KURAM VE UYGULAMADA EĞİTİM BİLİMLERİ EDUCATIONAL SCIENCES: THEORY & PRACTICE

Received: February 24, 2015
Revision received: April 5, 2016
Accepted: November 3, 2016
OnlineFirst: December 30, 2016

www.estp.com.tr
DOI 10.12738/estp.2017.2.0098 • April 2017 • 17(2) • 395–409

Research Article

Reliability of Scores Obtained from Self-, Peer-, and Teacher-Assessments on Teaching Materials Prepared by Teacher Candidates

Funda Nalbantoğlu Yılmaz
Nevşehir Hacı Bektaş Veli University

Abstract

This study aims to determine the reliability of scores obtained from self-, peer-, and teacher-assessments in terms of teaching materials prepared by teacher candidates. The study group of this research constitutes 56 teacher candidates. In the scope of research, teacher candidates were asked to develop teaching material related to their study. One class teacher and two teacher candidates (peers) randomly selected from the class took part in rating the teaching materials prepared by each teacher candidate. In addition to teacher- and peer-assessments, all teacher candidates who had prepared materials assessed their own material using the same criteria. The form used by the teacher, individuals, and two peers for rating teaching materials contains 10 criteria. Generalizability theory (G-theory) was used to determine the reliability of scores obtained from the self-, peer-, and teacher-assessments related to the teaching materials that teacher candidates had prepared. According to the results of the research, an insignificant difference between rater types was determined, and the reliability of scores obtained from self-, peer-, and teacher-assessments on teaching materials is within acceptable limits.

Keywords

Self-assessment • Peer assessment • Rater • Generalizability theory • Reliability

1 **Correspondence to:** Funda Nalbantoğlu Yılmaz (PhD), Department of Educational Sciences, Nevşehir Hacı Bektaş Veli University, Nevşehir 50300 Turkey. Email: fundan@nevsehir.edu.tr

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

Expected learning outcomes required by the 21st-century education system are critical thinking, problem solving, communication skills, active learning, participation, social interaction, self-monitoring, and self-regulation, as well as providing students with these skills (Greeno, Collins, & Resnick, 1996, p. 135, as citied in Sung, Chang, Chang, & Yu, 2010; Hack & Kendall, 2005; Sharma, 2000). This change in direction for learning outcomes affects teaching methods and in-class measurement and evaluation processes (Sung et al., 2010). The process of in-class measurement and evaluation, beyond just measuring knowledge, should have students actively participate in the assessment process. Students' ability to use knowledge when there is an issue, not just their knowledge, should be measured. In line with this, in-class measurements and evaluations should not only be knowledge-oriented, but also oriented towards participation, sharing, and social interaction.

Looking at the higher education programs in Turkey, it is seen that measurement and evaluation is often conducted on whether or not the students have gained the relevant program behaviors and it is conducted mostly in the way the lecturer governs. However, as the National Qualifications Framework for Higher Education in Turkey (Council of Higher Education [Yükseköğretim Kurulu], 2010) states, there are a number of competencies that individuals need to earn, such as critically evaluating knowledge and skills related to one's field, being able to identify and direct learning needs, and taking responsibility as an individual and a team member. Thus, involving not only the class teacher but also the students in the assessment process has become more and more a widespread need. However, making formative assessments for determining students' learning shortcomings concerning vocational skills is also required in higher education. This is important because it determines learning shortcomings related to skills and provides feedback concerning students' strengths and weaknesses related to the skills they are to gain. In this respect, self- and peer-assessments are important in higher education programs, as they provide professional knowledge and skills by giving feedback on strengths and weaknesses related to a skill, make one responsible for their own learning as an individual or as a team member, provide individuals with active participation in assessment by supporting cooperation and communication.

Self-assessment is when individuals judge their own learning. It is a method that provides students with an active role in determining their own learning. Self-assessment constitutes an important part of students' learning by contributing to their behavioral development. On the other hand, peer-assessment is when a student's work is assessed by their friends in accordance with specific criteria (Bound, Cohen, & Sampson, 1999, p. 103, as citied in Kutlu, Doğan, & Karakaya, 2010). Peer-assessment is a method for increasing responsibility among students. In peer-assessments, students also learn how their works are graded while grading their friends (Poon, McNaught, Lam, & Kwan, 2009).

