Journal of Educational Technology & Online Learning
Volume 4 | Issue 4 | 2021
http://dergipark.org.tr/jetol

# Online peer assessment in teacher education

Fatma Betül KURNAZ [a] *

[a] Karabuk University, Turkey

Suggested citation: Kurnaz, F. B. (2021). Online peer assessment in teacher education. *Journal of Educational Technology & Online Learning*, *4*(4), 835-853.

**Article Info**

*Keywords:*

Online peer assessment
Higher education
Teacher education
Online learning

Research Article

**Abstract**

It has become necessary to monitor the change in learners' skills during the education carried out in electronic environments. In this study, pre-service teachers made a presentation in their teaching practice, and a formative assessment was given to ensure active participation of the observer pre-service teachers in the online peer assessment process. The observer pre-service teachers were asked to evaluate their peers' performances using a rubric. Based on the quantitative data collected and analyzed, questions about the experiences of the participants were created, and the opinions of the participants were obtained through e-mail. The research findings were obtained through the sequential explanatory mixed method. The study revealed that the observer pre-service teachers could evaluate different performances consistently. The research also showed that the validity of the assessments was significantly low especially in the evaluation of low and medium level performances. The qualitative findings confirmed the quantitative findings.

## 1. Introduction

Teaching is a profession equipping students with certain knowledge and skills using various methods and techniques. Thus, pre-service teachers need assistance in planning of teaching, classroom management, evaluation of teaching, and choosing the appropriate teaching methods (Lemlech, 1995). They should also be supported in terms of communication skills, classroom discourse, and professional motivation (Tomlinson, 1998). Pre-service teachers' active involvement in the assessment processes may help them acquire professional and social skills (Carr, 2020; Taskiran, 2021; Zeng, 2020).

Since December 2019, when the Covid-19 pandemic broke out, daily life has undergone an inevitable change, and education has been carried out online in most countries to ensure social distancing. As a result, the structure of education and training processes have remarkably changed, which has necessitated the monitoring of the change in learners' skills using different methods. In particular, it intends to reveal the effectiveness of pre-service teachers' peer assessment of presentation performances in an online course.

### *1.1. Participatory Assessment Processes in Teacher Education*

Most teachers perceive assessment merely as grading. However, assessment involves much more than this (Lu & Law, 2012). Learners' assessment of their own or peers' work, use of a rubric for this purpose, and their active participation in the assessment processes make assessment a part of the learning process. If pre-service teachers acquire this culture of assessment during their pedagogical education, they may employ

* Corresponding author: Karabuk University, Turkey.
e-mail address: betulkurnaz@karabuk.edu.tr
This study was partly presented as a proceeding at the 1st International Conference on Educational Technology and Online Learning Conference held between 22-24 September 2021.

Doi: http://doi.org/10.31681/jetol.987902
Received 27 Aug 2021; Revised 11 Dec 2021; Accepted 12 Dec 2021
ISSN: 2618-6586 . 

diverse teaching practices in the future. Self-assessment and peer assessment during pedagogical training may help teacher trainees understand the purpose and function of these assessments. Seeing the results of these assessments during their own learning may allow them a deeper insight into their teaching practice (Li et al., 2020; Reinholz, 2016; Zeng, 2020). Their perspectives on assessment can also be enriched through such experiences. Computer-mediated peer assessments are less discouraging for students because negative interactions are more limited in this mode, so they are more beneficial than peer assessments carried out in the classroom (Li et al., 2020).

### *1.2. Peer Assessment*

Peer assessment is the evaluation of the performance or a product of a student by other students of the same status using predetermined criteria. Since peer assessment involves the evaluation of a certain performance, it enables the assessor to reflect on the skills required by performance (Roscoe & Chi, 2007), reinforces students' knowledge on the subject (Black et al., 2003), improves autonomy, and provides guidance for students to realize their own mistakes (Topping, 2009). Peer assessment also provides cognitive (Nelson & Schunn, 2009) and pedagogical (Strijbos et al., 2010) benefits to the learner. Indeed, many studies in the literature investigated the reliability and validity of peer assessments (Cho et al., 2006; Chang et al., 2011; Jones & Alcock, 2014), the relationship between peer assessment and self-assessment (Reinholz, 2016), and the effect of emotional or cognitive assessments in peer assessment (Zhou et al., 2020).

If well-managed, peer assessment processes enable the assessors to identify what to assess and establish the respect framework during the assessment (Zhou et al., 2020). It is also stated in the literature that, when peer and self-assessment are used together, learning experience is enhanced (Nicol & MacFarlane-Dick, 2006) and peer assessment is perceived as encouraging by learners (Ohaja et al., 2013).

### *1.3. Objectives of the research*

Substantial evidence in the literature shows that assessment is an important part of the teaching profession and that pre-service teachers' participation in the assessment process during their pedagogical training increases their competence. As maintained earlier, peers who make the assessment evaluate performance critically. These assessments are expected to give more reliable and valid results in time. Therefore, research on how peer assessment changes in time can shed light onto peer assessment in general. In addition, because there is limited research on online peer assessments (Cheng et al., 2015; Iglesias Pérez et al., 2020; Li et al., 2020), it is hoped that this study is significant as it investigates both online peer assessment and the changes that takes place in assessment. Assessment needs to be carefully addressed since assessment quality is increasingly seen as a major element of education quality.

In this study, pre-service teachers were asked to make a presentation in their teaching practice, and the observer pre-service teachers to make online peer assessment. Their active participation was ensured by means of formative assessment approach. That is, the observer pre-service teachers were asked to evaluate their peers' performances using a rubric. In this study, based on the analysis of quantitative data, questions about the experiences of the participants were formed. The research findings were obtained using the sequential explanatory mixed method. The following questions were addressed in the study:

1. How reliable are the online peer assessments of the observer pre-service teachers?

2. To what extent do the online formative peer assessment of the observer pre-service teachers and that of the instructor differ?

3. How do the assessment of the observer pre-service teachers and that of the instructor differ according to different levels?

4. How do the observer pre-service teachers perceive the online peer assessment process?

## 2. Methodology

The study used the sequential explanatory mixed method, which is a type of mixed method. It started with the quantitative stage, and in the qualitative stage, data were collected and analyzed based on the findings obtained from the quantitative stage. In the first phase, the problem situation was defined, and the research questions were formulated. Data collection tools were prepared, and preliminary analyses on their reliability and validity were conducted (Creswell, 2014; Creswell & Clark, 2017).

While analyzing the quantitative data, salient results were determined, and significant and non-significant results were examined. The qualitative phase was designed based on these findings. First, the qualitative research questions were determined, and then, a sample was selected by purposeful sampling strategy. The extent to which qualitative results helped to explain quantitative data was also discussed.

### *2.1. Participants*

The research was conducted with a total of 56 participants studying teacher education at a state university in Turkey. Of them, 12 are male, while 44 are female. At the time the study, the participants had completed or were enrolled in undergraduate education in the child development program, which admitted senior students or graduates. In the fall semester, they completed the following must courses: Instructional Principles and Methods, Measurement and Evaluation in Education, Educational Psychology, and Introduction to Educational Science. Then, in the spring semester, the students took the applied courses, including Teaching Methods, Instructional Technologies, and Material Development. Upon completion of these, students were expected to do practice teaching, which requires preparation of a lesson on a specific subject in applied courses and demonstration of their teaching skills. This research was carried out during these applied courses.

### *2.2. Instrument*

An instrument was developed to evaluate the presentation skills of pre-service teachers in practice lessons. Online peer assessment was done using an analytical rubric. The steps followed in the creation of the rubric are summarized below.

