Self, Peer, and Teacher Assessment of Oral Presentations: An Investigation of Reliability and Validity[1]

Mohammad Salehi[2], Sharif University of Technology, Tehran, Iran

Samaneh Gholampour[3], Azad University, Tehran North Branch, Tehran, Iran

Abstract

While there is a burgeoning of short-term research in self, peer, and teacher assessment, there is a paucity of longitudinal studies in this kind of assessment. Furthermore, oral assessment in English as a foreign language (EFL) context is relatively underexplored. The current study was designed to investigate the reliability and validity in 30 oral presentations that spanned over a period of four months, including variables such as the students' gender, ability, and the topic of presentations. The results indicated that reliability indices of all 30 oral assessments were high, ranging from .70 to .95 as measured by Cronbach Alpha. In addition, validity indices improved over time, with the later presentations enjoying higher validity indices than the preceding ones. The gender of the assessors did not affect the total mean score of the presentations. Also, students' self-assessments were not significantly different from the scores of the teacher, although they appeared slightly higher. There was no significant correlation between self-assessment and teacher-assessment.

Resumen

El número de investigaciones a corto plazo sobre la auto-evaluación, la evaluación de pares y la evaluación por parte de los profesores ha aumentado, sin embargo, hay pocos estudios longitudinales sobre este tipo de evaluación. Además, la evaluación oral en el contexto del inglés como lengua extranjera (EFL) ha sido poco explorada. La presente investigación se diseñó para explorar la confiabilidad y la validez en 30 presentaciones orales que abarcaron un período de cuatro meses. El estudio incluyó variables como el género de los estudiantes, su capacidad y el tema de sus presentaciones. Los resultados indicaron que los índices de confiabilidad de las 30 evaluaciones orales eran elevados, oscilando entre 0,70 y 0,95 de acuerdo con el coeficiente de Alfa de Cronbach. Asimismo, los índices de validez mejoraron con el paso del tiempo, y las presentaciones posteriores tuvieron índices de validez más altos que las anteriores. Otro resultado fue que el género de los evaluadores no afectó a la puntuación media total de las presentaciones. Por medio de los puntajes de las presentaciones orales se pudieron predecir válidamente los puntajes de rendimiento de los alumnos. Finalmente, las auto-evaluaciones de los alumnos no fueron significativamente diferentes de las evaluaciones del profesor, aunque parecían ligeramente más altas. Por lo tanto, no se encontró una correlación significativa entre la auto-evaluación y la evaluación del profesor.

Introduction

Self-assessment is an integral part of self-regulated learning (Panadero & Alonso-Tapia, 2013). Self and peer assessment are alternative methods that are beginning to replace traditional assessment (Falchikov, 1988). In traditional assessment, teachers or practitioners are given the power of assessment, whereas self and peer assessment empower students to evaluate themselves and others. This can promote learner autonomy and enhance learning. In performance assessment, which departs from traditional assessment, peer assessment (PA) is defined as making decisions about performance against criteria (Adachi et al., 2018). Reliability and validity are essential in any type of assessment including performance assessment. Reliability and validity of self and peer assessment have already been investigated. For example, Jonsson and Svingby (2007) found that the use of a rubric can enhance the reliability of scoring. They concluded that reliability is not necessarily a precondition for validity in performance tests. Cho et al. (2006) concluded that aggregate ratings of at least four peer raters were as reliable and as valid as those of the instructors'. Salehi and Sayyar (2017) also concluded that peer assessment was both a reliable and valid measure of speaking and writing tasks. As applied to oral presentations, Salehi and Daryabar's (2014) study also indicated that peer assessment was both reliable and valid.

To date, however, there have been no studies to track the two notions of reliability and validity longitudinally in oral presentations. Therefore, this study was designed to investigate the reliability and validity of oral presentation assessment over a period of time. If reliability and validity indices change over time and from one presentation to the other, this will have implications for training and familiarizing students with self and

[1] This is a refereed article. Received: 9 April, 2021. Accepted: 29 October, 2021. Published: 9 June, 2022.
[2] m_salehi@sharif.ir, 0000-0002-0615-8584, Correspondent
[3] samaneh_gholampour@yahoo.com, 0000-0002-9827-846X

peer assessments. One study that comes close to the aims of the current study is that of Han (2018) who investigated English-Chinese bidirectional interpretation over three formative assessments. The researcher found that training was effective in improving bidirectional interpretation abilities. Another study that is also aligned with the purposes of the current study is that of Murillo-Zamorano and Montanero (2017) who investigated the long-term effect of peer assessment in oral presentations, but their study did not investigate the reliability and validity of assessments.

