ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2019; 10(4); 406-421

Investigation of the Reliability of Teachers, Self and Peer Assessments at Primary School Level with Generalizability Theory *

Eda GÜRLEN **   Nagihan BOZTUNÇ ÖZTÜRK ***   Emel EMİNOĞLU ****

Abstract

This study aims at determining the reliability coefficients of teacher, self and peer assessments carried out at primary school level. In line with this aim, an interdisciplinary approach is adopted, and the notion of helpfulness included within the scope of values education is addressed in connection with the practices followed in Turkish, social studies and music lessons. The study group consists of 30 students of the third graders from a public school in the city of Ankara. In the light of the aim of the study, the Generalizability Theory is used for the data analysis. It is found out at the end of the study that the variance component estimated for the main effect of the student is the largest component of the total variance in all three lessons. When G and Φ coefficients are examined, reliability coefficients are found to be over .80 in music, and over .90 in Turkish and social studies. According to G-Facet analysis results, when teacher and peer assessments are excluded from the analysis, respectively, G and Φ coefficients have a decreasing tendency whereas these coefficients increase when self-assessment is excluded from the analysis. Especially in the music lesson, the reliability coefficients obtained by excluding teacher and peer assessments from the analysis are found to be around .60, which is a remarkable result.

Key Words: Teacher assessment, self-assessment, peer assessment, Generalizability Theory.

INTRODUCTION

Evaluation, which is an important element of the education system, has important functions such as providing information about the effectiveness and efficiency level of the teaching process, determining the degree of achievement of the previously-set goals and revealing the strengths and weaknesses of the practices followed during lessons. Implementing the evaluation activities thoroughly ensures the continuous control of the education and thus makes it possible to find a quick remedy for the troubles that come out at any stage of education and produce robust solutions for problems. Moreover, it enables the identification and then the elimination of learning difficulties and deficiencies by monitoring student development. It also identifies sources of success and failure and helps to uncover elements that affect education positively and negatively. Thus, it becomes possible to support the practices that improve the quality of education and to take timely measures against obstacles and threats. By also shedding light on planning and orientation studies for the future, education can be improved efficiently and quickly (Çeçen, 2011; İşman & Eskicumalı, 2003; Kurudayıoğlu, Şahin & Çelik, 2008; Turgut & Baykul, 2015; Yaşar, 2017).

Teachers use different methods in order to make an assessment that can reveal every aspect of the change created by all educational activities. As a result of these methods, students are assessed from the perspectives of teachers and experts according to the criteria prepared by them. However, education and training are processes that come to life with interaction. The fact that the point of view of the students actively participating in this interaction is not included in the assessment activities constitutes

* A part of this article was presented as an oral presentation at the 17th International Primary Teacher Education Symposium.

** Assoc. PhD., Hacettepe University, Faculty of Education, Ankara-Turkey, edaerdem@hacettepe.edu.tr, ORCID ID: 0000-0002-1719-9840

*** PhD., Hacettepe University, Lifelong Learning Center, Ankara-Turkey, nagihanboztunc@hacettepe.edu.tr, ORCID ID: 0000-0002-2777-5311

**** Teacher, Talatpaşa Secondary School, Ankara-Turkey, emel.eminoglu@gmail.com, ORCID ID: 0000-0003-1712-8652

To cite this article:
Gürlen, E., Boztunç-Öztürk, N., & Eminoğlu, E. (2019). Investigation of the reliability of teachers, self and peer assessments at primary school level with generalizability theory. *Journal of Measurement and Evaluation in Education and Psychology, 10*(4), 406-421. doi: 10.21031/epod.583891

Received: 28.06.2019
Accepted: 21.08.2019

an important deficiency. Assessment activities conducted in this way will not become meaningful enough for the students who do not participate in the process and therefore will not perform their intended functions fully. The assessment activities can be meaningful only if the students use the assessment criteria for their own studies as well as other studies. In this way, students can realize that the assessment process is a deep learning experience. By comparing their own work included in an activity with other students' work related to the same activity, they can reach a more in-depth learning level and understand the working principles of the mind during the assessment process. Thus, they can also have an idea about how the teacher performs the assessment process. This can open the way for the teacher-student dialogue and enable the students to think about the arrangements to be made after the assessment and to take responsibility. Students who take responsibility for their own learning processes have the opportunity to become independent learners who think, direct, realize their own development, organize their own work, criticize themselves, and learn. An individual who has the ability to decide whether a behavior they exhibit meets the criteria related to that behavior will also have the ability to control their own behavior independently of any authority during their life. Therefore, assessment activities will contribute to the education of individuals who have gained autonomy for lifelong learning (Race, 2001; Sünbül, 2007; Wilson & Jan, 1993).

The world of education, which has discovered this aspect of assessment activities, tends to assess individual's learning through methods in which the individual is at the center. Such assessments, although they are more costly and time-consuming, contribute to the learning of the students and the professional development of the teachers by integrating learning and assessment. These activities represent not only a scoring exercise but also a dynamic process in which learning skills develop through active participation whereas in-depth learning turns out to be a possible phenomenon. Students' involvement in this process allows them to understand that assessment is not just a grading process. Students who are not adequately informed about the objectives and functions of the assessment may not be able to fully understand the points that their teaching activities are intended to achieve. When students are not fully aware of what is expected from them, their motivation to learn can be affected adversely. This may lead them to develop negative attitudes towards learning. Students who understand the purpose and necessity of the assessment activity can explore their strengths and weaknesses by approaching the assessment criteria more realistically. Self-discovering students focus directly on learning by taking responsibility for their own learning, and they turn out to be self-confident, critical and independent learners (Ballantyne, Huges & Mylonas, 2002; Boud, 1986; Cihanoğlu, 2008; Cram, 1995; Falchikov, 2001; Tekindal, 2014; Topping, Smith, Swanson & Elliot, 2000).

In order for such assessment activities to be carried out objectively, students should be included in the process from the first stage of assessment. Students should actively participate in the process of deciding on the type of assessment, determining which learning outcomes will be assessed, and establishing the criteria to be used. Teachers and students should discuss and agree on these issues. There should be a harmonious relationship among those who are involved in the assessment. Thus, students can realize the ideal behaviors expected from them, the reasons why they are expected to display these behaviours and the necessity of learning. Therefore, it will be possible to develop the skills to establish a criterion for a specific behavior and grading the quality of that behavior. The participation of students in these discussions will also be beneficial in terms of communication and self-expression skills. With all this learned, students can manage their own learning processes from the beginning till the end. They can decide on what is needed to raise their learning levels (Alıcı, 2010; Stiggins and Chappius, 2005; Woolfolk 2002).