In the development of the rubric, first, the performance reflecting the task was defined clearly. This preceded the rubric development. Behaviors to be displayed in the performance were listed. The best performance criteria that correspond to these behaviors were described, followed by the weakest performance criteria. Other criteria were determined after determining the best and the poorest performances.

Following the definition of the performance, the criteria were examined, and the related ones were combined. Then, the components of the task (using effective and appropriate body language, using the voice appropriately, etc.) were determined. It was done so in advance to increase the validity of the assessments and to make sure the students and teachers understand the same thing from the assessment criteria (Topping et al., 2000). After the performance criteria were re-organized and grouped in bands, each band was labelled (*content, compliance with spelling rules*, etc.). Then, the number and name of ratings were determined. Generally, three or five ratings are recommended depending on the student level and content (Popham, 1997). Since the task planned to be assessed in this study is not too complicated to define, a three-point assessment in one dimension and a four-point assessment in the other dimension was decided on. The necessity of using positive expressions that aim to improve the action was considered while determining the names of the bands in the rating scale (Stevens & Levi, 2013).The four-point scale included the adjectives *competent, good, almost good,* and *needs to be improved*, and the three-point scale included the adjectives of *competent, good,* and *needs to be improved*.

The rubric was evaluated by three experts from the field of educational science for clarity, understandability, and validity, and it was revised accordingly. The validity of the rubric was achieved

through content validity analysis, which focuses on the relationship between the criteria and the task and the extent to which they reflect the expected behavior. Expert opinion, literature review, and the compatibility of the criteria with the task and with each other provided evidence to content validity.

The reliability of the rubric was tested in a pilot-study, which did not involve the actual participants. This group consisted of pre-service teachers who completed their degree in geography and who received the same pedagogical training as the study group. In this group, the students were asked to make a ten-minute presentation on a topic of their choice in their field. Two experts in educational sciences, who were given training on the content of the task, the rubric, and its use, independently evaluated the performance of four pre-service teachers using the rubric. There was a 90% agreement between raters, which suggests that the inter-rater reliability was high.

*2.3. Procedure*

The pre-service teachers had received or were receiving undergraduate education in the child development program. For this reason, all the pre-service teachers had taken many courses on developmental psychology during their undergraduate education. It was decided that the content of the presentations should be related to developmental psychology, a subject all the pre-service teachers presumably knew well. Choosing different concepts related to developmental psychology (sensory development, identity development, psycho-sexual development periods, etc.), the pre-service teachers were to prepare a presentation to be delivered online. They were informed that they would use technologies such as video, PowerPoint, and jpeg during their presentation and that they were expected to make their presentations without reading from their notes, paying attention to aspects such as tone of voice.

The procedure lasted for ten weeks. During the first two weeks, training was given on the principles of effective presentations with reference to specific examples. Seven pre-service teachers made their presentations in the third week. The instructor gave reflective feedback on the performances by explaining the relationship between the criteria in the rubric and the performance. While giving feedback, the instructor shared the screen with the pre-service teachers so that they could understand how the assessment was done. The reflective feedback, or the instructor's explanations about the relationship between the criteria in the rubric and performance, constituted one stage of the training. For this reason, there was no peer evaluation of the presentations made in the third week.

The other pre-service teachers made their presentations for seven weeks, and the other pre-service teachers evaluated these performances. The instructor did not interfere with these assessments and did not provide any reflective feedback not to influence the pre-service teachers' assessments. The total number of presentations evaluated in seven weeks was 49.

The classes were held online on Microsoft Teams. The pre-service teachers uploaded their presentations to a class opened on the Google Classroom application. During the online class hour, the video recorded performances of the pre-service teachers who were assigned to that week were watched all together. All the pre-service teachers sent their assessments at the same time, right after watching the videos. That is, the participants assumed the role of both the assessor and the assessed.

Qualitative data were obtained through emails from the group, in which quantitative data were collected. Collecting qualitative data by e-mail is seen as a preferable method when all participants have access to e-mail and use it in their professional lives (James & Busher, 2006). In addition, the use of e-mail in obtaining qualitative data allows participants to explore and reconsider their insights into their evolving professional identities to move back and forth in their narratives, and reflect on and redesign their responses (Mann & Stewart, 2000). This medium was also preferred in the present study as all the participants of this study regularly use their university e-mail addresses in different activities during the classes. Four open-ended questions prepared in line with the quantitative data were sent to the participants by e-mail. The questions asked the pre-service teachers to reflect on their experiences and opinions about the online peer assessment

process, the effectiveness of the procedure, and their involvement as decision-makers. Eleven participants responded to the questionnaire and shared their responses with the researchers. The qualitative data includes the responses of these 11 participants.

### *2.4. Data analyses*

Although the pre-service teachers were required to attend all the assessment sessions, they had to be excused for one or two weeks if they had a valid excuse such as illness. In addition, since the lessons were conducted online, the pre-service teachers could not assess a few presentations due to technical problems such as unstable internet connection. For this reason, the data obtained from the pre-service teachers who could make all the assessments was analyzed. The number of pre-service teachers who showed full participation in the assessments varied across weeks. To minimize data loss, the data collected from the participants whose assessments were complete in a certain week were included in the analysis.

In data analysis, the reliability of the results was calculated by using the $G$ and $\Phi$ coefficients, and the validity was tested using Spearman correlation and Wilcoxon $t$-Test, with the instructor's assessments being a criterion.

In generalizability ($G$) theory, unlike in the classical test theory, a single reliability value can be reached by considering more than one error source at the same time. In this study, $G$ and $\Phi$ coefficients were used to analyze the data. Since all error sources are evaluated together and simultaneously in $G$ theory, it gives a comprehensive reliability coefficient. Also, it is believed that the traditional difference between reliability and validity in classical test theory can be eliminated by making reliable observations (Brennan, 2001; Shavelson & Webb, 1991). For this reason, the $G$ theory was used. Since the number of performances evaluated is more than one, the $G$ coefficient and $\Phi$ coefficient were calculated for the two-facet design ($s$ $x$ $i$ $x$ $r$), where the source of variability is items and raters. $G$ and $\Phi$ coefficients close to 1 means that the assessments are consistent.

The correlations between the assessment scores of the pre-service teachers and those of the instructor were examined using the Spearman correlation coefficient. The Wilcoxon $t$-Test was used to examine whether the two were significantly different. In this study, the Wilcoxon $t$-Test, which is a non-parametric test, was used as the number of presentations evaluated in a week was low and the parametric tests may be affected by the number of observations, which may in turn affect the generalizability of the results. These analyses provided evidence to the validity of online peer assessments.

After quantitative data analysis, qualitative data analysis was performed. Qualitative data were analyzed through thematic analysis (Braun & Clarke, 2006). Here, pre-service teachers' responses to the questions were examined in terms of which themes were associated with which situation and concepts. Thus, general themes and sub-themes were identified. A theoretical analysis approach was adopted in the qualitative data analysis. For this reason, frequency was not considered in the display of themes and sub-themes, and it was assumed that all the elements that the pre-service teachers associated with the online peer assessment process were of the same importance. The themes were intended to reflect all the data, and the responses were reviewed three times to reach reliable and valid results. Then, the themes and sub-themes were given names, and the relationship matrix was given.

### *2.5. Findings*

The findings revealed by the analysis of the quantitative and qualitative data are presented below under separate headings.

#### *2. 5. 1. Findings of quantitative data analysis*

When scores are more consistent across different raters and situations, the assessment is regarded to be more reliable (Moskal & Leydens, 2000). To investigate the reliability of online peer assessments in the study, the consistency between the scores given by the pre-service teachers was monitored in different

weeks, and *G* and *Φ* coefficients were calculated separately for each week. The findings obtained are presented in Table 1.