Self/peer assessments play an important role in teacher-training programs. One important objective of teacher training and education science is to teach teacher candidates how to do assessments. Studies (Hughes & Large, 1993, p. 302, as citied in Cheng & Warren, 1999; Koç, 2011) show that self/peer assessments improve teacher candidates' academic performances and also support their occupational life by providing them with the experience of assessment. In teacher-training programs, having individuals perform self/peer assessments of their performances within the scope of the courses they take makes them familiar with these forms and provides them with the skill of using self/peer assessments as a prospective teacher. In addition to these benefits, teacher candidates' self/peer assessments on a topic can also help them see their strengths and weaknesses related to their performance. When students are supported in participating in the process of self/peer assessments, their motivation for the course and its applicability are also thought to increase.

Beside the benefits mentioned above, bias in self/peer assessments comes to mind as a disadvantage of these applications. Because the issue in self/peer assessments is grading one's own performance and a friend's performance, respectively, the thought appears that grading could not possibly be objective. In fact, this is why self/peer-assessment usage is rare. Falchikov and Goldfinch (2000) said that self/peer assessments are rarely used by teachers when assessing in class because of bias and the scores' low reliability. However, studies on scoring reliability of teacher-, self-, and peer-assessments in the literature are seen to be limited (Donnon, McIlwrick, & Woloschuk, 2013; Farrell, Mariotto, Conger, Curran, & Wallander, 1979; Farrokhi, Esfandiari, & Schaefer, 2012; Gözen & Deniz, 2016; Jackson, 2014; Karakaya, 2015; Kraiger, 1989; Sung et al., 2010; Stefani, 1994; Topping, 1998; Webb, Shavelson, Kim, & Chen, 1989; Zhang, Johnston, & Bağcı Kılıç, 2008). For these reasons, examining the reliability of scores individuals receives from themselves, their peers, and their teachers are also seen to be necessary.

The reliability of scores obtained by self-, peer-, and teacher-assessments related to student performance is an indicator of the consistency of scores given by different individuals for the same situation. In this context, consistency between raters can be defined as grading that does not change from one rater to another. Similarity among scores for self-, peer-, and teacher-assessments of the same situation indicates a high reliability for the scores obtained from these different raters. Reliability is a good question when self-, peer-, and teacher-assessments are performed on the same issue because of the possible bias in self/peer assessments. In this respect, this study aims to determine whether scores for the same situation differ according to being self-, peer-, or teacher-assessments, as well as to examine the reliability of scores obtained from ratings done by oneself, two peers, and the class teacher using G-theory in terms of teaching materials prepared by the teacher candidates in the scope of a course.

The study also aims to determine the variation of reliability according to different numbers of raters by making a decision study using G-theory. In line with these objectives, the study seeks answers to the following questions:

1. Is there a difference among the scores obtained when rating is done by oneself, by two peers, and by the class teacher over teaching materials developed by teacher candidates within the scope of a course?
2. What is the reliability of scores obtained from these ratings (self, two peers, and teacher)?
3. How does the reliability of the scores change when they are obtained from the individual himself/herself, peer 1 and peer 2 as a couple with the teacher?
4. How do the conditions of rater-type (self, peer, teacher) affect reliability?
5. For the decision study, how does the reliability coefficient change in terms of the number of raters?

Method

As this study attempts to determine the reliability of scores obtained from self-, peer-, and teacher-assessments for teaching materials that teacher candidates all prepared on the same topic, it is a descriptive research.

Study Group

The study group consists of 56 students in a pedagogical formation training of a science group from Nevşehir Hacı Bektaş Veli University. Thirty-one (55.4%) of the teacher candidates who participated in the research are female and 25 (44.6%) of them are male. Candidates had graduated from an institution of higher education and were continuing to the second semester of a certificate program on pedagogical formation training. The principle of accessibility was taken into account in determining the study group. A class teacher was assigned for assessing each of the teacher candidates' teaching materials prepared in the study group. The teacher has experience in instructional technology and material development, with specialization in measurement and evaluation. In addition to the class teacher, one individual (self) and two peers engaged in the research as raters. For peer assessing the materials prepared by the teacher candidates, two students were selected randomly from the class. These two students prepared materials within the scope of this course but were not included in the study group. In order to keep the number of students in the study group, no more than two peers were engaged; reliability, however, was examined as a result of changes in the number of peers and teachers. Aside from

the teacher and two peers, each teacher candidate who had prepared material self-graded their own material. Also, the teacher candidates in the group had already been taught the knowledge and skills for self/peer assessments by the researcher within the scope of the course, "Measurement and Evaluation," which had been taken before "Instructional Technology and Material Development."

