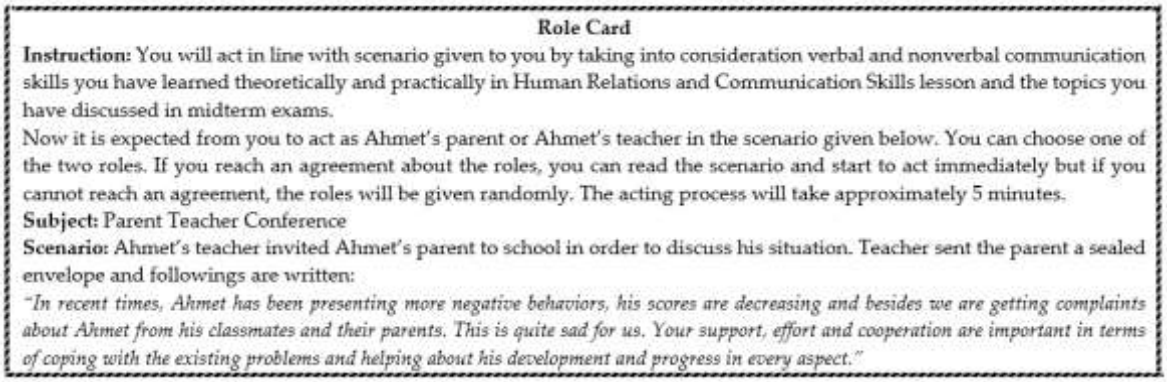
Taşdelen Teker, G., Şahin, M. G., & Baytemir, K. (2016). Using generalizability theory to investigate the reliability of peer assessment applications. *Journal of Human Sciences*, *13*(3), 5574-5586. doi:10.14687/jhs.v13i3.4155

2. Method

Participants: The participants for this study were 41 undergraduate students enrolled in a course called Human Relations and Communication from the department of Psychological Counselling and Guidance in the Faculty of Education at Amasya University in Turkey and the instructor of the course. All the students were sophomore and the instructor had five-year-experience of giving the course. 38 of 41 students were responsible for performing the scenarios given to them; the other three students and the instructor acted as raters by using the scoring rubric developed by researchers. The students who acted as rater was selected according to their academic achievement levels. To represent the all levels of students, high, medium and low academic level students were selected according to their midterm results of the course.

Since peer assessment requires students to use their knowledge and skills to correct, review and clarify other peers' work (Ballantyne et al, 2002) and the rater students of this study had little experience on peer assessment, some explanations about the process was given to them. In other words, the rater students were informed about peer assessment in detailed by the instructor.

Instrument: First of all, according to the course content, a role card, which contains main communication tasks, were prepared by the instructor. During the performances, there were two students on the stage and they randomly assigned to act as teacher or parent explained detailed on the role card given in the Figure 1. Their performances were graded by four raters: three of them are the peers and one of them is the instructor of the course.



Role Card

Instruction: You will act in line with scenario given to you by taking into consideration verbal and nonverbal communication skills you have learned theoretically and practically in Human Relations and Communication Skills lesson and the topics you have discussed in midterm exams.

Now it is expected from you to act as Ahmet's parent or Ahmet's teacher in the scenario given below. You can choose one of the two roles. If you reach an agreement about the roles, you can read the scenario and start to act immediately but if you cannot reach an agreement, the roles will be given randomly. The acting process will take approximately 5 minutes.

Subject: Parent Teacher Conference

Scenario: Ahmet's teacher invited Ahmet's parent to school in order to discuss his situation. Teacher sent the parent a sealed envelope and followings are written:

"In recent times, Ahmet has been presenting more negative behaviors, his scores are decreasing and besides we are getting complaints about Ahmet from his classmates and their parents. This is quite sad for us. Your support, effort and cooperation are important in terms of coping with the existing problems and helping about his development and progress in every aspect."

Figure 1. Role Card

A scoring rubric given in Table 1 based on the principles mentioned in the book of Human Relations and Communication (Voltan Acar, 2012) was developed by researchers to assess the quality of role plays of 2-student groups as parent and teacher. While developing the rubric, the opinions of four experts were taken. Two of the experts were from the field of guidance and psychological counselling and two of them were from measurement and evaluation in education field. Moreover, all of them had PhD degrees from their fields. According to the opinions and constructivist critics of the experts, the form was modified before application. The form broke down the eight tasks pertaining to each assignment. All performances acted by students as parent and teacher roles were scored by using the same scoring rubric which is 5 Likert type (strongly disagree to strongly agree).