**Table 1.**

*G* and *Φ* coefficients for inter-rater reliability

| | Component | df | Mean Squares | Variance | Variance % | *G* | *Φ* |
|---|---|---|---|---|---|---|---|
| **Week 1** | Rater | 49 | 12.58884 | 616.85299 | -- | | |
| | Item | 10 | 0.69055 | 6.90545 | 0.0 | | |
| | Individual | 6 | 0.17801 | 1.06805 | 0.0 | | |
| | Rater x Item | 490 | 2.21644 | 1086.0555 | 78.7 | 0,82 | 0,82 |
| | Rater x Individual | 294 | 0.39658 | 116.59429 | 0.0 | | |
| | Item x Individual | 60 | 0.48915 | 29.34909 | 0.0 | | |
| | Rater x Item x Individual | 2940 | 0.47225 | 1388.4171 | 21.3 | | |
| **Week 2** | Rater | 38 | 15.13614 | 575.17349 | -- | | |
| | Item | 10 | 0.96124 | 9.61239 | 0.0 | | |
| | Individual | 6 | 0.43179 | 2.59074 | 0.2 | | |
| | Rater x Item | 380 | 1.68284 | 639.47852 | 80.8 | 0,89 | 0,89 |
| | Rater x Individual | 228 | 0.32277 | 73.59108 | 0.9 | | |
| | Item x Individual | 60 | 0.28820 | 17.29204 | 0.0 | | |
| | Rater x Item x Individual | 2280 | 0.30725 | 700.52614 | 18.1 | | |
| **Week 3** | Rater | 43 | 16.12715 | 693.46724 | -- | | |
| | Item | 10 | 0.86133 | 8.61334 | 0.0 | | |
| | Individual | 6 | 0.26161 | 1.56966 | 0.0 | | |
| | Rater x Item | 430 | 1.90555 | 819.38666 | 77.3 | | |
| | Rater x Individual | 258 | 0.43306 | 111.72904 | 0.2 | 0,88 | 0,88 |
| | Item x Individual | 60 | 0.37449 | 22.46930 | 0.0 | | |
| | Rater x Item x Individual | 2580 | 0.42977 | 1108.8034 | 22.5 | | |
| **Week 4** | Rater | 54 | 7.62215 | 411.59622 | -- | | |
| | Item | 10 | 0.70144 | 7.01440 | 0.0 | | |
| | Individual | 6 | 0.54703 | 3.28217 | 0.2 | | |
| | Rater x Item | 540 | 1.05262 | 568.41417 | 51.3 | 0,86 | 0,85 |
| | Rater x Individual | 324 | 0.53589 | 173.62692 | 4.6 | | |
| | Item x Individual | 60 | 0.35257 | 21.15419 | 0.0 | | |
| | Rater x Item x Individual | 3240 | 0.48534 | 1572.5081 | 43.9 | | |
| **Week 5** | Rater | 49 | 14.35303 | 703.29870 | -- | | |
| | Item | 10 | 1.21018 | 12.10182 | 0.0 | | |
| | Individual | 6 | 0.19688 | 1.18130 | 0.0 | | |
| | Rater x Item | 490 | 1.58643 | 777.35273 | 74.5 | 0,89 | 0,89 |
| | Rater x Individual | 294 | 0.39902 | 117.31221 | 0.0 | | |
| | Item x Individual | 60 | 0.56812 | 34.08727 | 0.2 | | |
| | Rater x Item x Individual | 2940 | 0.40145 | 1180.2763 | 25.3 | | |
| **Week 6** | Rater | 45 | 23.25077 | 1046.2845 | -- | | |
| | Item | 10 | 0.73591 | 7.35912 | 0.0 | | |
| | Individual | 6 | 0.10399 | 0.62394 | 0.0 | | |
| | Rater x Item | 450 | 1.51063 | 679.78374 | 76.1 | 0,93 | 0,93 |
| | Rater x Individual | 270 | 0.36517 | 98.59684 | 1.7 | | |
| | Item x Individual | 60 | 0.42656 | 25.59345 | 0.1 | | |
| | Rater x Item x Individual | 2700 | 0.33885 | 914.90006 | 22.0 | | |
| **Week 7** | Rater | 45 | 7.42602 | 334.17081 | -- | | |
| | Item | 10 | 0.45246 | 4.52456 | 0.0 | | |
| | Individual | 6 | 0.09750 | 0.58498 | 0.0 | | |
| | Rater x Item | 450 | 1.45136 | 653.11180 | 83.0 | 0,81 | 0,81 |
| | Rater x Individual | 270 | 0.21169 | 57.15528 | 0.0 | | |
| | Item x Individual | 60 | 0.23777 | 14.26595 | 0.0 | | |

| | | | | |
|---|---|---|---|---|
| Rater x Item x Individual | 2700 | 0.24677 | 666.27950 | 17.0 |

The consistency between the assessments of the pre-service teachers varied between 0.81 and 0.93 in different weeks (Table 1). According to the theory of generalizability, reliability and generalizability coefficients exceeding .80 are generally considered to be acceptable (Brown et al., 2004). It can be interpreted that the values obtained in this study are reliable.

Examination of the sources of variability manifested that the rater-item interaction is greater than other sources of variance in all weeks. As can be seen in Table 1, in four weeks (Weeks 2, 3, 4, and 6), the rater-individual interaction is also a source of variance; however, its weight is lower than the rater-item interaction. This indicates that inter-rater consistency is high; however, it still shows that rater-item interaction is an important source of variance in assessment. Consistency in the assessments of the pre-service teachers is a measure of reliability as it shows the relationships within. Since these analyses do not include the assessments of the instructor, the coefficients obtained show the consistency among the assessments of the pre-service teachers.

The correlations between the pre-service teachers' assessment scores and the instructor's showed significant relationships at the level of .01. The Spearman rank differences correlation coefficients (*r*) ranged from 0.411 to 0.778. While Spearman's correlation coefficient demonstrates the consistency in the decisions of the pre-service teachers, it cannot show the rate of consistency (Şencan, 2005). Since averages are not used in the calculation of this coefficient, information about the similarity or difference between the two data groups cannot be obtained. It can only provide information about data interchange. The correlation coefficient indicates how the assessment scores of the pre-service teachers and those of the instructor changed together. Indeed, the positive correlation shows that, as the score assigned by the instructor increases, the score assigned by the pre-service teachers increases, and as the score assigned by the instructor decreases, so do the scores assigned by the trainee teachers. The positive and significant correlation value suggests that the assessments change in the same direction. However, it cannot provide information on whether the scores are similar or not. Similarly, the performances rated highly by the instructor also received high scores from the pre-service teachers, and the performances not rated highly by the instructor received low scores from the pre-service teachers. Whether the scores were similar or not was examined using the Wilcoxon *t*-Test.

Examining the difference between the decisions of the pre-service teachers and the decisions of the instructor with the Wilcoxon *t*-Test can provide important information about the accuracy of the decisions. In the study, the Wilcoxon *t*-Test was used to examine whether there were significant differences between the two. The *p* values for the results obtained are presented in Table 2.