Peer and self assessments are likely to be affected by demographic variables like gender, ability level, age, and socioeconomic status. Bolivar-Cruz and Verano-Tacoronte (2018) found that men's assessment is driven by incentives, whereas women's assessments are a function of their confidence as speakers. Langan et al. (2005) found that male peer assessors awarded higher scores to male speakers. Despite the fact that there are studies on gender, they are scarce (Torres-Guijarro & Begoechea, 2017) and inconsistent (Tucker, 2014). In another study, Langan et al. (2008) investigated the relationship between student characteristics and self, peer, and tutor evaluations of oral presentations. They found that females gave lower scores to themselves compared to those of tutors. Insomnia the night before the presentations gave rise to lower scores. Langan et al., (2005) also investigated the effect of participation in the development of assessment criteria: those who were involved in the development of criteria awarded lower marks than those who did not. These lower marks approximated those of the tutors. Although gender and some student characteristics have been investigated, differential ratings of students with high proficiency level and low-to-medium level of proficiency have not been investigated. It is particularly important to investigate this gap in the literature. If the logic behind peer assessment is to lift the burden of assessment from the shoulders of teachers (Bould & Holmes, 1995), then perhaps students with high language proficiency can be entrusted with the daunting task of assessment along and/or beside teachers.

Also unexplored is the predictive role of oral presentations in English as foreign language (EFL) contexts. Therefore, this study attempts to find out if oral presentation scores can predict end-of-course achievement scores in general English classes. If it is found that oral presentations can predict the end-of-course achievement scores, then trust can be placed on oral presentations as effective tools of assessment in cases such as absenteeism. Furthermore, the role of the topic of presentations in oral presentations has not been investigated so far. This is especially pertinent to English as foreign language contexts where variations in topics can affect the assessment of oral presentations.

In view of the existing gap in the literature of self, peer, and teacher assessments, the following research questions were raised

1-How do reliability and validity indices for oral presentation peer assessments change over time?

2-What are the effects of the variables of gender, ability level, and the topic of presentations on peer assessment?

3-What is the relationship between self and teacher assessment?

Literature Review

Why self-peer and teacher assessments?

There are numerous advantages associated with self and peer assessment as alternatives to traditional assessment. One advantage is learning benefits (e.g., Falchikov, 2004). Another advantage is that self and peer assessment can support lifelong learning (e.g., Sluijsman et al., 2001). Additionally, some studies refer to critical thinking skills as potential learning benefits of peer assessment (Adachi et al, 2018; Vickerman, 2009). In addition, peer and self-assessment can also enhance motivation (Chang et al., 2012) and promote learner autonomy (Tait-McCutcheaon & Knewstubb, 2017). Learner autonomy refers to independence in learning. In terms of self and peer assessment, students are empowered to be independent assessors. Peer assessment may also positively impact students' performance in later exams (Jhangiani, 2016), since students develop expectations of any subsequent exam that they will take. Self and peer assessment can attune students to course goals and objectives. Some researchers point to the benefit of accountability and self-reflection that can be derived from self and peer assessment (e.g., Barbera, 2009). This is especially important if students' assessments can carry some weight in the final evaluation of students. Another advantage of peer assessment referred to in the literature is feedback (Carnell, 2016). According to Chang et al. (2012), after peer assessment students can emulate their strengths and avoid others' weaknesses. This is especially important if feedback is provided right after the presentations. Self and peer assessment can provide a milieu in which giving and receiving feedback are facilitated (Yucel et al., 2014). Finally, some researchers (e.g., Boud & Holmes, 1995) believe that teacher workload can be alleviated by the inclusion

of self and peer assessment. In other words, teachers can delegate some workload to students. As mentioned above, this delegation is especially important if a few select students can be entrusted with assessment in place of teachers.

Reliability and validity of self-, peer and teacher assessments

Reliability refers to consistency in scoring (Ary et al., 2013; Rezaee & Salehi, 2008). According to classical true score theory, reliability is of different types (Bachman, 1990). However, aligned with the purpose of the current study, only inter-rater consistency is discussed. Cronbach's Alpha is used when more than two assessors are involved. According to Jonsson and Svingby (2007), a Cronbach value of .70 is considered acceptable. As Falchikov (2004) maintains, it is important to investigate the reliability and validity of self and peer assessment. Otherwise, the idea of involving students in assessment does not make sense. In traditional terms, validity is of four types: content validity, criterion-related validity, face validity, and construct validity[4]. In this research, our concern is criterion-related validity. Criterion-related validity is of two types: concurrent validity and predictive. Both types are investigated in the current study. In peer assessment, concurrent validity is often defined as the degree of a Pearson correlation between the scoring of a teacher, whose judgement is deemed as the criterion measure, and the mean scorings of peers whose judgements are gauged against the criterion measure. If there is a high correlation between the scores given by teachers and those of peers, then the peer assessment can be said to be valid. In other words, it is taken for granted that teachers' assessments are valid. This kind of assessment is referred to as peer assessment (PA) from teacher's perspective (Cho et al., 2006; Panadero et al., 2013). The other type of criterion-related validity is predictive validity which deals with the prediction of scores on a dependent variable. If it can be found that the final achievement scores can be predicted by oral presentation scores, then scores given to oral presentation can be said to be a valid indicator of ability.