Procedure: The instructor and the three peers (raters) who were assigned to evaluate all the performances were provided with the description of each assignment and its accompanying grading rubric. Before the ratings, some training was provided to peer raters by the instructor since they were not familiar with the peer assessment procedure before. This training was crucial since as

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

stated by Orsmond and Merry (1996) students were uncomfortable with peer assessment because they felt unqualified to mark others' work. Besides this finding, Falchikov (1995) and Mowl and Pain (1995) reported that the majority of their students found assigning marks to their peers' work difficult. Therefore, before the application, a brief explanation of the importance and the profits of peer assessment procedure was explained the peer raters.

Table 1. Scoring rubric

Tasks	Strongly disagree				Strongly agree
	1	2	3	4	5
1) Using the full message with its elements (perception, emotion, request)	( )	( )	( )	( )	( )
2) Making the appropriate emotion reflection	( )	( )	( )	( )	( )
3) Making the appropriate content reflection	( )	( )	( )	( )	( )
4) Exhibiting an efficient listening ability	( )	( )	( )	( )	( )
5) Using the open invitation to talk properly	( )	( )	( )	( )	( )
6) Using I language appropriately	( )	( )	( )	( )	( )
7) Not using negative communication patterns (preaching, accusing etc.)	( )	( )	( )	( )	( )
8) Using appropriate body language	( )	( )	( )	( )	( )

Study Design: In this study, the reliability of peer assessment was investigated by G theory and the analysis was done by using EduG. The design employed in this study conforms to what is known in generalizability terminology as a two-facet fully crossed G-study design (*sxtxr*), where *s* denotes the object of measurement (students), *t* denotes the tasks, and *r* denotes the raters. The cross symbol indicates that each student is rated by all the raters for all the eight tasks.

The final task of the course was composed of an assignment of role playing. There were two different roles: one was a parent and the other was a teacher. By acting as a teacher or as a parent, students were asked to act out by using the communication abilities mentioned during the semester in the course.

The three raters were chosen from 41 students according to their mid-term exam results: low, middle and high scored students, respectively. The role-players were matched randomly and as a result, there were 19 groups from 38 students. The three peer raters and the teacher rated the groups by using the scoring rubric.

The analysis of the data was carried out in three steps.

- In the first step, the ratings of three peer raters were investigated via G-study.
- In the second step, the overall ratings of three peer raters and the instructor ratings were examined. In other words, the mean of the three peer raters' ratings was calculated and a score which represents peer ratings was found. The data obtained from the mean score of the peer raters and the instructor's ratings were investigated via G-study.
- At the third step, to examine the peer assessment's reliability, a variety of D-studies were done by manipulating the number of raters and tasks. The D-studies were designed for the same universe of generalization as the universe of admissible observations in the G-study of the first stage of this research, and all D-studies were conducted under completely random effects *sxtxr* design.

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

3. Results

As stated before, the study conducted through three stages. Therefore, the results of each stage were given sequentially below.

Step 1: The estimation of variance components and reliabilities of scores obtained from peer ratings: In the study, 38 students (*s*) as the object of measurement were scored according to eight tasks (*t*) by three peer raters (*r*). Since the three peer raters scored all the tasks performed by all the students, the fully crossed two facets (tasks and raters) design (*sxtxr*) was applied. The results of the variance components for both main effects (*s*, *t* and *r*) and interactions (*st, sr, tr,* and *str*) were obtained through G study analysis according to this design are shown in Table 1.

Table 2. Analysis of Variance Results and Variance Component Estimates for Students, Tasks, Peer Raters and Interactions for Peer Assessments

Sources of variance	df	Sum of Squares	Mean of Squares	Variance	%
Student (*s*)	37	200.87	5.43	0.20	19.8
Task (*t*)	7	54.96	7.85	0.01	1.4
Rater (*r*)	2	1.58	0.79	-0.02	0.0
st	259	200.16	0.77	0.06	5.8
sr	74	31.59	0.43	-0.02	0.0
tr	14	84.98	6.07	0.14	14.2
str,e	518	308.52	0.59	0.59	58.7