The participation of students in assessment activities also contributes to the creation of a healthy teaching-learning environment. These activities give the teacher information about how the student thinks and, therefore, can learn. They enable teachers to recognize students in different aspects including affective characteristics. Thus, they guide the teacher in organizing teaching activities. They also help the student to understand how the teacher thinks. When students get involved in the process using similar ways of thinking, they feel that they become part of the learning environment. When students fulfil their potentials, their academic self-concept develops in a positive way; and they become more self-confident. They get proud of what they have learned by seeing their achievements and their level of development over time. Thus, they get happy and turn out to be willing to learn (Bahar, 2006; Stiggins and Chappius, 2005).

Assessment activities carried out with the participation of students allow students to become aware of other students' learning after they become aware of their own learning. Starting from themselves, the students take responsibility for each other's learning and gain the ability to assess other individuals. That's why self-assessment and peer-assessment are the two most important types of assessment that enable students to improve in this way.

Self-assessment means that students make judgments about the extent to which they fulfill these criteria by applying the assessment criteria for their own studies. Thus, students discover what they know, what they can do, how they feel, and how they learn. In this discovery process, students who have the opportunity to use their high-level thinking skills from a critical point of view are provided with the skill to make sense of themselves objectively. By becoming familiar with their strengths and weaknesses, students become aware of their learning problems. They can produce solutions to their own learning problems by using the detailed information they have acquired about their own learning paths. They develop an ability to plan their future studies and work by judging their learning experiences. Therefore, it can be said that a student who has the ability to evaluate his/her achievement will reach the competency level necessary to achieve greater success. Thus, students should be supported to form a set of productive and realistic objectives with an action plan based on the feedback resulting from the self-assessment (Alıcı, 2010; Boud, 1986; Kutlu, Doğan & Karakaya, 2008; Mistar, 2011; Stiggins, 1997; Tekindal, 2014).

From the perspective of cognitive and constructivist learning theories, it is seen that self-assessment helps the learner to structure the knowledge. According to these theories, newly-acquired information can be meaningful for students only when they associate the new pieces of information with the already existing ones. Self-assessment contributes to establishing a link between the existing knowledge and understanding and the new ones by giving meaningful feedback to students based on the criteria they have internalized before. In this way, students learn by constantly comparing their knowledge and understanding with their learning objectives. This shows that self-assessment is also effective in establishing learning goal orientation. Learning objectives require a certain degree of internal processing of information. Self-assessment contributes to the motivation of the learning type as it improves internal control, knowledge, understanding, and skills so that students can be aware of their progress towards understanding the information fully. (McMillan, trans. 2015).

Self-evaluation is closely related to the development of an individual's reflection ability. Reflection involves one's self-monitoring as an external observer and the development of decision-making skills for better action in the future (Osterman & Kottkamp, 1993). Students' developments in reflective behaviors and skills constitute the most important point in self-assessment. In order to make progress in this regard, it is necessary to clearly define which behaviors and skills will be assessed and the corresponding trends. In order to obtain reflective comments about students' work, what is expected from them should be clearly stated. Simple examples can be used to visualize trends in this field. It can be started by questioning the accuracy of the answers given by the students to the questions about the lesson. Afterwards, questions such as why the answer is not correct, what the wrong answer exactly tells the student, and what needs to be done in order to give the correct answer can be asked (McMillan, trans. 2015).

Self-assessment tools can be prepared in different ways. They may range from a format that is prepared in a draft form of checklists and questions to a format that questions the reflections they have produced from a composition before; however, what is important is that students should take responsibility for their learning by determining what they have learned and in which areas they have problems whatever the chosen self-assessment tool is (Bahar, Nartgün, Durmuş & Bıçak, 2008). Also, students' self-assessments should be kept in students' personal development files (Woolfolk, 2002).

There are a number of factors that prevent self-assessment from being performed in a healthy way. Such factors include students who are biased about assessing their own learning because of having difficulty in making objective interpretations, who overestimate or underestimate their own abilities, who are able to make self-evaluation because of being unaware of their own abilities, who do not consider themselves sufficient to perform self-assessment or who believe that assessment should only be done by the teacher. In this case, continuous self-assessment, clarification of how students can make self-assessment and encouraging students to self-assessment will be effective in eliminating these factors (Alıcı, 2010; Tekindal, 2014).

On the other hand, when peer assessment is in question, students evaluate the performances or quality of the products belonging to others by applying the relevant criteria to the work of other students of similar status. Thus, they learn new pieces of information together and from each other via examining and criticizing different works. Peer review involves providing students with feedback from their peers about the quality of their work. Peer feedback encourages working together and learning together. Students increase awareness about their own learning needs by seeing their strengths and weaknesses. They can even get to know each other better than their teachers and give more detailed feedback. In this respect, peers can provide feedback to a greater number of students than the teacher in crowded classrooms. Thus, they can develop each other’s talents and skills. However, students’ mastering in performing an effective peer review requires a lot of practice. The assessment criteria should also be clear, appropriate, and discussed with the students. (Ballantyne et al., 2002; Falchikov, 1986, 2001; Topping et al., 2000; Tekindal, 2014).

Peer review has turned out to be a part of our success development since the first years of our lives. When children get informal feedback from their peers, this contributes to their social development to a great extent. The social development of the students can be accelerated significantly when the power of peer feedback is included in the planned assessment activities. Students have the opportunity to improve the quality of their products through teamwork. They can see the mistakes and deficiencies in their studies from the point of view of their friends although they do not realize these mistakes and deficiencies on their own. Thus, the defects can be corrected, and the works can be carried to higher levels. It is no doubt that students also develop a number of social skills such as communication, cooperation, discussion besides improving their products of studies in such a process. Students learn to criticize each other constructively and accept criticism with tolerance. When they work together in this way, they can see themselves as a member of the community and develop a sense of belonging. They grow up as individuals who can use what they learn from their peers both in their own personal development and in the development of the society as a whole (Alıcı, 2010; Tekindal, 2014).

Initially, peer assessment, as well as self-assessment, may be difficult to perform objectively. Students are more likely to behave subjective when evaluating their peers whom they like and who are more popular than others in the class. However, when these studies are carried out routinely at regular intervals, students will start to carry out better assessments. The purpose, importance and implementation steps of peer assessment should be clearly explained to the students in order to improve peer-assessment process. It should be emphasized that it is necessary to make a distinction between the students’ features to be assessed and other qualities of these students that will be excluded from the peer-assessment process. Peer-assessment will be more objective when students start to feel that they are working together and not competing. Moreover, it is possible to carry out the peer-assessment process more objectively when students do not know whose product is being assessed, and assessment is done by more than one student or the students to assess a product are chosen randomly (Bahar et al., 2008; Alıcı, 2010).