Data Collection

Research was carried out during the course, "Instructional Technology and Material Development." The teacher candidates were asked to prepare visual teaching materials (posters, models, etc.) related to science that could be used when they become a teacher. Teacher candidates' prepared materials weren't restricted to the behavior to be learned, the subject, or the grade level. At the end of the material development process, teacher candidates brought the teaching materials they had developed to class and gave information on the characteristics of the material, the behavior to be learned by the material, the subject, and grade level. After each teacher candidate presented their prepared teaching material, the class teacher and two randomly selected teacher candidates (peers) from class assessed the teaching materials simultaneously using the same rating scale. After each presentation, teacher candidates who prepared the teaching material self-assessed their own material using the same form as the class teacher and peers. The form used for rating was introduced to the class beforehand. It had been examined for any unclear criteria, and a sample grading had also been performed.

The rating scale used for scoring of the teaching materials was prepared by reviewing the related literature (Demirel & Altun, 2012; Seferoğlu, 2010; Uzunboylu, 2011). Criteria were presented to two experts in measurement and evaluation for their opinions in terms of the clarity and expediency of the criteria. Some modifications were made in accordance with the experts' opinions. The rating scale, which had been prepared in accordance with the literature review and expert opinion, consists of 10 criteria. In relation to the prepared teaching materials, these criteria are appropriate to the defined target behavior, subject, and selected grade levels. The criteria examine if the teaching material provides students with the opportunity to apply and practice the target behavior, as well as if the material is interesting, creative, visual, and easy to use. The criteria for the rating scale are scored as yes (2), partially (1), and no (0), depending on if the material has these features. The analysis made within the scope of the study provided the reliability coefficient of its design, as well as evidence of the rating scale's reliability.

Data Analysis

There are many methods for determining the reliability of assessments obtained for the same situation from different raters. In this study, G theory was used to determine

the reliability of scores obtained from self-, peer-, and teacher-assessments on the teaching materials prepared by teacher candidates. G theory is a statistical theory used to determine the reliability of behavioral measures (Shavelson & Webb, 1991). In classical test theory (CTT), reliability that depends on a particular and single source of error (the measurement tool, raters, time, etc.) is calculated separately with distinct formulas. G theory, in comparison with CTT, can be considered as an extension of CTT in that it takes multiple sources of error interactions into account at the same time when determining reliability, which is its most important advantage. Also, G theory provides superiority to CTT by allowing the conditions of variables in different situations to be modified, allowing the study's reliability to increase (Brennan, 2001; Shavelson & Webb, 1991). In this respect, G theory was used in the study to determine the reliability of scores obtained from self-, peers-, and teacher-assessments comparing the same criteria on teaching materials prepared by individuals within the scope of a course.

The reliability coefficient (generalizability [G] and index of dependability/phi [Φ]) can be calculated using G theory. These coefficients, similar to calculating the reliability coefficient in CTT, take and interpret values between 0 and 1, just as the reliability coefficient in CTT. *The G coefficient* is used for relative decisions. Just like the ratio of variance in true scores to variance in observed scores is defined in CTT, *G coefficient* is the ratio of universe score variance to observed score variance. Observed score variance is formed from universe score variance and relative error variance (σ^2 [δ]). Φ is used in absolute decisions and is the ratio of universe score variance to the sum of universe score variance and absolute error variance (σ^2 [Δ]) (Shavelson & Webb, 1991).

In the study, the teaching materials prepared by each teacher candidate were assessed one-by-one under 10 predefined criteria by the individual presenting the material, two peers, and the class teacher. These assessments have been considered in terms of the facet of rater type. In these cases, rater type shows the differences in individuals' assessment qualities. The conditions of rater type interact with the other effects (teacher candidates and criteria) in the study. Therefore, rater types are crossed with teacher candidates and criteria. There are studies in the literature on individual self-ratings on a variety of topics where they have been handled under the similar conditions of measurement as peers and teachers/specialists, and due to this conditions of measurement interaction with other facets, analysis is based on G theory with designs where all facets are crossed examined (Farrell et al., 1979; Gözen & Deniz, 2016; Jackson, 2014; Kraiger, 1989; Sung et al., 2010; Webb et al., 1989; Zhang et al., 2008). In addition to the support from the literature, self-assessment were drawn from the data set, and two situations in which students are objects of the measurement and analysis of the self-assessment uses the crossed design (s x r x c, where s = 56 students, r is self-rating, c is 10 criteria) and nested design ([r : s] x c) with students. The results obtained from the comparison are given in Table 1.