As can be seen from Table 2, students (object of measurement of the study) account for the largest percentage (19.8 %) of the variance among the main effects. This result exhibits ideal case in measurements and can be interpreted as the differences between the students was revealed. It is desired that the variance coming from the object of measurement is high and the values for the other sources of variance are as low as possible. While the variance of the task main effect was 1.4 % and is rather small. From this result, it can be stated that there were not much differences between the tasks given to the students. Since the variance value of the rater main effect was negative, it was given as 0.00. When the estimated variance component is negative, it was stated that taking the negative variance as 0.00 is more appropriate as proposed by Cronbach (Shavelson & Webb, 1991; Brennan, 2001; Atilgan, 2004). The reasons of this situation are having small sample size or inappropriate measurement design. Here the design is not problematic. Therefore, maybe the sample size was the reason of negative variance. Moreover, from the 0.00 variance of rater main effect, the interpretation of such as there was no variation among raters could be made.

It can be seen from Table 2 that two way interactions of student-by-task account for 5.8 % of the total variance. As it is seen clearly, the value of 5.8 % demonstrates that each task's level of difficulty is the level of differentiation from student to student. This is inevitable in cases where the probability of differences that can stem from students' earlier experiences and attitudes is high. The fact that the 0.0 % of the total variance stems from the student-rater interaction demonstrates that the raters' scoring did not differ from one student to another. As another interaction, task-by-rater yielded 14.2 % of total variance. This value is the third highest value in the Table 1 and it indicates that the raters' scoring changed from task to task. Finally, the three way-interaction, students-by-tasks-by-raters, is also named as "residual" or "error" in the ANOVA model used here. If the measurement results are reliable in a research, this value of residual is desired to be as small as possible. According to Table 2, the three-way interaction accounted for 58.7 % of the total variance. Although, this is the largest variance value in Table 2, according to the G theory, this value of variance is desired to be as small as possible (Güler, Kaya Uyanık & Taşdelen Teker, 2012). Since the peer assessment is not widespread enough, there could be some random errors due to the

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

students who carried out the evaluations that made the residual variance so high. This value also can signal that the change in scores might have emerged due to different sources of variation which were not available in the study. The G and Phi coefficients obtained from peer assessment scores were 0.86 and 0.83, respectively. Therefore, the evaluations could be stated as having high reliabilities.

Step 2: The estimation of variance components and reliabilities of scores obtained from peers-instructor ratings: By taking the mean of peer ratings, a score for a representative of peer raters was obtained. To examine the consistency of this score with the instructor's one, G study was carried out to the data obtained from instructor's ratings and the mean of peer ratings. The variance components and variance results of peers-instructor ratings, the fully crossed design was applied with 38 students (*s*), eight tasks (*t*) and two raters (*r*) (peers ratings mean as the first rater and the instructor's as the second one). The results of the variance components for both main effects (*s*, *t* and *r*) and interactions (*st, sr, tr* and *str*) were obtained through G study analysis according to this design are shown in Table 3.

Table 3. Analysis of Variance Results and Variance Component Estimates for Students, Tasks, Raters and Interactions for Peer and Instructor Assessments

Sources of variance	df	Sum of Squares	Mean of Squares	Variance	%
Student (*s*)	37	109.73	2.97	0.15	17.2
Task (*t*)	7	64.93	9.28	0.02	1.7
Rater (*r*)	1	0.73	0.73	-0.02	0.0
st	259	141.25	0.55	0.25	2.2
sr	37	17.34	0.47	-0.01	0.0
tr	7	56.12	8.06	0.19	22.2
str,e	259	131.02	0.51	0.51	56.6

The values obtained in Table 3, were not very different from the ones in Table 2. Therefore, the interpretations of the results of Table 3 were similar to the interpretations of Table 2 results stated above. As can be seen in Table 3, the biggest variance value was obtained for *str* interaction effect (56,6 %) as in Table 2. The variance value of the task-by-rater interaction effect was the second higher value. It means that the raters' scoring changed from task to task. Finally the third highest variance value was obtained from the student main effect. It indicates that the differences among the students were revealed. The other variance sources' values are quite small means there are no differences. The G and Phi coefficients obtained from peer-instructor assessment were 0.82 and 0.76 respectively. As a result of this, it could be stated as, the reliability of peer-instructor was relatively high.

Step 3: D-Studies of peer ratings: To determine the most appropriate number of peer rater for peer assessment, the G (generalizability) and Phi "Φ" (dependability) coefficients obtained by using the variance values obtained from the first step of this study given in Table 2 by increasing and decreasing the rater numbers by D-study was given in Table 4. Besides, to determine the most appropriate task number for three-rater situations, one more D-study was conducted by changing the levels of tasks.