When self-assessment and peer-assessment are used together, they help and develop each other. However, students should be able to use their assessment skills actively and correctly in order to achieve this development. This is closely related to providing students with the opportunity to grow up in a culture of assessment and evaluation. Researches show that performing such activities routinely from the first year of primary education contributes significantly to critical thinking skills (Alıcı, 2010). In addition to this, when the related literature is examined, it is seen that such assessment should be carried out continuously in order to handle this process in a healthy way. Therefore, students’ participation in assessment activities should be ensured from the first stages of education. Assessment activities, which include both students’ and teachers’ perspectives, will provide more detailed data and develop more effective solutions. Examining the students’ point of view by comparing them with the teachers’ point of view will guide the development of assessment activities in this field. In this case, it is important to determine whether the primary school students in the first stages of education differ from the teachers who are experts in the field in terms of evaluating their own and their peers’ work according to certain criteria. If so, identifying the scope of this difference is important to determine where we are in the field of assessment. When the related literature is reviewed, it is clear that there are numerous studies on self- and peer assessment in Turkey, but there are a limited number of studies that examined self- and peer assessment through comparing the reliability of these assessments. Considering that reliability is one of the significant limitations of such assessments, it is thought that addressing this issue is important in terms of revealing the level reached in studies that are have been carried out about assessment involving students’ participation. Therefore; this study aims at determining the reliability of the scores obtained from teacher and student (self and peer) assessments in primary school level. For this purpose, the researchers examined the change of reliability of scores obtained via self- and peer-assessment while evaluating the exemplar event-driven performance works that were done in Turkish, social sciences and music lessons at third grade of a primary school. It is thought that the study will contribute to more efficient assessment studies by examining the self-assessment and peer-assessment skills of primary school students.

METHOD

This study is a descriptive study since it is aimed to determine the reliability of teacher, self and peer assessments performed in the performance works done in Turkish, social sciences and music lessons at the third grade of primary school.

Study Group

The study group consists of 30 third grade students (14 boys and 16 girls) studying at a primary school in the city of Ankara. It was decided during the study that five students to be selected randomly among 30 students would score for peer assessment. As a result, the remaining 25 students were included in the study as the measurement object.

Data Collection Tools

The assessment, self-assessment and peer-assessment scales prepared by the Ministry of National Education and included in the teacher’s guide books were used as data collection tools after being simplified in accordance with performance works that had been prepared in line with the expert opinions (5 classroom teachers, 2 Turkish teachers, and 1 music teacher).

Writing Skills Assessment Scale included in the teacher’s guide book which has been used since 2013-2014 academic year upon the approval of the Ministry of National Education was used to assess the writing skills of the students in the Turkish lesson (Milli Eğitim Bakanlığı-MEB, 2013). Taking into consideration the length of time of the implementation and the performance task, a grading scale that consists of four criteria was created by selecting and arranging critical criteria among the ten measures included in the scale in accordance with the opinions of experts. The generated grading scale is given in Table 1.

Table 1. Grading Scale Used in Turkish Performance Task

Criterion	**Rating**		
	1	**2**	**3**
Following spelling rules			
Writing meaningful and normative sentences			
Writing events in order of occurrence			
Including the main idea in writings			

Discussion Scale included in the teacher's guide book which has been used since the 2015-2016 academic year upon the approval of Ministry of National Education was used to assess the discussion skills of the students in the social sciences lesson (MEB, 2017a). Taking into consideration the length of the time of the implementation and the prepared performance task, a grading scale consisting of four criteria was created by selecting and arranging critical criteria among the ten criteria in the scale in accordance with expert opinions. The generated grading scale is given in Table 2.

Table 2. Grading Scale Used in Social Sciences Performance Task

Criterion	**Rating**		
	1	**2**	**3**
Being able to express his/her idea clearly			
Interpreting the questions correctly and giving appropriate answers to the questions			
Following the rules of discussion			
Controlling the tone of voice and gestures			

Analytical-Rate Grading Scale for Song/Folk/March Performances included in the teacher's guide book which has been used since the 2017-2018 academic year upon the approval of Ministry of National Education was used to assess the singing performances of the students in the Music lesson (MEB, 2017b). Taking into consideration the length of time of the implementation and the performance task, the grading scale consisting of four criteria was created by selecting and arranging critical criteria among the six criteria included in the grading scale in accordance with expert opinions. The generated grading scale is given in Table 3.

Table 3. Grading Scale Used in Music Performance Task

Criterion	**Rating**		
	1	**2**	**3**
Singing the lyrics of the song correctly			
Singing the tune of the song correctly			
Paying attention to the rhythm of the song			
Using his/her voice correctly and effectively			

As a result, grading scales, which consist of four criteria and each of which is specific to the course, were used in each lesson. Grading scales are rated with three different smiley icons in accordance with the age group. The data set was prepared by the researchers as 1-2-3, which is the scoring equivalence of smiley icons.

Data Collection Procedure

An interdisciplinary approach has been used in this study. The issue of helpfulness within the scope of values education has been addressed in relation to the practices in three lessons. It is thought that it will be possible to examine the same subject from the angels of different methods in different disciplines by means of adopting such an approach, and in this way, it will be possible to obtain more detailed data. Moreover, it is aimed to help the students make a healthier assessment by organizing different knowledge and skills to form a meaningful whole and get students gain this meaningful whole. At the same time, it is thought that it will be possible to examine the differences in the perspectives of teachers and students about the assessment of different disciplines.

In the research, the same students' group was asked to do both peer and self-assessment in three different lessons. While the students' group remained the same, it was ensured that different teachers made the assessment in different lessons. In this case, firstly, the teacher and the students were informed about the type of assessment before the research started. Teacher assessments were conducted by two classroom teachers and one music teacher, each working in a public school with expertise and experience in the field. While the teacher of the class in which this study was being conducted made the assessment in social sciences lesson, the teacher of a different class made the assessment in Turkish lesson. In music lesson, the music teacher, who is also one of the researchers of this study, made the assessment. It was decided that music lesson should be conducted by a music teacher who had received a music education as Music lesson requires special skills and the scoring should be done as neutrally as possible. Since the music teacher is one of the researchers of this study, she already has detailed information about the grading scale and the scoring process. On the other hand, the classroom teachers that were to do scoring in Turkish and social sciences lessons were informed about the types of assessment and the grading scales in advance. For this purpose, classroom teachers were given training on how to do assessments using a grading scale, and they were provided with the opportunity to examine exemplary implementations with the researchers. In the process of informing students, short training was given on teacher assessment, self-assessment, peer assessment, and grading scales.