The role of demographic variables

Demographic variables like gender, age, friendship effect, achievement level have been postulated as factors affecting peer assessment (e.g., Falchikov, 2004). Cho et al. (2006) refer to factors like race and friendship as demographic factors affecting peer assessment. Level of achievement is another factor that is likely to affect assessment. Low achievers tended to over mark themselves compared to high achieving students. Another study that was conducted with the respect to the role of demographic variables is that of De Grez et al. (2012) who found an interaction effect for the role of gender in peer assessment in that male assessors gave higher marks to female presenters than vice versa. The reason, the authors admit, might be unknown and attributed to the generosity of males. But further research is warranted into the issue.

The use of a rubric

Using a rubric is essential to self and peer assessment. There are advantages and disadvantages associated with using a rubric. As for advantages, Falchikov & Goldfinch (2000), assert that peer assessment based on well-defined criteria appears to be more valid than peer assessment with individual dimensions. As a further advantage associated with using rubrics, Yucel et al. (2014) maintain that, because of their engagement with a rubric, students can develop a cognizance of what is expected. This is especially important when the criteria in the rubrics are clearly aligned with the end of course evaluation. A case in point can be courses in EFL with a focus on speaking skills. In such cases, oral presentations can prepare students for final achievement tests. Thirdly, if students can grasp the criteria that exist in the rubrics, this can better help promote learning (Yucel et al., 2014). In other words, students will be attuned to the criteria which embody learning goals. Despite the benefits associated with using rubrics, some studies (e.g., Jones & Wheaden, 2015) found that not using criteria resulted in good reliability and validity indices.

Oral presentations

Two studies have used rubrics with oral presentations. The first study is that of Murillo-Zamorano & Montanero (2017) who were interested in probing the long-term effects of peer assessment. They had two groups of students at their disposal: one group with a rubric and the other without one. The one with a rubric improved more than the one without a rubric. This study is important in two ways which are in alignment with the purpose of the current study. One is the use of a rubric. Another one is the long-term effect of using the rubric. However, in their study, the reliability and validity of using a rubric in peer and

[4] Content validity deals with the representativeness of the material in the test. Face validity deals with appropriateness of a test in terms of appearance. Construct validity deals with operational definitions of constructs.

teacher assessment of peer assessment were not taken into account. Another study that of De Grez et al. (2012) who compared teachers' assessments and those of peers in oral presentations that were video recorded. They found that teachers and peers agreed on certain aspects of the rubric employed. But they also disagreed in some other aspects of the criteria. To estimate reliability, they used intra-class correlation. But the interesting point is that they interpreted it as reliability not as validity, as the common practice does (Topping, 1998).

Rubrics have been deemed useful for assessing ther skills in EFL settings as well. Wang (2016), for example, found that students embraced the rubric as a very useful tool. The researcher based the research on reflective journals produced by the students and also retrospective interviews with six participants. In another study conducted by Lindblom-Ylanne et al. (2006) the use of a rubric for assessing writing tasks was deemed very appropriate. In their study, self, teacher and peer assessments converged, which the authors attribute to the good design of the study which included a good rubric for assessing writing tasks.

Methodology

Setting and participants

The study was undertaken in two intact (non-randomized) General English classes in Sharif University of Technology in Tehran, Iran. The students had registered for a general English course which was held two times a week for three hours for an academic semester. One class was held in the morning (N=30) with ten female and twenty male students. The other class was taught in the afternoon (N=28) with six female and twenty-two male students. The students ranged from 18-20 years of age. Teaching these two classes was part of the workload of the researcher. Giving oral presentations was a requirement of the course. The textbook *Inside Reading* was used in this General English course. This book focuses on a series of passages followed by vocabulary and reading comprehension exercises. This book was complemented by classroom discussions and oral presentations, and the students had to take a final exam which consisted of reading comprehension and vocabulary items. Ten percent of the final scoring was dedicated to the oral presentations. As Magin and Helmore (2001) rightly mention, students will not take the assessments seriously unless they are counted towards their final grade.