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

Table 4. G and Phi coefficients of D Studies

Number of Raters													
2		**3**		4		5		6		10		15	
G	*Φ*	**G**	***Φ***	G	*Φ*	G	*Φ*	G	*Φ*	G	*Φ*	G	*Φ*
0.82	0.78	**0.86**	**0.83**	0.89	0.86	0.90	0.88	0.91	0.89	0.93	0.92	0.94	0.93
Number of Tasks													
3		4		6		**8**		10		12		15	
G	*Φ*	G	*Φ*	G	*Φ*	**G**	***Φ***	G	*Φ*	G	*Φ*	G	*Φ*
0.70	0.65	0.76	0.72	0.82	0.79	**0.86**	**0.83**	0.87	0.86	0.90	0.88	0.92	0.90

Note: The bold rater and task numbers were the actual ones used in the study.

As can be seen from Table 4, the G and Phi coefficients for the application of peer assessment for 38 students, eight tasks and three raters fully crossed random design were estimated as 0.86 and 0.83 respectively. As a result of decreasing the number of raters from three to two, the G and Phi coefficients also decreased to 0.82 and 0.78. It was seen that, as expected, the increase of the rater number increases both generalizability and dependability values. When the number of raters were increased to 6 and above, the coefficients become 0.90 and above. On the other hand, increasing the rater number to 15 did not affect the values too much. As is apparent from Table 2, the rater's main affect's variance to the total variance was 0.00, the rater number did not affect the reliability so much. Moreover, increasing the number of raters above three does not increase reliability value significantly. Therefore, increasing the number of raters will not bring any benefits to the similar studies to be performed in the future; and it would not be a practical way in cases where it is difficult to act as peer rater in both small and crowded classes.

At the last line of the Table 4, there are D-study results obtained by changing the task numbers. It can be seen that, when the number of tasks was four and above, the obtained reliability was also 0.70 and above. Therefore, it can be concluded that, if there are more than three tasks, the reliabilities will be computed as higher than 0.70. With six, eight and ten tasks, the reliabilities increase to 0.80 and above. Finally, with 12 and above tasks, the reliability values reach to 0.90 and above.

4. Discussion

This study was carried out for determining the quality and effectiveness of peer assessment which can be used in education. In this context, the performances of students which were based on acting out as a teacher or as a parent according to given scenarios containing the communication tasks were graded by both peers and the instructor by using developed scoring rubric and the obtained data were analyzed by using generalizability theory.

According to the results of the first and second steps of the study, the three-way-interaction variance components were the highest. Although literature supports the use of peer assessment, there are several problems and limitations that have been repeatedly associated with the process. Many of these arise as a result of the 'newness' of peer assessment as a formal assessment tool in higher education. For example, academic staff and students generally have little experience with this form of assessment (Ballantyne, Hughes, & Mylonas, 2002). Indeed, peer assessment was also a new method to the participants of this study. Because of this reason, the residual variance is the largest variance value in both Table 2 and Table 3.

Another reason of the high residual variance could be that students often lack confidence in their own and peers' abilities as raters. Ballantyne et al.'s (2002) study reveals that once first-year students have gained experience in using peer assessment, they should be more comfortable in using this technique in subsequent years of studies. Second- and third-year students, however, are likely to be more confident in participating in assessment processes than first-year students and consequently be more attuned to the requirements and standards expected when undertaking

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

assessment tasks. Such confidence and experience may result in higher levels of satisfaction with the peer assessment process than the less experienced first-year students. This project indicates that if peer assessment is used with first-year students, the process needs to be structured very carefully. Therefore, this technique should be used from the first year of students to make him/her more experienced, comfortable and confident in both his/her own and their peers' abilities as assessors. Furthermore, usage of peer assessment at primary and high school students may lead more positive effects for further applications.

When the results of the study were investigated by means of reliability, quite high reliability coefficients were obtained from the first and the second step of the study. If the tasks (scoring criteria) were clarified to peers clearly by means of their means and indicators, the evaluations made by peers would become more reliable. In other words, by providing a description of the scoring criteria in advance, rubrics may positively impact interrater reliability (Moskal & Leydens, 2000). Moreover as Eckes (2009) stated, the interrater reliability increases when raters are trained (cited in Karakaya, 2015). Therefore, in the study, since the peer raters were trained by means of usage of scoring rubrics, the reliability coefficients were quite high. The reliability of peer ratings (G: 0.86 and Phi: 0.83) obtained from the first step of the study was very close to (even higher) reliability of the second step of the study where the instructor's ratings were included (G: 0.82 and Phi: 0.76). To summarize the reliability results of the study it can be concluded that both the peer ratings were reliable and the peers' ratings were consistent with the instructor's one.