**Table 2.**

The *p* values obtained from the Wilcoxon *t*-Test regarding the difference between the assessment scores of the pre-service teachers and those of the instructor

| | **WEEK** | | | | | | |
|---|---|---|---|---|---|---|---|
| **Assessor** | **1** | **2** | **3** | **4** | **5** | **6** | **7** |
| 1 | **.091** | **.343** | .046 | **.075** | .017 | .026 | **.058** |
| 2 | **.203** | **.246** | .027 | **.063** | .018 | .041 | **.115** |
| 3 | **.462** | **.104** | .028 | .027 | .027 | .041 | **.276** |
| 4 | .027 | **.223** | .018 | .028 | .018 | .040 | .046 |
| 5 | **.833** | **.916** | .034 | .028 | **.172** | **.088** | **.089** |
| 6 | **.398** | **.667** | .018 | **.063** | **.528** | **.395** | **.075** |
| 7 | **.865** | **.854** | .046 | **.396** | **.546** | **.201** | **.207** |
| 8 | **.054** | **.345** | .018 | .034 | **.071** | .018 | **.066** |
| 9 | **.141** | **.596** | .046 | .028 | **.141** | .027 | **.080** |
| 10 | **.865** | **.916** | .034 | .043 | **.072** | .016 | **.307** |
| 11 | **.463** | **.246** | .034 | **.916** | **.279** | **.861** | **.340** |
| 12 | .042 | **.345** | .018 | .027 | .018 | .026 | **.078** |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 13 | **.091** | **.461** | .018 | .027 | .027 | **.089** | **.174** |
| 14 | **.446** | **.933** | .018 | .028 | .026 | .039 | **.113** |
| 15 | **.062** | **.078** | .018 | .028 | .027 | **.062** | **.062** |
| 16 | **.058** | **.068** | .017 | .034 | .028 | .024 | **.140** |
| 17 | **.225** | **.104** | .027 | .108 | .027 | **.066** | **.588** |
| 18 | **.144** | **.461** | .042 | .028 | .018 | .026 | **.223** |
| 19 | **.102** | **.715** | **.108** | .027 | .027 | **.058** | **.140** |
| 20 | **.202** | **.673** | .018 | .028 | .027 | **.389** | **.234** |
| 21 | **.141** | **.892** | .017 | .028 | .028 | **.290** | **.115** |
| 22 | **.072** | **.496** | **.063** | .028 | **.167** | .027 | **.066** |
| 23 | **.463** | **.273** | **.051** | **.127** | **.498** | **.089** | **.236** |
| 24 | **.307** | **.285** | .042 | .046 | .042 | **.673** | **.798** |
| 25 | **.933** | **.141** | **.149** | .042 | **.112** | **.063** | **.068** |
| 26 | .046 | **.273** | .018 | .043 | .028 | **.230** | **.063** |
| 27 | .017 | **.225** | .018 | .042 | .042 | **.058** | **.078** |
| 28 | **.173** | **.599** | .027 | .028 | **.072** | **.752** | **.108** |
| 29 | **.345** | **.605** | **.201** | **.063** | .018 | .027 | **.086** |
| 30 | **.496** | **.276** | **.351** | .046 | **.172** | **.112** | .041 |
| 31 | **.735** | **.916** | .034 | .046 | .042 | **.595** | **.107** |
| 32 | **.105** | **.136** | .028 | **.125** | .027 | .027 | **.088** |
| 33 | .018 | **.786** | .018 | .026 | **.461** | **.680** | **.138** |
| 34 | **.128** | **.713** | .017 | **.091** | .027 | **.174** | **.340** |
| 35 | **.092** | **.596** | .028 | .033 | .027 | **.068** | **.202** |
| 36 | **.237** | **.276** | .028 | .041 | **.225** | **.357** | .042 |
| 37 | **.075** | **.144** | .018 | .028 | .027 | .041 | **.114** |
| 38 | **.201** | **.786** | .018 | .028 | .026 | **.172** | **.128** |
| 39 | .028 | **.340** | **.115** | .028 | .028 | **.414** | .043 |
| 40 | .028 | **.715** | .018 | .027 | **.051** | **.599** | **.089** |
| 41 | **.115** | **.599** | .018 | .034 | **.063** | .042 | **.080** |
| 42 | **.107** | **.080** | .018 | .027 | **.340** | .017 | **.063** |
| 43 | **.088** | **.225** | .028 | .028 | .042 | **.114** | **.068** |
| 44 | **.674** | **.345** | .027 | .028 | **.115** | **.105** | **.089** |
| The number of consistent assessments | 37 | 43 | 7 | 9 | 18 | 27 | 40 |

The *p* values given in Table 2 provide information on the significance of the difference between the assessments of the instructor and those of the pre-service teachers. In this study, *p* values are expected to be higher than .05. A *p* value greater than .05 indicates that the instructor's and the pre-service teachers' assessments do not differ significantly.

An analysis of the number of assessments consistent with the decisions taken by the instructor in different weeks shows that the pre-service teachers made effective assessments in the 1st, 2nd, and 7th weeks, while they made ineffective assessments in the 3rd and 4th weeks. In the 5th week, almost half of the evaluations were correct, and in the 6th week, more than half of the evaluations were correct. In fact, the accuracy of the pre-service teachers' assessments was expected to increase throughout the process. However, this did not occur in this study. In other words, the pre-service teachers could not make decisions with increasing accuracy in the seven-week period.

The pre-service teachers were observed to have difficulties in making the right decision in the 3rd and 4th weeks. This may be attributed to the fact that there were more students with different performance levels in these weeks than in other weeks and the presentation performances were examined according to the instructor's assessments. Firstly, the average and standard deviation values of the scores obtained from the instructor's assessments in different weeks were examined. The findings are given in Table 3.

**Table 3.**

Mean and standard deviation values for the instructor's assessments

| | Week1 | Week2 | Week3 | Week4 | Week5 | Week6 | Week7 |
|---|---|---|---|---|---|---|---|
| **Mean** | 29,14 | 31,57 | 24,57 | 25,0 | 27,9 | 29,7 | 31,7 |
| **Standard deviation** | 6,0 | 7,1 | 6,2 | 6,6 | 4,3 | 5,3 | 4,5 |

As can be seen in Table 3, the mean scores of instructor assessments in the 3rd and 4th weeks are lower than the other weeks. The mean score has decreased in the fifth week, and the standard deviation has narrowed. Thus, it can be concluded that, as the quality of the performance decreased, the pre-service teachers had difficulty making the right decisions in their assessment. To examine this, performances are divided into different levels, taking into consideration the minimum and maximum values that can be obtained from the rubric. While determining the assessment criteria, narrow score ranges were preferred as this could reveal the performance differences better. The performance bands were defined according to the instructor's assessments. The pre-service teachers evaluated with the scores of 11-15 (performance level 1), 16-20 (performance level 2), 21-25 (performance level 3), 26-30 (performance level 4), and 30 or higher (performance level 5) were grouped, and the Wilcoxon *t*-Test was repeated for each group. Since there were no pre-service teachers who received scores in the range of 11-15, the performance analysis was conducted based on four levels. The *p* values obtained are presented in Table 4.