Table 1
Inclusion of Self-Assessed Scores Crossed and Nested Format into the Design

s x *r* x *c*			(*r* : *s*) x *c*		
Sources of Variance	σ²	%	Sources of Variance	σ²	%
s	0.06557	30.3	*s*	0.06557	30.3
r	……	….	*r* : *s*	….	…
c	0.01377	6.4	*c*	0.01377	6.4
sr	……	….	*sc*	0.13703	63.3
sc	0.13703	63.3	*rc* : *s*	….	…
rc	…….	….			
src	…….	….			
G = 0.83 *Φ* = 0.81			*G* = 0.83 *Φ* = 0.81		

Note. *s* is student/teacher candidate; *r* is self-ratings, *c* is criteria and $n_s = 56$; $n_r = 1$; $n_c = 10$.

In both cases, the obtained coefficients were found to be similar, $G = 0.83$; $\Phi = 0.81$. In this respect, self-rating was confirmed to be able to be crossed with other effects, in addition to the support of the literature. As such, the *s* x *r* x *c* design was used where *s* is student/teacher candidate, *r* is rater type, and *c* is criteria and the teacher candidates were determined as objects of measurement. *G* and *Φ*, the reliability coefficients, are calculated as follows for the *s* x *r* x *c* design formed for this research.

The raters, the criteria in the measuring tool, and their interactions that cause

$$G = \frac{\sigma^2(\tau)}{\sigma^2(\tau) + \sigma^2(\delta)} = \frac{\sigma_s^2}{\sigma_s^2 + \frac{\sigma_{sr}^2}{n_r} + \frac{\sigma_{sc}^2}{n_c} + \frac{\sigma_{src}^2}{n_r n_c}} \quad (1)$$

$$\Phi = \frac{\sigma^2(\tau)}{\sigma^2(\tau) + \sigma^2(\Delta)} = \frac{\sigma_s^2}{\sigma_s^2 + \frac{\sigma_r^2}{n_r} + \frac{\sigma_c^2}{n_c} + \frac{\sigma_{rc}^2}{n_r n_c} + \frac{\sigma_{sr}^2}{n_r} + \frac{\sigma_{sc}^2}{n_c} + \frac{\sigma_{src}^2}{n_r n_c}} \quad (2)$$

differences in measurements are the source of errors that affect reliability. For example, as raters' variations decrease when measuring, errors also decrease and reliability increases. Thus the reliability coefficient obtained in terms of the scope of this study presents evidence for both the reliability of scores for different raters and of the measurement tool. However, the reliability of scores from different raters has been highlighted in the analysis made under the study's aim. The study was conducted in two stages. In the first stage, the generalizability study was performed under the *s* x *r* x *c* design and the variance components were estimated. By using the estimated variance components, it was determined whether the scores obtained from the self-, two peers-, and teacher-assessments were different. In the second stage, the variance components obtained in the first stage were used to calculate the reliability coefficients (*G* and

Φ) for the scores obtained from self-, peer-, and teacher-assessments. In the *G*-facet analysis, each condition of rater-type (self, two peers, and teacher) was excluded from analysis and the differences in *G* and *Φ* were examined. In the decision study, reliability was examined by changing the number of raters. To examine the effect of different numbers of peers and teachers on reliability, the condition of *self* was excluded by using the reduction tab of the analysis program (EduG program). With multiple peers and the teacher available as the conditions of rater types, the variable was randomly assigned. EduG was used for all these calculations.

Findings

Estimated variance components for the *s* x *r* x *c* design is reported in Table 2. As mentioned in Table 2, the *s* x *r* x *c* design has seven sources of variance components.

Table 2
Estimated Variance Components for the s x r x c *Design*

Sources of Variance	Sum of Squares	*df*	Mean of Squares	Variance (σ^2)	%
Students (*s*)	146.29598	55	2.65993	0.05852	23.6
Rater Types (*r*)	4.87991	3	1.62664	0.00138	0.6
Criteria (*c*)	32.37545	9	3.59727	0.01214	4.9
sr	35.09509	165	0.21270	0.00828	3.3
sc	116.94955	495	0.23626	0.02659	10.7
rc	20.84777	27	0.77214	0.01147	4.6
src	192.92723	1485	0.12992	0.12992	52.3
Total	549.37098	2239			100%

Among the sources of variance given in Table 2, the source of variance belonging to the teacher candidates explains 23.6% of the total variance. A large variance component related to teacher candidates points out those teacher candidates who differ from each other in terms of skills with regard to creating teaching materials. This is actually a desirable situation. In a sense, it shows that teacher candidates' individual differences with regard to the related skill can be revealed.