As a result of D studies conducted to determine the most appropriate number of peer raters, quite high reliability coefficients (0.82 for G and 0.78 for Phi coefficients) were estimated with *just* two raters. It can be concluded that if the peers are trained well, it will be possible to have reliable rating results. When the number of tasks was differed, both the G and Phi coefficients increased to 0.70 and above with four and more number of tasks. Because of the reliable results obtained from peer and peers-instructor assessments, the peer assessment, which help students to gain the ability by means of evaluating others objectively and preparing them to real life situations, can be recommended as to be widely used in education system. Moreover, not only to reveal the effectiveness of peer assessment but also the reliability of this process, researchers can study peer assessment not only at higher education but also at different grade levels.

5. Conclusions and recommendations

Unlike objective tests, the grading of subjective tests like essays, performance tasks, open ended questions etc., are quite time consuming activities. Peer assessment can be used especially at universities to substantially decrease the workload of the staff of the courses. Since many study results showed that peer assessment was reliable (Gugiu & Gugiu, 2012; Sung et al, 2010) as well as this study, this effective method should be used not only for the benefit of the staff of the courses but also for the students' personal development in terms of gaining the skill of evaluating others objectively and critical thinking ability.

Many studies proposed that for peer assessment to be used at schools by teachers and by students with the orientation of their teachers, it should firstly be used by preservice teachers during their educations at education faculties (Koc, 2011; Bayat, 2010; Bal, 2009; Coklar & Odabasi, 2009; Mamur, 2011; Dogan & Kutlu, 2011). Therefore, it can be proposed that not only the peer assessment method but also all the other alternative methods mentioned briefly to preservice teachers during their education at universities' education faculties and after graduation to elementary and high school teachers by means of in-service training should be used frequently.

Our study had some limitations, which should be considered when determining the generalizability of the results. Although the sample size of the study was relatively small (N=38) which could be the result of getting negative variance results of the first and the second stage of the study given at Table 2 and Table 3 , G theory allows meaningful results. In this study, the participants were all undergraduate sophomore students. Perhaps the results would be different for

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

other levels of undergraduate students especially for higher grades, graduate students and high school and elementary school students due to potential differences in maturity levels and critically analyzing a peer's performance. Therefore, similar studies could be done by other reserachers by taking into account different samples of students.

The design of the study was a two-facet fully crossed G-study design which means all the students were rated by all raters for all tasks. However, if there are too many students to rate, the rating cannot be done by the same raters. In other words, the design could be nested instead of crossed as in this study. Besides this situation, the number of students evaluated by raters could be change from one rater to another. In this time, the design become unbalanced. The reliability of peer assessment could be done for nested and unbalanced designs for other studies if possible.

References

Alfallay, I. (2004). The Role of Some Selected Psychological and Personality Traits of The Rater in the Accuracy of Self-and Peer- Assessment. *System,* 32, 407-425.

Aryadoust, V. (2016). Gender and Academic Major Bias in Peer Assessment of Oral Presentations. *Language Assessment Quertly an International Journal,* 3(1), 1-24.

Atılgan, H. (2004). *Genellenebilirlik Kuramı ve Çok Değişkenlik Kaynaklı Rasch Modelinin Karşılaştırılmasına İlişkin Bir Araştırma*. Unpublished Doctoral Dissertation. Hacettepe University: Ankara.

Bal, A. P. (2009). T*he Evaluation of Measurement and Evaluation Approaches Used in Fifth Grade Mathematics Instruction in Terms of Students' and Teachers' Opinions*. Unpublished Doctoral Dissertation, Cukurova University, Adana, Turkey.

Ballantyne, R., Hughes, K., & Mylonas, A. (2002). Developing Procedures for Implementing Peer Assessment in Large Classes Using an Action Research Process. *Assessment & Evaluation in Higher Education,* Vol. 27, No. 5.

Baştürk, R (2008). Applying The Many-Facet Rasch Model to Evaluate Powerpoint Presentatiton Performance in Higher Education. *Assessment&Evaluation in Higher Education*, 33(4), 431-444.