**Table 4.**

*p* values obtained from Wilcoxon *t*-Test regarding the difference between the pre-service teachers' assessment scores and the instructor's assessment scores according to the level of performance

| | Performance level | | | |
|---|---|---|---|---|
| **Assessor** | **2** | **3** | **4** | **5** |
| 1 | .018 | .007 | .003 | **.251** |
| 2 | .018 | .007 | .004 | **.775** |
| 3 | .018 | .005 | .007 | **.684** |
| 4 | .018 | .005 | .003 | .008 |
| 5 | .018 | .028 | .007 | **.628** |
| 6 | .018 | .012 | **.050** | **.096** |
| 7 | .046 | **.113** | **.283** | **.521** |
| 8 | .018 | .005 | .006 | **.058** |
| 9 | .028 | .008 | .004 | **.862** |
| 10 | .018 | .016 | .010 | **.450** |
| 11 | **.107** | .017 | **.065** | .020 |
| 12 | .018 | .005 | .003 | .021 |
| 13 | .034 | .008 | .003 | **.160** |
| 14 | .018 | .005 | .005 | **.826** |
| 15 | .018 | .005 | .004 | .006 |
| 16 | .018 | .005 | .003 | **.097** |
| 17 | .043 | .011 | .008 | **.392** |
| 18 | .027 | .007 | .007 | **.087** |
| 19 | .018 | .018 | .003 | **.439** |
| 20 | .018 | .005 | .003 | **.301** |
| 21 | .034 | .007 | .003 | **.139** |
| 22 | .034 | .005 | .010 | **.260** |
| 23 | **.150** | .012 | .016 | **.736** |
| 24 | .028 | .008 | .005 | **.185** |
| 25 | **.063** | **.058** | .009 | **.191** |
| 26 | .041 | .005 | .005 | **.087** |
| 27 | .028 | .005 | .005 | .034 |
| 28 | .018 | .004 | .026 | **.757** |
| 29 | **.051** | .007 | .004 | **.792** |
| 30 | **.121** | **.066** | .014 | **.391** |

| | | | | |
|---|---|---|---|---|
| 31 | .018 | .007 | .016 | **.155** |
| 32 | .017 | .005 | .025 | **.249** |
| 33 | .026 | .019 | **.059** | **.315** |
| 34 | .046 | .005 | .036 | **.841** |
| 35 | .026 | .005 | .003 | **.794** |
| 36 | **.093** | .014 | .007 | **.256** |
| 37 | .018 | .005 | .004 | **.061** |
| 38 | .018 | .005 | .006 | **.629** |
| 39 | .018 | .005 | .005 | **.359** |
| 40 | .028 | .005 | .012 | **.504** |
| 41 | .027 | .005 | .009 | **.215** |
| 42 | .018 | .011 | .008 | .026 |
| 43 | .046 | .008 | .007 | **.091** |
| 44 | .028 | .021 | .004 | **.984** |
| The number of consistent assessments | 6 | 3 | 4 | 38 |

As can be seen in Table 4, the majority of the pre-service teachers could not make effective decisions in evaluating low and medium level performances; however, as the performance level increased, they could make more effective decisions in their assessments. For this reason, it may be concluded that it is the level of the assessed performance, rather than the number of assessments and the assessment process, that affects the accuracy of the pre-service teachers' assessments in different weeks.

*2. 5. 2. Results of qualitative data analysis*

Although performance assessments are carried out using rubrics, they are not free of the judgments of the evaluator. The major source of information about these judgments is the evaluators themselves. Evaluators' perceptions in online peer assessment can significantly influence the research results. Therefore, four questions were sent to the participants through e-mail ten days after the assessment period was completed. This section presents the findings regarding the experiences and perceptions of the participants about the assessment procedure followed in the study.

The results of the qualitative data analysis conducted through thematic analysis are summarized in Figure 1. As can be seen here, the views of the pre-service teachers on online peer assessment are categorized under five themes: (i) contribution, (ii) accuracy of the assessment, (iii) difficulties in the process, (iv) concerns, and (vi) the effects of the process on performance.

As regards the contribution of online peer assessment, the participants emphasized its effects on their learning. They reported that they associated the learning content and learning skills with unbiased evaluation, using the criteria improving presentation skills and enhancing practical training. They also stated that online peer assessment increased their attention in the lesson. As a matter of fact, they started to detect performance-related mistakes more easily, evaluated their own performance more effectively, understood the expected performance better, and participated in the lesson more. Below are some of the opinions of the participants on this issue.

*"I learned to evaluate both my peers and myself using correct criteria and methods."*

*"We had to listen and understand carefully since we would evaluate and grade after the presentation."*

*"I transferred what I learned in this lesson to the teaching practice lesson, and I realized how much I learned, especially about preparing a presentation. "*

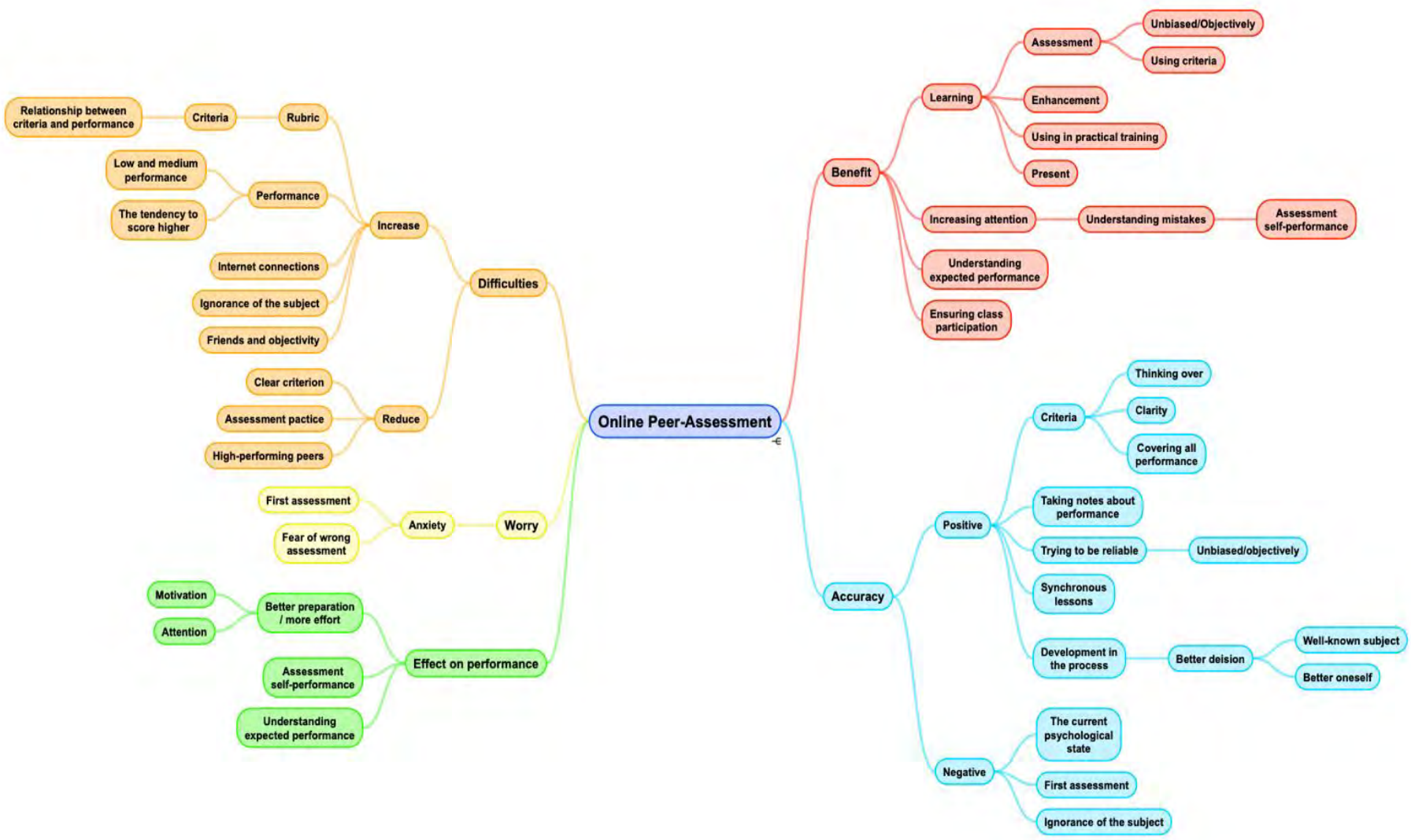


**Fig. 1.** Findings of the qualitative analysis of participants' views and experiences about the online peer assessment practice