In the Turkish lesson, students were allowed to watch a cartoon film that was telling a fairytale based on the importance of helpfulness. The film was stopped at half, and the students were asked to write the end of the fairytale. After all the students completed their studies, they went to the blackboard one by one and read the rest of the fairytale as they had completed. Since writing rules were also included in the assessment, students' papers were examined by the peer students and the teacher immediately after each student finished reading. In this way, the writing skills of the students were assessed by the teacher and the students.

In the social sciences lesson, students were allowed to watch a short film that was explaining how charity can create a cycle by awakening the sense of helpfulness in people. Then, the students were asked to discuss in groups the positive and negative results that charity could produce based on the events they had watched. Groups of four students were established as they wished and the two groups mutually had the opportunity to discuss the topic. After each group finished discussion, the students' ability to discuss within the group was assessed by using grading scales prepared by teachers and students.

In the music lesson, a song that teaches the importance of helpfulness was taught to students by using ear-to-ear teaching method. Then, the students were asked to sing the song individually. The song performance of the students was assessed by the teacher and the students.

Data Analysis

In this study, it is aimed to determine the reliability of the scores obtained as a result of teacher, self and peer assessment. When the literature is examined; it is clear that Classical Test Theory (CTT), Generalizability (G) Theory and Item Response Theory (IRT) are employed to identify the reliability of the measurement results (Güler, 2011). Especially when it is focused on the studies that try to determine the reliability between different raters, it is seen that G Theory or IRT-based methods have been preferred more frequently compared to CTT (Atılgan, 2005; Börkan, 2017; Büyükkıdık & Anıl, 2015; Farrokhi, Esfandiari & Dalili, 2011; Farrokhi, Esfandiari & Schaefer, 2012; Karakaya, 2015; Matsuno, 2009; Nalbantoğlu-Yılmaz, 2017; Taşdelen-Teker, Şahin & Baytemir, 2016; Yıldıztekin, 2014). If a comparison is made on the basis of CTT and G-theory, it is seen that only one error source is allowed to be estimated in the reliability determination studies based on the CTT, while all error sources can be included in the analysis in the reliability analysis based on G theory. In addition to his; in G theory, the sources of error can be addressed separately and the interactions of error sources can be determined as a result of the analysis (Brennan, 2001; Güler, 2009; Güler, Kaya-Uyanık & Taşdelen-Teker, 2012; Shavelson & Webb, 1991). The study carried out by Taşdelen-Teker and Güler (2019) shows that G Theory is frequently used especially in inter-rater reliability and standard-setting studies. Due to these advantages and application areas of G Theory, in this study, G Theory was preferred in order to determine the reliability between different types of raters (teacher-self-peer).

In accordance with the purpose of this study, by using the students (s) who are the measurement objects and the rater type (r) and criterion (c) variability sources, the analysis was conducted on full crossed random two-facet design (sxrxc). The peer assessment, which is one of the rater types, was included in the analysis and the average score of those given to 25 students by the chosen 5 was taken. For the three courses covered in the study, the predicted variance components, G and Phi coefficients were calculated to determine the main and common effects of the variables that constitute the sources of variability. In addition, G and Phi coefficients were also calculated by using G-facets analysis when the rater types were excluded from the analysis respectively. The analyses were performed using the EduG 6.1 package program.

RESULTS

In this section, estimated variance components, reliability values and G-Facet components done according to rater type of teacher, self and peer assessment scores are given under separate titles for Turkish, social sciences and music lessons respectively.

1. Turkish Lesson

For the G study of sxrxc pattern which is completely crossed in Turkish lesson; the estimated variance components and percentages of total variance explanation are given as the main effects of s, r and c, and the common effects of sr, sc, rc, and src in Table 4.

Table 4. Estimated Variance Components for the Turkish Lesson

Sources of Variance	**Sum of Squares**	***df***	**Mean of Squares**	**Variance (σ^2)**	**%**
Student (s)	77.16853	24	3.21536	0.25745	64.0
Rater Type (r)	4.10027	2	2.05013	0.01712	4.3
Criterion (c)	3.32627	3	1.10876	0.01068	2.7
sr	6.17307	48	0.12861	0.00674	1.7
sc	7.12373	72	0.09894	-0.00090	0.0
rc	1.86453	6	0.31076	0.00836	2.1
src,e	14.63547	144	0.10164	0.10164	25.3
Total	114.39187	299			100%

It is seen in Table 4 that the estimated variance component (0.258) explains the 64.0% of the total variance for the main effect of student (s) in Turkish lesson. The main effect of the student has the biggest share in the total variance. Therefore, it can be concluded that the assessment process can determine the differences between students.

It is also clear that the estimated variance component (0.017) for the main effect of rater type (r) explains 4.3% of the total variance. The main effect of the rater type is the variance component which has the third-largest share in the total variance. According to this, it can be said that the scores given by the teacher, self and peers differ slightly.

It is seen that the estimated variance component (0.011) for the main effect of criterion (c) explains 2.7% of the total variance. In this case, it can be said that the given scores differ slightly from one criterion to another.

When the common effect values are examined, it is seen that the estimated variance component (0.007) for the common effect of student-rater type (sr) explains 1.7% of the total variance. The common effect of the student-rater type (sr) has the second-lowest variance of the total variance. In this case, it can be said that the scores given to students by different types of raters do not change much.

It is seen that the estimated variance component (-0.001) for the common effect of student-criterion (sc) explains 0.0% of the total variance. Student-criterion (sc) common effect has the lowest variance in the total variance, having a negative value. In cases where variance is negative, Cronbach et al. (1972) suggested that the variance value be zero (as cited in Doğan & Anadol, 2017). The reason for the variance being negative may be that the study group is small, or the measurement pattern is not suitable (Taşdelen-Teker et al., 2016). In this study, since there was no problem with the pattern, the finding is thought to be related to the size of the study group. Based on that, when the total variance of the student-criterion (sc) common effect is considered to be zero, it can be said that this effect does not contribute to the total variance. In short, students' performances do not differ according to criteria.

It is seen that the estimated variance component (0.008) for the common effect of rater type-criterion (rc) explains 2.1% of the total variance. This finding shows that there is a slight difference in the scoring from one criterion to the other according to the rater type.

As is seen, student-rater type-criterion (residual) common effect variance component (0.102) explains 25.3% of the total variance. This ratio is the second-largest value in the total variance. However, the share of the student-rater type-criterion (residual) common effect variance component in the total variance is expected to be small (Shavelson & Webb, 1991). As a result, this situation may indicate that the student-rater type-criterion common effect and/or the random error in the measurement can be large.

When G and Phi coefficients are examined, G coefficient is found to be .96 based on relative error variance, and Phi coefficient is found to be .93 based on absolute error variance. It can be said that these values are quite high values within the acceptable limits of the reliability coefficient (Brennan, 2001).