Bayat, O. (2010). İngilizce Yazılı Anlatım Derslerinde Uygulanan Akran ve Öz Değerlendirme EtkinliklerineYönelik Öğrenci Görüşleri. *Dil Dergisi*, 150, 70-81.

Baykul, Y. (2000). *Eğitimde ve Psikolojide Ölçme*, Ankara: ÖSYM Yayınları

Brennan, R. L. (1992). *Elements of Generalizability Theory*. New York: Springer-Verlog.

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

Crocker, L., & Algina, J. (1986). *Introduction to Classical and Modern Test Theory*. Harcourt Brace Javanovich College Publishers, USA.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The Dependability of Behavioral Measurements: Theory of Generalizability for Scores and Profiles*. New York: Wiley.

Coklar, A. N., & Odabasi, H. F. (2009). Determining the Assessment and Evaluation Self-Efficacies of Teacher Candidates Regarding Education Technology Standards. *Ahmet Kelesoglu Education Faculty (AKEF) Journal,* 27, 1-16.

Dogan, C. D., & Kutlu, O. (2011). Factors Related Learning Which Effect Pre-Service Teachers' Preferences on Alternative Assessment Methods. *Kastamonu Education Journal*, 19(2), 459-474.

Donnon, T. , McIlwrick, J. & Woloschuk, W. (2013). Investigating the Reliability and Validity of Self and Peer Assessment to Measure Medical Students' Professional Competencies. *Creative Education*, 4, 23-28. doi: 10.4236/ce.2013.46A005.

Esfandiari, R. (2015). Rater Errors Peer-Assessors: Appliying the Many-Facet Rasch Measurement Model. *Iranian Journal of Applied Linguistics*, 18(2), 77-107.

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

Falchikov, N. (1998). Involving Students in Feedback and Assessment : A Report from the Assessment Strategies in Scottish Higher Education (ASSHE) project, in: S. BROWN (Ed.) *Peer Assessment in Practice*, SEDA Paper 102, Birmingham, SEDA.

Falchikov, N. (1995). Peer Feedback Marking: Developing Peer Assessment. *Innovations in Education & Training Interational,* 32(2), 175-187, DOI: 10.1080/1355800950320212.

Falchikov, N. (1986). Product comparisons and process benefits of collaborative peer group and self assessments. *Assessment & Evaluation in Higher Education,* 11:2, 146-166, DOI: 10.1080/0260293860110206

Goodrich Andrade, H. (2001). The Effects of Instructional Rubrics on Learning to Write . *Current Issues in Education*, 4(4), 1-22.

Gugiu, M. & Gugiu, P. C. (2012). Assessing the Reliability of Peer Evaluation of Undergraduate Research Papers Through the Use of Generalizability Theory. American Political Science Association, Conference Paper, 1-34.

Güler, N., Kaya Uyanık, G. & Taşdelen Teker, G. (2012) *Generalizability Theory*. Pegem Akademi Publishing, Ankara, Turkey.

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.

Hafner, J. C. & Hafner, P. M. (2007). Quantitative Analysis of the Rubric as an Assessment Tool: An Empirical Study of Student Peer-Group Rating, *International Journal of Science Education*, 25(12), 1509-1528.

Han, K. S., Mun, G., S. & Ahn, J. Y. (2009). Comparing the Use of Self and Peer Assessment: A Case Study in A Statistics Course. *Communications of the Korean Statistical Society*, 16(6), 979-987.

Karakaya, İ. (2015). Comparison of Self, Peer and Instructor Assessments in the Portfolio Assessment by Using Many Facet Rasch Model. *Journal of Education and Human Development*, 4(2), 182-192.

Koç, C. (2011). Sınıf Öğretmeni Adaylarının Öğretmenlik Uygulamasında Akran Değerlendirmeye İlişkin Görüşleri. *Kuram ve Uygulamada Eğitim Bilimleri*, 11(4), 965-1989.

Kutlu, Ö., Dogan, C. D., & Karakaya, İ. (2009). *Ogrenci Basarisinin Belirlenmesi: Performansa ve Portfolyoya Dayali Durum Belirleme.* Ankara: Pegem Academy Publishing.

Lee, G., & Frisbie, D. A. (1999). Estimating Reliability Under a Generalizability Theory Model for Test Scores Composed of Testlets. *Applied Measurement in Education*. 12(3), 237-255.