The themes that emerged from the participant responses regarding the factors increasing the accuracy of the assessments are as follows: (i) reflection on performance by the help of criteria, (ii) wide coverage of the criteria, (iii) notes taken during the performance, (iv) effort for objectivity, (v) attendance to synchronous classes, and (vi) advantage of familiarity to topics. They stated that their psychological state during the assessments, uncertainty in their initial assessments, and unfamiliarity to the subject reduced the accuracy of the assessments. Below is the opinion of one participant on the issue:

*"Initially I was very anxious while assessing the performance of my peers. I was asking myself if I could make the right assessment or not. However, I think that I made the right assessments because I had the opportunity to practice a lot before the actual assessment process."*

The themes pertaining to the difficulties the participants experienced in their online peer assessment practices are as follows: (i) occasional inability to relate the criteria to the performance, (ii) difficulty to assess low and medium level performances, (iii) tendency to give high scores, (iv) internet connection problems, (v) lack of familiarity to the subject, and (vi) difficulty to remain subjective when evaluating friends. Participants stated that the clarity of the criteria, increasing experience with peer assessment, and the assessment of high-level performances reduced the difficulties they experienced in online peer assessment. Below are some of the opinions of the participants on this issue:

*"Sometimes the internet problems I experienced at home negatively affected my assessment."*

*"It was easy to evaluate the friends who prepared for their presentations well and showed successful performances. However, I realized that I had difficulty in evaluating poor performances."*

*"I understood the presentation more clearly if I was familiar with the subject, but I had difficulties in evaluating some aspects of the presentations on subjects I did not know very well. For example, are the examples suitable for the subject? Were the selected examples correct? Does it fit the theoretical content? Evaluating these aspects were difficult in some subjects."*

The participants stated that they were more anxious during the first assessments due to fear of making wrong assessments. They also stated that online peer assessments helped them make better preparations, increased their performance, allowed them to evaluate their own performance more realistically, and clarified what is expected of them in the presentation. Below are some of the opinions of the participants on this aspect:

*"Being evaluated by my friends and knowing that this community of friends consists of my colleagues has made me study and research more for an almost perfect presentation."*

*"Thanks to the peer assessment procedure, I noticed the mistakes in my friends' presentations and performances. Thus, I had the opportunity to correct my mistakes and did my best to perform with as few mistakes as possible. As we evaluated each performance after watching them, I believe that we continuously reinforced what we learned and our learning became permanent."*

### *2.6. Discussions*

This study investigated the effects of and participants' perception into an online classroom practice which involved peer assessment of oral presentations. The practice was conducted with pre-service teachers who were undergoing their practice teaching training.

As regards the reliability of multiple assessments, the quantitative data analysis revealed that the pre-service teachers could evaluate different performances consistently. This result is in concordance with the findings of the research by Iglesias Pérez et al. (2020) pointing to the high reliability of peer assessments. The use of rubric in assessments increases inter-rater reliability (Jonsson & Svingby, 2007), which is confirmed by the findings obtained from the qualitative data analysis. According to the participants, the use of set criteria in the assessments, the clarity of the criteria, and the conformity between the criteria and the expected

performance increased consistency of their assessments. The participants attributed the accuracy of the assessments to taking notes during the evaluation of the performance, their efforts to make an objective and reliable evaluation, and their synchronous participation in classes.

Reliability is not the only critical concept in evaluating performance. Validity should also be established in authentic forms of assessment (Jonsson & Svingby, 2007). However, this may be difficult as far as performance assessments are concerned. In this study, the validity of the assessments was measured based on the relationship between teacher assessments and peer assessments. The analysis of the quantitative data revealed that online peer assessments may sometimes fail to produce valid results when the instructor's assessments are used as the criteria. In fact, the validity of the assessments proved significantly low in the evaluation of low and medium level performances. The qualitative findings revealed that the participants had difficulty especially in the initial assessments. They stated that sometimes they knew little about the subject, which affected the accuracy of their assessments. Qualitative findings were parallel to the quantitative findings.

The predetermined, clear, and precise criteria provide the evaluators with deep insights into peer assessments (Iglesias Pérez et al., 2020; Reuse-Durham, 2005); during these assessments, the criteria users learn about their own work (Lu & Law, 2012), and the effect size increases significantly when the evaluators are trained and the evaluation is done online (Li et al., 2020). In this study, consistent results were obtained in the assessments probably because the pre-service teachers were clearly informed about how to use the rubric and what to expect of the performance. This suggests that, in peer assessments, consistency increases when the purpose and function of peer assessment and the content of the task are well-explained. However, the consistency of the results does not guarantee their validity (Crocker & Algina, 1986). For this reason, in peer assessments, the accuracy of the assessments, as well as the consistency of the results, should be carefully examined.

The purpose of assessment is important in interpreting reliability and validity coefficients (Jonsson & Svingby, 2007). For example, while reliability and validity are more important in high-stake assessments, the contribution of assessment to teaching is more important in classroom assessments. While reliability is regarded as a prerequisite to validity in large-scale assessments, this does not necessarily apply to classroom assessments. Class decisions based on an assessment can easily be changed if they seem to be wrong (Black, 1998). Hence, lower levels of reliability are acceptable, at least with relatively low-risk assessments. In brief, validity tends to be more important in authentic classroom assessments. When an assessment that produces highly reliability results is ineffective in distinguishing between different performance levels, it is likely to be useful in classroom assessments (Gearhart et al.,1995).

Based on this argument in the literature, it can be concluded that, the ability of the pre-service teachers to make assessments by distinguishing between low and intermediate level students is a marked indicator of the validity of the peer assessments. That is, the aim of the peer assessment is to make more accurate decisions and to better evaluate their own performance based on these decisions, rather than produce reliable results. The results of this study show that what makes the assessors make the right decisions is not about the process alone or the procedure of multiple assessments, which is confirmed by participant responses. The participants stated that the following factors create difficulties in the online peer assessment process: (i) difficulty in establishing a relationship between performance and the criteria, (ii) challenge of evaluating medium and low-level performances, (iii) tendency to give high scores, (iv) unstable internet connection (v) unfamiliarity to the subject, and (vi) difficulty to remain objective when assessing friends.

The quantitative and qualitative findings show that making multiple assessments or hoping that these assessments will be valid over time may jeopardize the potential benefit from peer assessments. In this study, the instructor provided a theoretical training on performance content only in the first two weeks and a practical training on performance assessments in the third week. This initial training was not sufficient for peer assessments to produce valid results. In future studies on online peer assessment, the validity of

the assessments can be investigated considering this situation; it is advised that feedback sessions and training practices be spread over time, rather than conducted in a specific period. It is also recommended that assessors be exposed to some content to gain experience. The instructor's giving descriptive and reflective feedback after each peer assessment can enhance the assessment assessors' experience.

Some findings suggest that, in peer assessment, grading alone does not provide a significant gain for students, while descriptive or reflective feedback does (Li et al., 2020). Zeng (2020) stated that reflective feedback should include topics beyond teaching and presentation and that the reflection process should be structured to contribute to learning. Training is effective in increasing the consistency between the judgments of the assessors; however, it should be noted that it will not completely eliminate differences in decisions (Stuhlmann et al., 1999; Weigle, 1999).

In their study conducted with university students, Jones and Alcock (2014) stated that peer assessments give reliable and valid results. They further maintained that in peer assessments, inter-rater reliability is generally lower than it is between experts because students have less experience than experts and less subject matter knowledge. In the study, it was stated that the correlation between peer assessments was lower than the correlation between experts' assessments. This finding is consistent with the findings of the present study.