As a result of the G-facets analysis, the reliability coefficients obtained when each of the rater types is not included in the analysis respectively are given in Table 5.

Table 5. G-Facets Analysis of Rater Types

Facet	Level	G	Φ
Rater Types (n_r = 3)	Teacher Assessment	.92	.88
	Self Assessment	.97	.94
	Peer Assessment	.92	.86

As is clear in Table 5, the G and Φ coefficients decrease slightly when the teacher or peer assessments are excluded from the analysis. However, the obtained reliability coefficients are quite high. As a result of excluding the self-assessment from the analysis, both G and Φ coefficients increase.

2. Social Sciences Lesson

For the G study of sxrxc pattern, which is completely crossed in the Social Sciences lesson, the estimated variance components and total variance explanation percentages are given as s, r and main effects and sr, sc, rc, and src common effects in Table 6.

Table 6. Estimated Variance Components for Social Sciences Lesson

Sources of Variance	Sum of Squares	*df*	Mean of Squares	Variance (σ^2)	%
Student (s)	109.41813	24	4.55909	0.35791	74.6
Rater Type (r)	1.01840	2	0.50920	0.00034	0.1
Criterion (c)	0.59987	3	0.19996	-0.00254	0.0
sr	9.62827	48	0.20059	0.03713	7.7
sc	8.33013	72	0.11570	0.02120	4.4
rc	1.95973	6	0.32662	0.01098	2.3
src,e	7.50027	144	0.05209	0.05209	10.9
Total	138.45480	299			100%

It is seen in Table 6 that the estimated variance component (0.358) explains the 74.6% of the total variance for the main effect of student (s) in social sciences lesson. As a result of obtaining the highest variance ratio from the student variable, it can be concluded that the assessment process can identify the differences between students.

It is seen that the estimated variance component (0.000) for the main effect of the rater type (r) explains 0.1% of the total variance. The main effect of the rater type is the variance component which has the second smallest share in the total variance. According to this, it can be said that the scores given by the teacher, self, and peer show almost no significant difference.

It is observed that the estimated variance component (-0.003) for the main effect of criterion (c) explains 0.0% of the total variance. The main effect of the criterion has the lowest variance in the total variance while it gets a negative value. If the total variance of this variable is considered as zero, it can be said that this effect does not contribute to the total variance. In short, the scoring does not differ according to the criteria.

When the common effect values are examined, it is seen that the estimated variance component (0.037) for the common effect of student-rater type (sr) explains 7.7% of the total variance. The student-rater type (sr) of the common effect has the third-highest variance in the total variance. In this case, it can be said that the scores given to students by the different rater types vary.

It is clear in Table 6 that the estimated variance component (0.021) for the common effect of student-criterion (sc) explains the 4.4% of the total variance. Student-criterion (sc) common effect has the lowest third variance in total variance. As a result, students' performances differ slightly according to the criteria.

It is seen that the estimated variance component (0.011) for the common effect of rater type-criterion (rc) explains 2.3% of the total variance. While this indicates that the rater-criterion (rc) common effect has the lowest third variance value, it can be said that the rater type may differ slightly from criterion to criterion.

Student-rater type-criterion (residual) common effect variance component (0.053) appears to explain 10.9% of the total variance. While this ratio appears to have the second largest value in the total variance, it may be an indicator that the student-rater type-criterion common effect and/or random errors in measurement may be large.

When G and Phi coefficients are examined; both G coefficient and Phi coefficient are found to be .94. It can be said that the obtained relative and absolute reliability coefficients are quite high within the acceptable limits.

As a result of the G-Facets analysis, the reliability coefficients obtained when each of the rater types is not included in the analysis respectively, are given in Table 7.

Table 7. G-Facets Analysis of Rater Types

Facet	Level	G	Φ
Rater Types (n_r = 3)	Teacher Assessment	.91	.90
	Self Assessment	.97	.96
	Peer Assessment	.89	.88

As is clear in Table 7, the G and Φ coefficients decrease slightly when the teacher or peer assessments are excluded from the analysis. This decrease was found to be slightly higher in peer assessment, but the obtained reliability coefficients are still quite high. It is seen that both reliability coefficients increased slightly when G and Φ coefficients are excluded from the analysis.

3. Music Course

For the G study of sxrxc pattern which is completely crossed in Music lesson, the estimated variance components and total variance explanation percentages are given as the main effects of s, r; and the common effects of sr, sc, rc, and src in Table 8.

Table 8. Estimated Variance Components for Music Lesson

Sources of Variance	**Sum of Squares**	***df***	**Mean of Squares**	**Variance (σ^2)**	**%**
Student (s)	71.25013	24	2.96876	0.21139	47.3
Rater Type (r)	5.35707	2	2.67853	0.02104	4.7
Criterion (c)	0.18600	3	0.06200	-0.00453	0.0
sr	17.49627	48	0.36451	0.06021	13.5
sc	13.77067	72	0.19126	0.02253	5.0
rc	2.00400	6	0.33400	0.00841	1.9
src,e	17.80933	144	0.12368	0.12368	27.7
Total	127.87347	299			100%

It is seen in Table 8 that the estimated variance component (0.211) for the main effect of the student (s) explains 47.3% of the total variance in music lesson. As a result of the scoring performed within the scope of music lesson, it can be concluded that differences between students can be identified.

It is seen that the estimated variance component (0.021) for the main effect of rater type (r) explains 4.7% of the total variance. Considering the main Moreover of the rater type; it can be said that the scores given by the teacher, self and peer vary.

It is observed that the estimated variance component (-0.05) for the main effect of criterion (c) explains 0.0% of the total variance. The main effect of the criterion has the lowest variance in the total variance while it gets a negative value. If the total variance of this variable is considered as zero, it can be said that this effect does not contribute to the total variance. In short, the scoring does not differ according to the criteria.

When the common effect values are examined, it is seen that the estimated variance component (0.060) for the common effect of student-rater type (sr) explains 13.5% of the total variance. In this case; while the student-rater type (sr) has the third-highest variance in the total variance, it can be said that with this finding, the scores given to students by different rater types differ.

While the student-criterion (sc) explains 5.0% of the total variance of the estimated variance component (0.023) for the common effect; it can be said that the scores given to the students differ according to the criteria. Considering the estimated variance component for the rater type-criterion (rc) common effect; it explains 1.9% of the total variance. According to this result; the scores obtained by the rater type according to the criteria differ slightly.

It is seen that the estimated variance component (0.008) for the common effect of rater type-criterion (rc) explains 1.9% of the total variance.