The variance component for rater types ($\sigma^2_r = 0.00138$) explains 0.6% of the total variance and has the smallest value when compared to the other components. The explanatory percentage of total variance, and thus its contribution to rater types, is quite close to zero. This shows that the difference among scorings carried out on teacher candidates' materials is too low. In other words, one can say that the scores given by the self, peers, and teacher in relation to the same situation are consistent with each other. The variance component for the interaction of teacher candidates and rater types ($\sigma^2_{sr} = 0.00828$, 3.3%) has the second smallest variance. This shows that the difference originating from common effect of the teacher candidates-rater types is low, and scores related to teacher candidates' materials don't change too much according to self-, peer-, or teacher assessments. The variance component for the interaction of

rater types and criteria (σ^2_{rc} = 0.01147, 4.6%) demonstrates that assessments carried out according to rater type differs a little bit from one criteria to another.

The variance component for criteria (σ^2_c = 0.01214) explains 4.9% of the total variance. This indicates that the effect criteria have is somewhat different. The variance component for the interaction of teacher candidates and criteria (σ^2_{sc} = 0.02659) explains 10.7% of the total variance, which indicates that the performance of teacher candidates differed somewhat from one criteria to another.

The residual variance component (σ^2_{src} = 0.12992, 52.3%) includes three-way interactions between teacher candidates, rater types, criteria and/or unmeasured variance sources. Having a large result in residual effect may originate from reasons such as including the variances of teacher candidates, rater type, and criteria in the residual effect; the possibility of unexplained variance sources and/or random errors could have played a part in the calculation process (some of the students being sick on the day the study was carried out, excitement, lack of attention, etc.).

Reliability coefficients for the teaching materials prepared by the teacher candidates, which belong to status of the scoring by the candidate's own (self), peers, and teacher pursuant to the 10 criteria and the rater type facet were examined dichotomously (self/teacher and peer/teacher). The reliability coefficients estimated in this way are given in Table 3.

Table 3
Reliability Estimates

	Teacher (1), Peers (2), Self (1)	Self-Teacher	Peer 1-Teacher	Peer 2-Teacher
G	0.88	0.83	0.73	0.82
Φ	0.86	0.77	0.70	0.77

As shown in Table 3, by grading the teaching materials prepared by 56 teacher candidates pursuant to the 10 criteria by individual teacher candidate who had prepared the material, two of the candidate's peers, and the class teacher, *G coefficient* acquired for relative decisions was calculated as 0.88, and *Φ* acquired for absolute decisions was calculated as 0.86. These are determined to be within acceptable limits, as criteria of reliability coefficients are considered acceptable at 0.80 (Cardinet, Johnson, & Pini, 2010; Shavelson & Webb, 1991). One can say that the reliability is high related to candidate's own self, two peers, and a teacher grading teaching materials prepared by teacher candidates pursuant to the 10 criteria.

Also, in the study, the conditions (levels) of rater-type were extracted from the data through the EduG program using the reduction tab. Thus, reliability coefficients were calculated by examining separately and dichotomously with the class teacher the candidates' self-scorings and peer-scorings related to candidates' teaching materials.

With the help of paired comparisons, it's possible to determine which condition of rater-type creates a difference in scoring. When considering that the teacher is more objective and experienced in scoring than the candidate's self or peers in this comparison, the teacher was used as the base in paired comparisons. Therefore, self- and peer-ratings weren't compared to each other; which one of these scorings differed more than the teacher's scorings of the teacher was attempted for determination between the self-ratings and the two-peer ratings. As can be seen in Table 3, G and Φ related to scorings of the second peer-teacher, and self-teacher are similar. But when reliability coefficients related to scorings of first peer-teacher are examined, one sees that the estimated G and Φ are less than the self-teacher, and second peer-teacher comparison scorings. From this point of view, one can say that the second peer- and self-ratings are more similar to the teacher with regard to scoring behavior, and the first peer tends to be more severe or lenient in scoring when compared to the self- and second peer scorings.

G-facets analysis, which provides the contribution of conditions of rater-type to the reliability, was performed and the findings are given in Table 4. G-facets analysis shows how each condition of the facet affects the reliability, as well as what the reliability will be when a related condition is not included in the design (Cardinet et al., 2010). Table 4 shows what the reliability coefficients will be for each condition of rater-type facet that is not included in the analysis.