Validity is about the correct use of assessment tools as far as performance assessment is concerned. This shows that assessors are an important factor affecting the validity of the results (Baartman et al., 2007). In performance evaluation, the main reason why reliability is associated with inter-rater reliability is closely linked with the evaluator's subjective judgments. However, it is not sufficient to examine the consistency between raters alone. At the same time, consistency must be maintained between varying tasks (Dunbar et al., 1991).

Although making more assessments might be thought to enable pre-service teachers to grasp the assessment process and relate the performance to criteria more effectively, in reality, the accuracy of the assessments was influenced by the quality of the performance evaluated more than the process. The related literature expresses that higher education students can make more accurate assessments in peer assessment than K12 students because of their stronger reflection skills (Falchikov & Boud, 1989). Thus, it was concluded that the assessors in the present study should have received more reflective feedback in evaluating the level of performance even if they studied in higher education or graduated from an undergraduate program. Presumably, feedback on the accuracy of their own assessments will enable the pre-service teachers to evaluate different performances more accurately.

One reason for the low validity of online peer assessments may be that the assessors are not knowledgeable and experienced in the subject they are assessing (Van den Berg et al.,2006). This is also supported by the qualitative findings obtained in this study. The participants stated that they had greater difficulty in making an assessment when they did not know the subject very well.

It is further stated in the literature that the effectiveness of students in peer assessment varies (Falchikov & Goldfinch, 2000). Orsmond et al. (1996) argue that since students have an immature understanding of how to interpret the given criteria, they may not find what they look for although they know what to look for. The differences between the judgments of the instructor and the pre-service teachers can, therefore, be attributed not only to performance, but to the students' insufficient understanding of the criteria used. Thus, associating the reflective feedback to be given after peer assessments with both performance and criteria may help eliminate these two obstacles. In this study, the participants stated that they sometimes had difficulty in establishing a connection between the performance and the criteria. This finding coincides with the related literature (Orsmond et al., 1996).

## 3. Conclusion and Suggestions

Assessors being unfamiliar to the subject they are assessing and inexperienced in peer assessment might account for the low validity of online peer assessments (Van den Berg et al., 2006), an assumption also supported by the qualitative findings of the study. The participants pointed to the greater difficulty they had in making assessments when their knowledge on the subject was limited.

In this study, it was observed that the pre-service teachers could evaluate varying performances consistently. However, the validity of the assessments proved significantly low in the evaluation of low and medium level performances. It may be because instructor's assessments were used as the criteria, which may not produce valid results in online peer assessments all the time. The initial three weeks of theoretical training and reflective feedback were obviously insufficient to produce valid online peer assessments.

In the qualitative part of the study, the pre-service teachers brought up the contribution of online peer assessment to their development, the accuracy of the assessments, the difficulties experienced in the process, the effect of this experience on their own performance, and the concerns they had about the procedure. The findings regarding the contributions of online peer assessment can be summarized as follows. The participants reported that online peer assessment:

- facilitates learning how to use the criteria and making assessments objectively,
- contributes to their own learning,
- improves their attendance to the lesson, helping them understand performance-related mistakes and learn how to evaluate their own performance,
- clarifies expectations from the task, and
- increases their participation in classes.

The findings regarding the accuracy of online peer assessment are grouped below. The participants believed that:

- using clear and comprehensive criteria increases their involvement in the performance, increasing accuracy of assessments,
- taking notes on the performance to be objective while making assessments led to accurate assessments,
- being familiar to the subject helped them make better decisions, also contributing to their development in the process, and
- viewing better performances added to accuracy.

As far as the negative aspects regarding the accuracy of online peer assessment are concerned, the pre-service teachers stated that

- their psychological mood during the assessment affected the accuracy of the assessments, and
- their inexperience and difficulty in making the right decisions on unfamiliar subjects decreased the accuracy of the assessments in the initial assessments.

As regards the difficulties experienced in the process, the participants stated the sources and solutions to difficulties:

- They had difficulties in relating the criteria to the performance, evaluating low and medium level performances, and assessing presentations on less-known subjects. Tendency to give high scores, internet connection problems, and difficulty to evaluate friends objectively were other problems that, they believed, negatively affected the assessment process.

- They stated that the clarity of the criteria, the increasing experience in making assessments, and the evaluation of high-level performances reduce these difficulties.

The pre-service teachers stated that online peer assessment caused some anxiety. The anxiety level was higher in the first runs, and they were afraid to make incorrect assessments. As far as the impact of online peer assessment on their own performance is concerned, the pre-service teachers stated that

- their motivation and attention increased, and
- their understanding of the performance descriptors improved,
- which, in turn, escalated their overall performance.

The results obtained in this study regarding the low validity of online peer assessments do not necessarily mean that peer assessments should be avoided in classroom practices. The related literature has substantial evidence to the advantages of it. The use of peer assessments in classroom practices has a positive effect on students' learning (McConlogue, 2015). Learners making peer assessments believe in the benefits of it as they can reflect on their own performance and improve it (Wanner & Palmer, 2018). Learners reflect on performance during peer assessment and learn to make constructive criticism (Wang et al., 2012).

Research on peer assessment processes reveal different results for different parties: those who assess and those who are assessed (Lu & Law, 2012). Assessment processes enable assessors to improve their skills in the task of assessing. The effect of peer assessments on the assessors' own performance was out of the scope of this study since there was not enough time for a preservice teacher to make more than one presentation. It is suggested that future studies investigate this aspect of peer assessment.

## References

Baartman, L. K., Bastiaens, T. J., Kirschner, P. A., & Van der Vleuten, C. P. (2007). Teachers' opinions on quality criteria for Competency Assessment Programs. *Teaching and Teacher Education, 23*(6), 857-867. doi: 10.1016/j.tate.2006.04.043.

Black, P. (1998). *Testing: friend or foe? Theory and Practice of Assessment and Testing*. Maidenhead, London: Falmer Press.

Black, P., Harrison, C., & Lee, C. (2003). *Assessment for Learning: Putting It Into Practice*. UK: Open University Press.

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology, 3*(2), 77-101. doi: 10.1191/1478088706qp063oa.

Brennan, R. L. (2001). *Generalizability Theory*. New York: Springer. doi: 10.1007/978-1-4757-3456-0.

Brown, G., Glasswell, K., & Harland, D. (2004). Accuracy in scoring of writing: Studies of reliability and validity using a New Zealand writing assessment system. *Assessing Writing, 9*, 105–121. doi: 10.1016/j.asw.2004.07.001.

Carr, J. (2020). Teacher Candidate Perceptions on Alternative Asynchronous Online Discussion Boards. *Journal of Educational Technology & Online Learning, 3*(3), 288-310. doi: 10.31681/jetol.752283.

Chang, C. C., Tseng, K. H., Chou, P. N. & Chen, Y. H. (2011). Reliability and validity of web-based portfolio peer assessment: A case study for a senior high school's students taking computer course. *Computers and Education, 57,* 1306–16. doi: 10.1016/j.compedu.2011.01.014.

Cheng, K. H., Liang, J. C., & Tsai, C. C. (2015). Examining the role of feedback messages in undergraduate students' writing performance during an online peer assessment activity. *The Internet and Higher Education, 25*, 78–84. doi: 10.1016/j.iheduc.2015.02.001.

Cho, K., Schunn, C. D., & Wilson, R. W. (2006). Validity and reliability of scaffolded peer assessment of writing from instructor and student perspectives. *Journal of Educational Psychology, 98*(4), 891. doi: 10.1037/0022-0663.98.4.891.

Creswell, J. W. (2014). *A Concise Introduction to Mixed Methods Research*. USA: Sage Publications Inc.

Creswell, J. W., & Clark, V. L. P. (2017). *Designing and Conducting Mixed Methods Research*. USA: Sage publications Inc.