While the student-rater type-criterion (residual) common effect variance component (0.124) explains 27.7% of the total variance, this value is the second largest value in the total variance. Therefore, it can be said that the common effect of student-rater type-criterion and/or random errors in measurement may be large.

When the G and Phi coefficients obtained in the analysis are examined; the G coefficient is found to be .85 and the Phi coefficient is .83. It is seen that the obtained reliability coefficients are within the accepted limits according to the literature (Brennan, 2001).

The reliability coefficients obtained when each of the rater types in G-facets analysis is not included in the analysis respectively, are given in Table 9.

Table 9. G-Facets Analysis of Rater Types

Facet	Level	G	Φ
Rater Types (n_r = 3)	Teacher Assessment	.63	.60
	Self Assessment	.97	.96
	Peer Assessment	.68	.62

In Table 9; there is a significant decrease in the G and Φ coefficients obtained by excluding teacher or peer assessments from the analysis. The obtained reliability coefficients are lower than the acceptable reliability coefficient limit in the literature. There is an increase in both G and Φ coefficients as a result of excluding the self-assessment type from the analysis.

DISCUSSION and CONCLUSION

This study aims at identifying the reliability of the scores given by third-grade elementary school students through self and peer assessment methods and that of the scores obtained as a result of teacher assessment. An interdisciplinary approach has been adopted for that purpose, and the notion of helpfulness has been associated with Turkish, Social Sciences and Music lessons within the scope of values education.

G-theory was used in the study as it was aimed to include more than one source of error in the analysis and to examine the sources of variance in detail. Thanks to the advantages of the relevant theory, both main and interactive effects of variance sources were examined, and relative as well as absolute reliability coefficients were estimated.

When Turkish lesson is in question, it is seen that the component explaining the total variance is the main effect of the student (s). The fact that the main effect of the student (s) has the largest percentage of explanation is desirable during the assessment process, because it is obtained that the differences between students can be revealed by the assessment process (Atılgan, 2005; Doğan & Anadol, 2017; Taşdelen-Teker et al., 2016). It is seen that the total variance is the second mostly explained component by the residues (src,e) following the main effect of the student (s). This result may be an indicator that the common effect of student-rater type-criterion (src,e) and/or random errors may be large. The cause of random errors in this lesson can be that students who do not encounter such practices frequently experience a lack of excitement and motivation. Considering the main effect of the criterion (c) variable; it is seen that it explains 2.7% of the total variance. When evaluated in terms of criteria, it can be said that student and teacher perspectives differ in some of the criteria within the scope of writing skills. Another noteworthy finding obtained in the context of the Turkish lesson is that the common effect of student-criterion (sc) does not contribute to the total variance. In short, students' performances do not differ according to the criteria included in the grading scale. In this case, it can be said that these criteria assess the same skills.

When the results related to the social sciences lesson are considered, the main component explaining the total variance was the student (s) main effect, and after that, the largest share in explaining the total variance belongs to residues (src,e). It can be said that differences between the students can be revealed in the assessments made within the scope of social sciences lesson with the biggest share of the main effect of the students. The sources of random errors that may occur in this lesson are thought to be that there might be distractions and noise generated in the classroom by the students who did not participate in the activity. The main effect of rater type (r) on estimated variance values of social sciences lesson has a relatively small share in total variance. In other words, it can be said that the scores given in teacher, self and peer assessments show almost no significant difference. In the research, considering that the teacher's assessment of the social sciences lesson is done by that classroom's teacher, the result obtained is thought to be based on the fact that the teacher knows the students in the classroom better and that the students can score more easily in an assessment environment made by the classroom teachers.

When the results related to the estimated variance components within the scope of music lesson are considered, the main effect of the student (s) is the component that explains most of total variance in music lesson as is the case in other lessons. In this respect, differences among the students have been revealed in the assessment made in Music lesson. In addition, the effect of residues (src,e) in explaining the total variance has the largest share following the main effect of the student (s). Among the reasons why residues in music lessons have a high share, the reaction from the class during the individual performance of some students, and the excitement of students unfamiliar with individual performance can be included as the sources of random errors. Another remarkable finding obtained in the context of music lesson is that the main effect of the measure (c) does not contribute to the total variance; in other words, the scoring does not differ according to the criteria. This situation can be explained by the fact that all of the criteria are directed towards singing skills and the level of musical ability of the students has the same effect on the skills related to the criteria.

When G and Φ coefficients are examined, it is seen that G and Φ coefficients obtained for all three lessons are considerably higher than the acceptable value of .80 in the literature (Brennan, 2001). When the G and Φ coefficients are handled on lesson base, it is seen that the coefficients obtained in music lessons are lower than the coefficients obtained in Turkish and Social Sciences lessons. The coefficients obtained in the Turkish and Social Sciences lessons are above .90, and they are very close to each other. In the study, the fact that the teacher assessments in Turkish and social sciences were made by the classroom teachers and the assessment in music lesson was conducted by the music teacher can be considered as a factor affecting the reliability.

When the values obtained as a result of G-facet analysis of rater types are evaluated, if teacher and peer assessments are not included in the analysis for all three lessons, G and Φ coefficients decrease. While the new G and Φ coefficients obtained as a result of these decreases are still higher than the acceptable reliability coefficient for Turkish and social sciences lessons, they are below the acceptable limits for music lesson. When the scores obtained at the end of self-assessment were not added to the analysis, G and Φ coefficients obtained in all three lessons increased. This increase was more in music lesson than it was in other lessons. In the inclusion of peer assessment scores in analysis, the scores of the five raters were averaged. In short, the five raters acted as if they were one rater. In this case, even if one of the peers had not scored very accurately, it may have increased the reliability with the average of the others. But in self-assessment, students may have scored in favor of themselves because they only scored for themselves. When the age characteristics of the students are taken into consideration, instead of exhibiting a biased behavior by giving higher scores to their friends, they are thought to be as careful as possible. In this case, peer assessment and teacher assessment can be expected to be close to each other while self-assessment can be expected to be different from them. A similar result was observed in Salmaner’s (2015) study, which examined self, teacher, and peer scores with the multi-surface Rasch measurement model. In this study, Salmaner worked with 5th grade students, and as a result of the analysis, he found out that the most generous raters were self-raters and the strictest raters were teachers or peer raters. When the age group is taken into consideration, it can be said that students’ desire to succeed or the anxiety of failure might have created a tendency to give themselves higher scores.