Luft, J. (1997). Design Your Own Rubric. *Educational Leadership,* 20(5), 25-27.

Mamur, N. (2011). The Qualification of Pre-Service Visual Art Teachers About Measurement and Evaluation Tools and Approaches of Their Branch. *Turkish Educational Sciences Journal*, 9(3), 597-626.

Marcoulides, G. A. (1989). The Application of Generalizability Analysis to Observational Studies. *Quality and Quantity*, 23, 115–127.

Marcoulides, G. & Simkin, M. G. (1992). Evaluating Student Papers: The Case for Peer Review*, Journal of Education for Business,* 67(2), 80-83.

Marty, C. M., Henning, J. M., & Willse, J. T. (2010). Accuracy and Reliability of Peer Assessment of Athletic Training Psychomotor Laboratory Skills, *Journal of Athletic Training*, 45(6), 609–614.

Mcdowell, L. (1995) The Impact of Innovative Assessment on Student Learning. *Innovation in Education and Training International*, 32(4), 302–313.

Moskal, B. & Leydens, J. (2000). Scoring Rubric Development: Validity and Reliability. *Practical Assessment, Research and Evaluation*, 7(10), 1-11.

Mowl, G. & Pain, R. (1995). Using Self and Peer Assessment to Improve Students' Essay Writing: A Case Study from Geography. *Innovation in Education and Training International*, 32(4), 324–335.

Nunally, J. C. & Bernstein, I. H. (1994). *Psychometric Theory*. New-York: Mc-Graw-Hill.

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

Orsmond, P. & Merry, S. (1996). The Importance of Marking Criteria in the Use of Peer Assessment. *Assessment and Evaluation in Higher Education*, 21(3), 239–250.

Popham, J. W. (1997). What's Wrong and What's Right With Rubric. *Educational Leadership*, 55(2), 72-75.

Reed, M. W., & Burton, J. K. (1985). Effective and Ineffective Evaluation of Essays: Perceptions of College Freshmen. *Journal of Teaching Writing,* 4(2), 270-283.

Reynolds, C. R., Livingston, R. L., & Willson, V. L. (2009). *Measurement and Assessment in Education.* Upper Saddle River, NJ: Pearson/Merrill Publishers.

Sadde, P. M. & Good, E (2006). The Impact of Self and Peer-Grading on Student Learning. *Educational Assessment,* 11(1), 1-31.

Searby, M. & Ewers, T. (1997) An Evaluation of The Use of Peer Assessment in Higher Education: A Case Study in The School Of Music, Kingston University. *Assessment and Evaluation in Higher Education*, 22(4), 371–383.

Semerci, Ç. (2011a). Mikro Öğretim Uygulamalarinin Çok Yüzeyli Rasch Ölçme Modeli ile Analizi. *Eğitim ve Bilim,* 36 (161), 14-25.

Semerci, Ç. (2011b). Doktora Yeterlilikler Çerçevesinde Öğretim Üyesi, Akran ve Öz Değerlendirmelerin Rasch Ölçme Modeliyle Analizi. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi.* 2(2), 164-171.

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability Theory: A Primer.* USA: SAGE Publications.

Strang, K. D (2013). *Exploring Summative Peer Assessment During A Hybridundergraduate Supply Chain Course Using Moodle.* Paper Presented 30th Ascilite Conference, 1-4 December, Sydney.

Sung, Y., Chang, K., Chang, T., & Yu, W. (2010). How Many Heads are Better Than One? The Reliability and Validity of Teenagers' Self- and Peer Assessments, *Journal of Adolescence*, 33, 135–145.

Şahin, M. G., Taşdelen Teker, G. & Güler, N. (2016). An Analysis of Peer Assessment through Many Facet Rasch Model. *Journal of Education and Practice*, 7(32), 172-181.

Topping, K. J. (1998) Peer Assessment Between Students in Colleges and Universities. *Review Of Educational Research*, 68 (3), 249–276.

Topping, K. J., Smith, E. F., Swanson, I. & Elliot, A. (2000) Formative Peer Assessment of Academic Writing between Postgraduate Students. *Assessment and Evaluation in Higher Education*, 25(2), 146–169.

Topping, K. J. (2009). Peer Assessment. *Theory into Practice*, 48(1), 20-27.

Voltan Acar, N. (2012). *Insan Iliskileri ve Iletisim.* Ankara: Nobel Academic Publishing.

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