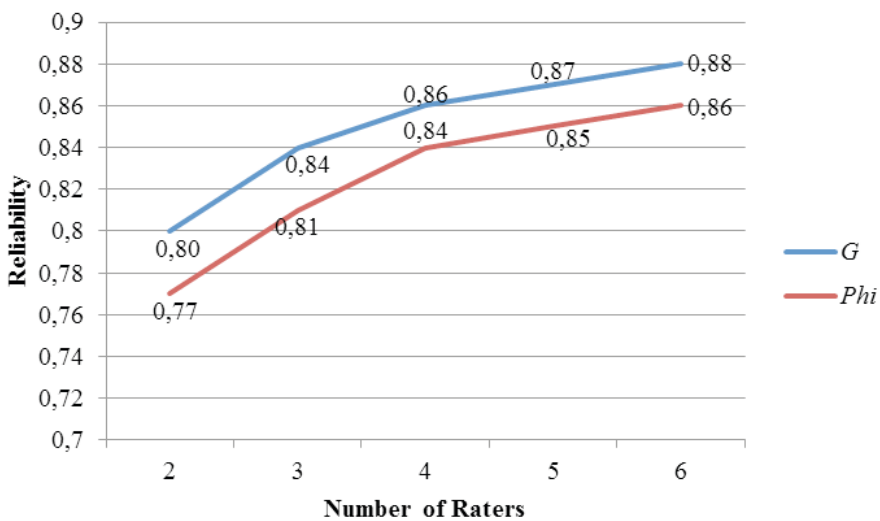
Table 4
G-*Facets Analysis of Rater Types*

Facet	Level	G	Φ
Rater Types	Self	0.84	0.81
($n_r = 4$)	Peer 1	0.87	0.83
	Peer 2	0.84	0.81
	Teacher	0.88	0.87

According to Table 4, when excluding the second peer's rating or self-rating from the analysis, the acquired reliability coefficients show a small decrease against the calculated G and Φ when examining all levels. Also, when the first peer is excluded, the calculated reliability coefficients do not show a significant difference against the calculated coefficients when examining all conditions. When the teacher is excluded from the analysis, the reliability reaches its highest value against other rater-types. However, the calculated reliability coefficients when excluding the teacher from analysis do not change compared to the coefficients that are calculated when examining all conditions. Additionally, when each condition of rater type is excluded from the analysis separately, all the acquired reliability coefficients are within acceptable limits (≥ 0.80). As can be seen in Table 4, excluding each condition of rater type from the analysis didn't overtly change the reliability. This supports that no significant difference exists for the self- and peer-scores with the score the teacher gives to an individual.

According to Table 4, *G* was calculated as 0.84 and *Φ* was calculated as 0.81 for other conditions of rater-type (two peers, one teacher) when extracting the condition of self-rating from the design. Thus, what the reliability will be when self-rating scores are not used was determined in the research. In this determination, reliability coefficients are within acceptable limits. The reliability coefficients obtained when self-rating, a conditional facet of rater-type, is not included in the design (*s* x *r* x *c*) decrease *G* by 0.04 points and *Φ* by 0.05 points. These differences between the reliability coefficients are low. Also, this small difference can also be expected when using self-rating due to the number of conditions' rising from 3 to 4. This supports that individuals' self-scores don't affect the scoring, and therefore self-rated scores can be used as a variable condition along with peers' and teacher's scores in a totally "crossed" design.

In the research, a decision study was also carried out by changing the number of individuals who performed scoring, and the change in reliability coefficients was examined. In this decision study, the condition of self was extracted using the reduction tab in the program. The decision study was carried out over other rater-type conditions (peers and teacher) in this situation, analyzing the contribution that changes in the number of individuals who are scoring provides to the reliability. Changes for the estimated *G* and *Φ* in the decision study are given in Graph 1.



Graph 1. Decision study for raters (Peer and teacher rating).

As shown in Graph 1, when the number of total raters is three (one teacher and two peers), *G* is 0.84, and *Φ* is 0.81. When the number of total raters is two, *G* and *Φ* are estimated as 0.80 and 0.77, respectively. From this perspective, when the number of raters decreases, reliability is seen to decrease a little and *Φ* is no longer within acceptable limits. As such, one can say that reducing the number of raters who are

scoring less than two peers and one teacher ($n = 3$) is not convenient. On the other hand, reliability increases when the number of raters increases, according to the research data. But when examining this increase in reliability, considering the work load in increasing the number of raters, this increase (in teachers or peers) is not deemed suitable.