When the literature is examined, it is observed in the studies carried out on the comparison of teacher, peer, and self-assessment at primary school level that students cannot make fully objective assessments; it is generally seen that self-assessments give the most generous scores (Salmaner, 2015; Sarıtaş, 2015). Börkan (2017) also scored the presentation performance of the students by using a grading key in a four-day peer review study with university students. As a result of the study, it was concluded that peer raters generally rated their friends in a very generous manner; and the strictness/generosity levels differed from each other when the raters were compared among themselves. Matsuno (2009) conducted another study in which peer assessment and self-assessment were handled together with teacher assessment. Matsuno (2009) conducted this study with 91 Japanese students between the ages of 19-21 and four teachers. In this study, especially high-performing students gave lower scores than estimated in the self-assessment process, whereas the raters were more tolerant and consistent in the peer assessment process. Regardless of their writing skills, they scored low on high-performing students and high on low-performing students. It was seen that most of the peer raters were consistent and showed less biased interactions than the self-assessment and teacher raters. Farrokhi et al. (2011) conducted a study to determine the tendency of centralism in self-assessment, peer assessment and teacher assessment using the multi-faceted Rasch model. 194 evaluators assessed 188 written compositions with a six-analytical scale and concluded that there was a centrality among peers and self-evaluation. In 2015, Karakaya made a comparison between self-evaluation, peer evaluation and teacher evaluations in evaluating portfolio files of teacher candidates. The findings of the study indicated that the raters were more tolerant in the self-assessment and more rigid in the peer assessment, and it generally found a statistically significant difference between the evaluators. In another study conducted by Nalbantoğlu-Yılmaz (2017) with 56 teacher candidates, it was aimed to determine whether there were differences in self-evaluation and peer evaluations related to a project and to reveal the reliability of the grades given by the teacher candidates and their peers and the scores given by their teachers. As a result of the study, no significant difference was found between the evaluators. It showed that the reliability of self-assessment, peer-assessment and teacher assessments were within acceptable limits.

As is seen in the study results, students should be provided with more opportunities to assess their own works and works of their peers; thus, they should be encouraged to make use of high-level thinking skills such as critical thinking and problem-solving. In this study, a different teaching method called case method was used in order to provide the students with the opportunity to use what they have learned in their daily life, and hence, help them internalize what they have learnt and turned them into a part of permanent learning. Also, the students were asked to make use of alternative assessment skills such as self-assessment and peer-assessment, and thus, the effort made by the students to understand the learning processes deeply was revealed at the end of the study.

The findings of the study show that there should be more space for activities to develop high-level thinking skills such as discussion, critical thinking, and problem-solving which support students' self-assessment and peer-assessment skills. It should be given importance to provide the students with these skills at an early age and to educate individuals who can think scientifically. In addition to the case studies conducted to improve self-assessment and peer-assessment skills, different practices such as problem-based learning and project-based learning should be included more in the curriculum. The interdisciplinary link should be established to contribute to the more effective implementation of curricula.

Choosing the teaching methods appropriate to the level of the students can enable the students to use their self-assessment and peer assessment skills more efficiently. By taking into account the characteristics of student development, appropriate assessment criteria should be determined together with the students to learn the subject. For this purpose, students should have more information about alternative assessment methods. Students should be given performance tasks for self-assessment and peer-assessment, and they should take responsibility for and develop an awareness of their learning.

In this study, the reliability of the rater types in Turkish, social sciences and music lessons was investigated based on an interdisciplinary approach. In different studies, course types and grade levels can be changed, and all teacher assessments can be made by the same teacher as well. The results of such a study can reduce the sources of error that would interfere with the comparison between lessons. In the study, the size of the study group was determined to be 30, but similar studies can be repeated on larger groups of students. The reasons for the low reliability values obtained in music lesson in this study can be examined in detail in different studies.

REFERENCES

Alıcı, D. (2010). Öğrenci performansının değerlendirilmesinde kullanılan diğer ölçme araç ve yöntemleri. S. Tekindal (Ed.), *Eğitimde ölçme ve değerlendirme* (2. Baskı) içinde (ss. 127-168). Ankara: Pegem Akademi.

Atılgan, H. (2005). Genellenebilirlik kuramı ve puanlayıcılar arası güvenirlik için örnek bir uygulama. *Eğitim Bilimleri ve Uygulama Dergisi*, 4(7), 95-108. Retrieved from http://www.ebuline.com/pdfs/7Sayi/7_6.pdf

Bahar, M. (2006). *Fen ve teknoloji öğretimi*. Ankara: Pegem A Yayıncılık.

Bahar, M., Nartgün, Z., Durmuş, S., & Bıçak, B. (2008). *Geleneksel-alternatif ölçme ve değerlendirme öğretmen el kitabı*. Ankara: Pegem A Yayıncılık.

Ballantyne, R., Huges, K., & Mylonas, A. (2002). Developing procedures for implementing peer assessment in large classes using an action research process. *Assessment and Evaluation in Higher Education*, *27*(5), 427-441. doi: 10.1080/0260293022000009302

Börkan, B. (2017). Akran değerlendirmesinde puanlayıcı katılığı kayması. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi*, *8*(4), 469-489. doi: 10.21031/epod.328119

Boud, D. (1986). *Implementing student self-assessment*. Sydney: Higher Education Research and Development Society of Australasia.

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

Büyükkıdık, S., & Anıl, D. (2015). Performansa dayalı durum belirlemede güvenirliğin genellenebilirk kuramında farklı desenlerle incelenmesi. *Eğitim ve Bilim*, *40*(177), 285-296. doi: 10.15390/EB.2015.2454

Çeçen, M. A. (2011). Türkçe öğretmenlerinin seviye belirleme sınavı ve Türkçe sorularına ilişkin görüşleri. *Mustafa Kemal Üniversitesi Sosyal Bilimler Enstitüsü Dergisi, 8*(15), 201-211. Retrieved from http://www.acarindex.com/dosyalar/makale/acarindex-1423909379.pdf

Cihanoğlu, M. O. (2008). *Alternatif değerlendirme yaklaşımlarından öz ve akran değerlendirmenin işbirlikli öğrenme ortamlarında akademik başarı, tutum ve kalıcılığa etkileri* (Yayımlanmamış doktora tezi). Dokuz Eylül Üniversitesi Eğitim Bilimleri Enstitüsü, İzmir.

Cram, B. (1995). Self-assessment: From theory to practice. Developing a workshop guide for teachers. In G. Brindley (Ed.), *Language assessment in action* (pp. 271-350). Sydney: National Centre for English Language Teaching and Research, Macquerie University.

Doğan, C. D., & Anadol, H. Ö. (2017). Genellenebilirlik kuramında tümüyle çaprazlanmış ve maddelerin puanlayıcılara yuvalandığı desenlerin karşılaştırılması. *Kastamonu Eğitim Dergisi*, *25*(1), 361-372. Retrieved from https://dergipark.org.tr/tr/pub/kefdergi/issue/27737/309180

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

Falchikov, N. (2001). *Learning together; Peer tutoring in higher education*. London: Routledge-Falmer.