Crocker, L., & Algina, J. (1986). *Introduction to Classical and Modern Test Theory*. USA: Thomson Learning.

Dunbar, S. B., Koretz, D. M., & Hoover, H. D. (1991). Quality control in the development and use of performance assessment. *Applied Measurement in Education, 4*, 289-303. doi: 10.1207/s15324818ame0404_3.

Falchikov, N., & Boud, D. (1989). Student self-assessment in higher education: A meta-analysis. *Review of Educational Research, 59*(4), 395–430. doi: 10.3102/00346543059004395.

Falchikov, N., & Goldfinch, J. (2000). Student peer assessment in higher education: A meta-analysis comparing peer and teacher marks. *Review of Educational Research, 70*, 287–322. doi: 10.3102/00346543070003287.

Gearhart, M., Herman, J. L., Novak, J. R., & Wolf, S. A. (1995). Toward the instructional utility of large-scale writing assessment: Validation of a new narrative rubric. *Assessing Writing, 2*, 207-242. doi: 10.1016/1075-2935(95)90013-6.

Iglesias Pérez, M. C., Vidal-Puga, J., & Pino Juste, M. R. (2020). The role of self and peer assessment in higher education. *Studies in Higher Education,* 1-10. doi: 10.1080/03075079.2020.1783526.

James, N., & Busher, H. (2006). Credibility, authenticity and voice: Dilemmas in online interviewing. *Qualitative Research, 6*(3), 403-420. doi: 10.1177/1468794106065010.

Jones, I., & Alcock, L. (2014). Peer assessment without assessment criteria. *Studies in Higher Education, 39*(10), 1774-1787. doi: 10.1080/03075079.2013.821974.

Jonsson, A., & Svingby, G. (2007). The use of scoring rubrics: Reliability, validity and educational consequences. *Educational Research Review, 2*(2), 130-144. doi: 10.1016/j.edurev.2007.05.002.

Lemlech, J. K. (1995). *Becoming a Professional Leader*. New York: Scholastic Inc.

Li, H., Xiong, Y., Hunter, C. V., Guo, X., & Tywoniw, R. (2020). Does peer assessment promote student learning? A meta-analysis. *Assessment & Evaluation in Higher Education, 45*(2), 193-211. doi: 10.1080/02602938.2019.1620679.

Lu, J., & Law, N. (2012). Online peer assessment: Effects of cognitive and affective feedback. *Instructional Science, 40*(2), 257-275. doi: 10.1007/s11251-011-9177-2.

Mann, C. & Stewart, F. (2000) *Internet Communication and Qualitative Research*. A Handbook for Research Online. London: Sage.

McConlogue, T. (2015). Making Judgements: Investigating the process of composing and receiving peer feedback. *Studies in Higher Education, 40*(9), 1495–1506. doi: 10.1080/03075079.2013.868878.

Moskal, B. M., & Leydens, J. A. (2000). Scoring rubric development: Validity and reliability. *Practical Assessment, Research & Evaluation, 7,* 71-81. doi: 10.7275/q7rm-gg74.

Nelson, M. M., & Schunn, C. D. (2009). The nature of feedback: How different types of peer feedback affect writing performance. *Instructional Science, 37*(4), 375–401. doi: 10.1007/s11251-008-9053-x.

Nicol, D. J., & MacFarlane-Dick, D. (2006). Formative and self-regulated: A model and seven principles of good feedback practice. *Studies in Higher Education, 31*(20), 199-218. doi: 10.1080/03075070600572090.

Ohaja, M., Dunlea, M., & Muldoon, K. (2013). Group marking and peer assessment during a group poster presentation: The experiences and views of midwifery students. *Nurse Education in Practice, 13*(5), 466-470. doi: 10.1016/j.nepr.2012.11.005.

Orsmond, P., Merry, S., & Reiling, K. (1996). The importance of marking criteria in the use of peer assessment. *Assessment and Evaluation in Higher Education, 21*(3), 239–250. doi: 10.1080/0260293960210304.

Popham, W. J. (1997). What's wrong-and what's right- with rubrics. *Educational Leadership, 55*(2), 72-75. Retrieved from http://www.ascd.org/publications/educational-leadership/oct97/vol55/num02/What's-Wrong%E2%80%94and-What's-Right%E2%80%94with-Rubrics.aspx.

Reinholz, D. (2016). The assessment cycle: A model for learning through peer assessment. *Assessment & Evaluation in Higher Education, 41*(2), 301-315. doi: 10.1080/02602938.2015.1008982.

Reuse-Durham, N. (2005). Peer evaluation as an active learning technique. *Journal of Instructional Psychology, 32*(4), 328–345. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.522.838&rep=rep1&type=pdf.

Roscoe, R. D., & Chi, M. T. (2007). Understanding tutor learning: Knowledge-building and knowledge-telling in peer tutors' explanations and questions. *Review of Educational Research, 77*(4), 534-574. doi: 10.3102/0034654307309920.

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability Theory: A primer.* London: Sage Inc.

Stevens, D. D., & Levi, A. J. (2013). *Introduction to Rubrics: An assessment Tool to Save Grading Time, Convey Effective Feedback, and Promote Student Learning*. Virginia: Stylus Publishing, LLC.

Strijbos, J.-W., Narciss, S., & Dünnebier, K. (2010). Peer feedback content and sender's competence level in academic writing revision tasks: Are they critical for feedback perceptions and efficiency? *Learning and Instruction, 20*(4), 291–303. doi: 10.1016/j.learninstruc.2009.08.008.

Stuhlmann, J., Daniel, C., Dellinger, A., Denny, R.K., & Powers, T. (1999). A generalizability study of the effects of training on teachers' abilities to rate children's writing using a rubric. *Journal of Reading Psychology, 20*, 107-127. doi: 10.1080/027027199278439.

Şencan, H. (2005). *Sosyal ve Davranışsal Ölçmelerde Güvenirlik ve Geçerlik*. Ankara: Sözkesen Matbaacılık.

Taskiran, A. (2021). Project-based online learning experiences of pre-service teachers. *Journal of Educational Technology and Online Learning, 4*(3), 391-405. doi: 10.31681/jetol.977159.

Tomlinson, P. (1998). *Understanding Mentoring*. Buckingham: Open University Press.

Topping, K. J. (2009). Peer assessment. *Theory into Practice, 48*(1), 20-27. doi: 10.1080/00405840802577569.

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*, 149–169. doi: 10.1080/713611428.

Van den Berg, I., Admiraal, W. & Pilot, A. (2006). Design principles and outcomes of peer assessment in higher education. *Studies in Higher Education, 31*, 341–356. doi: 10.1080/03075070600680836.

Wang, Y., Li, H., Feng, Y., Jiang, Y. & Liu, Y. (2012). Assessment of programming language learning Based on Peer Code Review Model: Implementation and experience report. *Computers & Education, 59*(2), 412–422. doi: 10.1016/j.compedu.2012.01.007.

Wanner, T., & Palmer, E. (2018). Formative self-and peer assessment for improved student learning: the crucial factors of design, teacher participation and feedback. *Assessment and Evaluation in Higher Education, 43*(7), 1032–1047. doi: 10.1080/02602938.2018.1427698.

Zeng, L. M. (2020). Peer review of teaching in higher education: A systematic review of its impact on the professional development of university teachers from the teaching expertise perspective. *Educational Research Review, 31* (100333), 1-16. doi: 10.1016/j.edurev.2020.100333.

Zhou, J., Zheng, Y., & Tai, J. H. M. (2020). Grudges and gratitude: The social-affective impacts of peer assessment. *Assessment & Evaluation in Higher Education, 45*(3), 345-358. doi: 10.1080/02602938.2019.1643449.