Discussion

The aim of the research has been to determine whether or not there is a difference between scores for an individual's project that are self- or peer-assessed compared against the same situation that has been teacher-assessed, as well as to determine the reliability of scores that given by candidates themselves, their peers and teacher for the teaching materials that teacher candidates prepared within the scope of a lesson. From the analyses performed in line with this aim, scores related to teacher candidates' materials were determined to not change when comparing the ratings from the candidate's self, peers, or the teacher; the performed scorings did not show much difference. Also, the reliability coefficients of scores acquired by self-, peer-, and teacher-assessed scorings of teaching materials in accordance with these same criteria were observed to be at an acceptable level. Based on this result, one can say that the difference among scores related to teacher candidates' self-grading, their peers' grading, and the teacher's grading of teacher candidates' teaching materials is low, and the reliability of scores given by these various observers is high. Regarding the topic related to self-, peer- and teacher-scorings, previous studies show that Stefani (1994) resulted in student's scorings being as reliable as the teacher's; Topping (1998) had the result that peer assessment is as reliable as teacher assessment; and Zhang et al. (2008) found that reliability related to self- and peer-scorings is high. However, in contrast with the findings acquired in this study, the studies of Farrell et al. (1979); Gözen and Deniz (2016); Kumandaş and Kutlu (2013); and Poon et al. (2009), which all compared scores acquired from teacher candidate-, self-, and teacher-ratings, observed differences between self-scores and teachers' scores. Karakaya (2015); Kraiger (1989); and Webb et al. (1989) determined that there are differences among self-, peer-, and expert-scoring behaviors. These differences that occurred among the studies may have arisen from: study groups' characteristics, individuals made to score without feeling any scoring anxiety, the training provided to individuals before the application, or clarity of the criteria used. Showing students how to perform scoring, having them experience it through a case study, and having the students understand the criteria are important steps for the objectivity and reliability of self- and peer-scorings. The research carried out by Dochy, Segers, and Sluijsmans (1999) also supports this. Thus, having consistency among scores acquired from self-, peer-, and teacher-ratings in studies may be linked to teacher candidates who have gained the knowledge and skills related to self- and peer-assessments in the scope of a lesson they had previously taken, as well as the clarity of the criteria in the rating scale.

This research also determined that the reliability estimations of self/teacher and peer/teacher are dichotomous with regard to scoring behaviors and that the second peer's and self are more similar to the teacher's than the first peer's. Although the research determined no major difference between the scorings, one can say that small differences arose from the first peer's tendency to make a more severe or lenient scoring compared to others. In the decision study, reliability was observed to increase when the number of raters increased when comparing the data used in the research; yet this increase does not affect reliability much when considering the labor and work load involved in increasing the number of raters. Therefore, one can say it is not convenient to increase the number of teachers or peers.

In the research, the condition of self-rating was extracted; reliability coefficients were also calculated for other conditions of the related facets (peers and the teacher). As a result of these calculations, the difference between the reliability coefficients was determined to be low for both when extracting self-ratings from the study and when self-ratings are included. This difference between the reliability coefficients can be said to be normal due to the difference in the number of conditions belonging to rater-types in both situations. Accordingly, self-rated scores can be said to cross with other variables because of scarcely any scoring differences between rater-types. In different studies that determine differences based on rater type or that determined self-rated scores as non-objective, care should be taken when using self-rated scores for individuals through crossing. Webb et al. (1989) found differences between rater types in their study. Accordingly, they calculated reliability by examining rater types separately.

Self/peer assessments occupy an important position both for their contribution with regard to having teacher candidates gain skills related to teachers' education, as well as the skill of providing consistent and objective scoring. In the study, individuals' themselves, peers, and teacher were used in scoring teaching materials developed in the Instructional Technologies and Material Design lesson. Future studies may research the differences in rater characteristics for other lessons and/or applications of the teacher training program.

References

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

Cardinet, J., Johnson, S., & Pini, G. (2010). *Applying generalizability theory using EduG.* New York, NY: Routledge.

Cheng, W., & Warren, M. (1999). Peer and teacher assessment of the oral and written tasks of a group project. *Assessment & Evaluation in Higher Education, 24*(3), 301–314. http://dx.doi.org/10.1080/0260293990240304

Demirel, Ö., & Altun, E. (Eds.). (2012). *Öğretim teknolojileri ve materyal tasarımı* [Instructional technology and material design]. Ankara, Turkey: Pegem Akademi Yayıncılık.

Dochy, F., Segers, M., & Sluijsmans, D. (1999). The use of self-, peer and coassessment in higher education: A review. *Studies in Higher Education, 24*(3), 331–350. http://dx.doi.org/10.1080/03075079912331379935

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*(6A), 23–28.