Farrokhi, F., Esfandiari R., & Dalili, M. V. (2011). Applying the many-facet rasch model to detect centrality in self-assessment, peer-assessment and teacher assessment. *World Applied Sciences Journal*, *15*(Innovation and Pedagogy for Lifelong Learning), 70-77. Retrieved from https://pdfs.semanticscholar.org/dd21/ba5683dde8b616374876b0c53da376c10ca9.pdf

Farrokhi, F., Esfandiari R., & Schaefer, E. (2012). A many-facet rasch measurement of differential rates severity/leniency in three types of assessment. *JALT Journal*, *34*(1), 79-102. Retrieved from https://pdfs.semanticscholar.org/d79d/75e55050f9b977ffecd079ba5aadcdc10443.pdf?_ga=2.184016357.2134357192.1569916691-1527179006.1569916691

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). Rasgele veriler üzerinde genellenebilirlik kuramı ve klasik test kuramı'na göre güvenirliğin karşılaştırılması. *Education and Science*, *36*(162), 225-234.

Güler, N., Kaya-Uyanık, G., & Taşdelen-Teker, G. (2012). *Genellenebilirlik kuramı*. Ankara: Pegem Akademi.

İşman, A., & Eskicumalı, A. (2003). *Eğitimde planlama ve değerlendirme* (4. Baskı). İstanbul: Değişim Yayınları.

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. doi: 10.15640/jehd.v4n2a22

Kurudayıoğlu, M., Şahin Ç., & Çelik, G. (2008). Türkiye'de uygulanan Türk edebiyatı programındaki ölçme ve değerlendirme boyutu uygulamasının değerlendirilmesi: Bir durum çalışması. *Ahi Evran Üniversitesi Kırşehir Eğitim Fakültesi Dergisi, 9*(2), 91-101. Retrieved from http://kefad.ahievran.edu.tr/InstitutionArchiveFiles/f44778c7-ad4a-e711-80ef-00224d68272d/d1a3a581-af4a-e711-80ef-00224d68272d/Cilt9Sayi2/JKEF_9_2_2008_91_101.pdf

Kutlu, Ö., Doğan, D., & Karakaya, İ. (2008). *Öğrenci başarısının belirlenmesi, (performansa ve portfolyoya dayalı durum belirleme)*. Ankara: Pegem Akademi.

Matsuno, S. (2009). Self-, peer-, and teacher-assessments in Japanese university EFL writing classrooms. *Language Testing*, *26*(1), 75-100. doi: 10.1177/0265532208097337

McMillan, H. J. (2015). *Sınıf içi değerlendirme.* (Çev: A. Arı). Ankara: Pegem A Yayıncılık.

Milli Eğitim Bakanlığı. (2013). *İlköğretim Türkçe 3 öğretmen kılavuz kitabı*. Ankara: Milli Eğitim Bakanlığı.

Milli Eğitim Bakanlığı. (2017a). *İlkokul hayat bilgisi öğretmen kılavuz kitabı 3. sınıf*. Ankara: Milli Eğitim Bakanlığı.

Milli Eğitim Bakanlığı. (2017b). *İlköğretim müzik 4 öğretmen kılavuz kitabı*. Ankara: Milli Eğitim Bakanlığı.

Mistar, J. (2011). A study of the valıdıty and relıabılıty of self-assessment. *Teflin Journal*, *22*(1), 45-58. Retrieved from http://journal.teflin.org/index.php/journal/article/viewFile/18/20

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*(2), 395-409. doi: 10.12738/estp.2017.2.0098

Osterman, K. F., & Kottkamp, R. B. (1993). *Reflective practice for educators: Improving schooling through professional development*. Newbury Park, CA: Corwin Press.

Race, P. (2001). *A briefing on self, peer and group assessment*, Retrieved from https://blogs.shu.ac.uk/teaching/files/2016/09/id9_briefing_on_self_peers_and_group_assessment_snas_901.pdf

Salmaner, R. (2015). *Yazma becerilerinin değerlendirilmesinde öz akran ve öğretmen puanlarının çok yüzeyli rasch ölçme modeliyle incelenmesi* (Yayımlanmamış yüksek lisans tezi). Gazi Üniversitesi Eğitim Bilimleri Enstitüsü, Ankara.

Sarıtaş, S. (2015). *Problem çözme becerilerinin değerlendirilmesinde öz, akran ve öğretmen puanlarının çok yüzeyli Rasch ölçme modeli ile incelenmesi* (Yayımlanmamış yüksek lisans tezi). Gazi Üniversitesi, Eğitim Bilimleri Enstitüsü, Ankara.

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

Stiggins, J. R. (1997). *Student-centered classroom assessment*. New Jersey, NJ: Merrill, Prentice Hall, Inc.

Stiggins, R., & Chappius, J. (2005). Using student-involved classroom assessment to close achievement gaps. *Theory into Practıce, 44*(1), 11–18. Retrieved from https://www.jstor.org/stable/3496986?seq=1#metadata_info_tab_contents

Sünbül, A. M. (2007). *Öğretim ilke ve yöntemleri*. Konya: Çizgi Kitabevi.

Taşdelen-Teker, G., & Güler, N. (2019). Thematic content analysis of studies using generalizability theory. *International Journal of Assessment Tools in Education*, *6*(2), 279-299. doi: 10.21449/ijate.569996

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. Retrieved from https://j-humansciences.com/ojs/index.php/IJHS/article/view/4155/2035

Tekindal, S. (2014). *Okullarda ölçme ve değerlendirme yöntemleri* (4. Basım). Ankara: Nobel Akademik Yayıncılık.

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), 149-169. doi: 10.1080/713611428

Turgut, M. F., & Baykul, Y. (2015). *Eğitimde ölçme ve değerlendirme* (7. Baskı). Ankara: Pegem Akademi.

Wilson, J., & Jan, W. L. (1993). *Thinking for themselves: Developing strategies for reflective learning*. Australia: Eleanor Curtain Publishing.

Woolfolk, A. (2002). *Educational psychology*. New York, NY: Pearson.

Yaşar, M. (2017). Ölçme ve değerlendirmenin önemi. S. Tekindal (Ed.), *Eğitimde ölçme ve değerlendirme* (5. Baskı) içinde (ss. 2-8). Ankara: Pegem Akademi.

Yıldıztekin, B. (2014). *Klasik test kuramı ve genellenebilirlik kuramından puanlayıcılar arası tutarlılığın farklı yöntemlere göre karşılaştırılması* (Yayımlanmamış yüksek lisans tezi). Hacettepe Üniversitesi Eğitim Bilimleri Enstitüsü, Ankara.