Falchikov, N., & Goldfinch, J. (2000). Student peer assessment in higher education: A meta-analysis comparing peer and teacher marks. *Review of Educational Research, 70*(3), 287–322.

Farrell, A. N. D., Mariotto, M. J., Conger, A. J., Curran, J. P., & Wallander, J. L. (1979). Self-ratings and judges' ratings of heterosexual social anxiety and skill: A generalizability study. *Journal of Consulting and Clinical Psychology*, *47*(1), 164–175.

Farrokhi, F., Esfandiari, R., & Schaefer, E. (2012). A many-facet RASCH measurement of differential rater severity/leniency in self assessment, peer assessment, and teacher assessment. *Journal of Basic and Applied Scientific Research*, *2*(9), 8786–8798.

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

Hack, C., & Kendall, G. (2005). Bioinformatics: Current practice and future challenges for life science education. *Biochemistry and Molecular Biology Education, 33*(2), 82–85. http://dx.doi.org/10.1002/bmb.2005.494033022424

Jackson, L. (2014*). Validity and rater reliability of peer and self assessments for urban middle school students* (Master's thesis, University of Wisconsin, Milwaukee, WI). Retrieved from http://dc.uwm.edu/etd/696/

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 ilişkin görüşleri [Opinions of class teacher candidates related to peer reviews in teacher applications]. *Kuram ve Uygulamada Eğitim Bilimleri*, *11*, 1965–1989.

Kraiger, K. (1989). *Generalizability theory: An assessment of its relevance to the Air Force job performance measurement project*. Brooks AFB, TX: Training Systems Division, Air Force Human Resources Laboratory. Retrieved from http://www.dtic.mil/dtic/tr/fulltext/u2/a207107.pdf

Kumandaş, H., & Kutlu, Ö. (2013). Okul öncesi öğretmen adaylarının kendi sunum becerilerine ilişkin öz değerlendirmeleri ile eğitici değerlendirmesinin karşılaştırılması [A comparison of educational evaluation and self-assessment related to preschool teacher candidates' own presentation skills]. *Eğitim Bilimleri ve Uygulama*, *12*(23), 43–55.

Kutlu, Ö., Doğan, C. D., & Karakaya, İ. (2010). *Öğrenci başarısının belirlenmesi performansa ve portfolyoya dayalı durum belirleme* [Determining student success by determining the situation based on performance and portfolios]. Ankara, Turkey: Pegem Akademi Yayınları.

Poon, W., McNaught, C., Lam, P., & Kwan, H. S. (2009). Improving assessment methods in university science education with negotiated self- and peer-assessment. *Assessment in Education: Principles, Policy & Practice, 16*(3), 331–346.

Seferoğlu, S. S. (2010). *Öğretim teknolojileri ve materyal tasarımı* [Instructional technology and material design]. Ankara, Turkey: Pegem Akademi Yayınları.

Sharma, G. L. (2000). Effectiveness of problem-solving teaching technique on the evolvement of higher level learning outcomes. *Psycho-Lingua, 30*(2), 99–105.

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

Stefani, L. A. J. (1994). Peer, self and instructor assessment: Relative reliabilities. *Studies in Higher Education, 19*(1), 69–75. Retrieved from http://dx.doi.org/10.1080/03075079412331382153

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*(1), 135–145. Retrieved from http://dx.doi.org/10.1016/j.adolescence.2009.04.004

Topping, K. J. (1998). Peer assessment between students in college and university. *Review of Educational Research, 68*(3), 249–276.

Uzunboylu, H. (Ed.). (2011). *Öğretim teknolojileri ve materyal tasarımı* [Instructional technology and material design]. Ankara, Turkey: Pegem Akademi Yayıncılık.

Webb, N. M., Shavelson, R. J., Kim, K. S., & Chen, Z. (1989). Reliability (generalizability) of job performance measurement: Navy machinist mates. *Military Psychology*, *1*(2), 91–110.

Yükseköğretim Kurulu. (2010). *Yükseköğretimde yeniden yapılanma: 66 soruda Bologna süreci uygulamaları* [Restructuring in higher education: Applications of the Bologna process in 66 questions]. Ankara, Turkey: Author.

Zhang, B., Johnston, L., & Bağcı-Kılıç, G. (2008). Assessing the reliability of self- and peer-rating in student group work. *Assessment & Evaluation in Higher Education*, *33*(3), 329–340.