Out of 58 presentations, only thirty presentations were included in the study. The first ten presentations were used as lead ins to the study. It took time for the students to get used to self and peer assessment. Ten more presentations were also excluded because they were delivered one month after the course was over and on the day of the final exam. The pressure of the final exam was one of the reasons those ten presentations were omitted. Therefore, only thirty oral presentations were included in the analyses. Eight more presentations were excluded because of the fact that too many students were absent in those sessions. The same teacher taught the two classes. Therefore, teacher variability was not a problem. Informed consent was obtained from all the participants. Sharif University of Technology Language Center Ethical Committee approved the study. The purpose of the study was not revealed to the students.

Instrumentation

A rubric consisting of thirteen criteria used in a previous study (Peng, 2010; Salehi & Sayyar, 2017) was reused with modifications in this study. For example, *originality* was added as some students copied materials from the internet. Students gave a score on a Likert scale from 0 to 5. Maximum possible score a presenter could get was 65 (13×5). The rubric was distributed in each of the thirty presentations to both the presenter (self-assessment) and other students (peer assessment). The language of the rubric was English, and the meaning of some unfamiliar words was explained to the students (e.g., rehearsal). Eight of the thirteen criteria were related to *content* and five to *delivery*.

Data collection

Data were collected after each presentation. On the rubric, spaces were provided for the students to write additional comments about the presentations. Time allotment for each presentation was 15-20 minutes. Finishing the presentations in time was one of the thirteen criteria in the rubric; therefore, the presenters did their best to finish the presentations within the time limitation. Peer assessment was not conducted anonymously.

Indices of validity

In the current study, two indices of validity were employed:

Descriptive Statistics

If self and peer assessment are within one standard deviation (SD) of teacher's assessment, they are said to be valid (Kwan and Leung, 1996). Therefore, in this study, the total means of self, peer and teacher assessments were calculated. The same thing was done for every single presentation. The degree of similarity in the total mean of the scores would reveal the validity indices of self and peer assessments.

Correlational Analysis

According to Melvin and Lord (1995), if teachers' ratings can be viewed as a criterion measure, then correlations with teacher ratings can serve as the criterion-related concurrent validity for the peer ratings. In this way the concurrent validity of the oral presentations was calculated. Considering the fact that the researcher had experience in the area of teaching EFL, his evaluations could be deemed as a yardstick against which the ratings of peers could be compared.

Data Analysis

To answer the first research question, the researchers estimated validity and reliability in the following way: To arrive at validity, descriptive statistics of self, peer, and teacher assessments were computed. Furthermore, the scores given by the teacher were correlated with the mean scores of the peers. To compute reliability, Cronbach's alpha was computed for the thirty presentations.

To answer the second research question, independent samples t-tests were employed for both gender and the ability level. High ability level students were defined as those students who were very proficient in the English language as determined by the teacher in classroom interaction. Independent samples T-tests were performed for each presentation to see if females and males rated the presenters differently. Finally, an independent sample T-Test was conducted to see if female and male presenters received differential ratings from the peers. The same procedure was repeated for the ability level. As for the topic of presentations, a one-way ANOVA was conducted for three categories of topics: purely scientific topics, non-scientific topics, and topics that dealt with applied sciences. An example of a non-scientific topic is "great men in the history of Persia". An example of a purely scientific topic is "torsion in physics or black holes". Finally, an example of a topic dealing with applied sciences is "social robots". Finally, after t-tests and an Analysis of Variance, a Univariate analysis was run with three independent variables of gender, ability level and topic and scores given by peers to see if there were any main and interaction effects for the independent and the dependent variables.

To answer the third research question, a Pearson Product Moment correlation was run to correlate the scores the presenters gave themselves with those of the teacher.

Results

Research question 1: How do reliability and validity indices for oral presentation peer assessments change over time?

To answer the first research question, reliability and concurrent validity indices for the presentations were computed. The validity indices as calculated by correlating the scores given by the teacher with those of the peers. As Figure 1 below shows, these varied from one presentation to the other in that the later presentations yielded higher validity indices than the preceding ones. The mean scores of the presentations of the teacher were also significantly correlated with those of the peers, speaking to the validity of the peers' presentation scores (r=.68). In terms of reliability indices, Cronbach alphas of .70 are usually considered acceptable. Reliability scores for most presentations were satisfactory, but presentations that had low reliability did have low validity as well. For example, presentations 3 and 4 had low reliability (smaller than .70), and they also had low validity. But it does not seem to be the case that reliability is a prerequisite for validity. Unlike validity indices, reliability indices did not seem to have improved over time; for example, the very first presentation showed a good reliability index, but presentation 29 did not. However, there were some inconsistencies in the data. Presentations 1-13 do have low validity indices except for number 6 and number 8. Presentations from 14 onward do have good validity indices except for 23, 26, and 29.

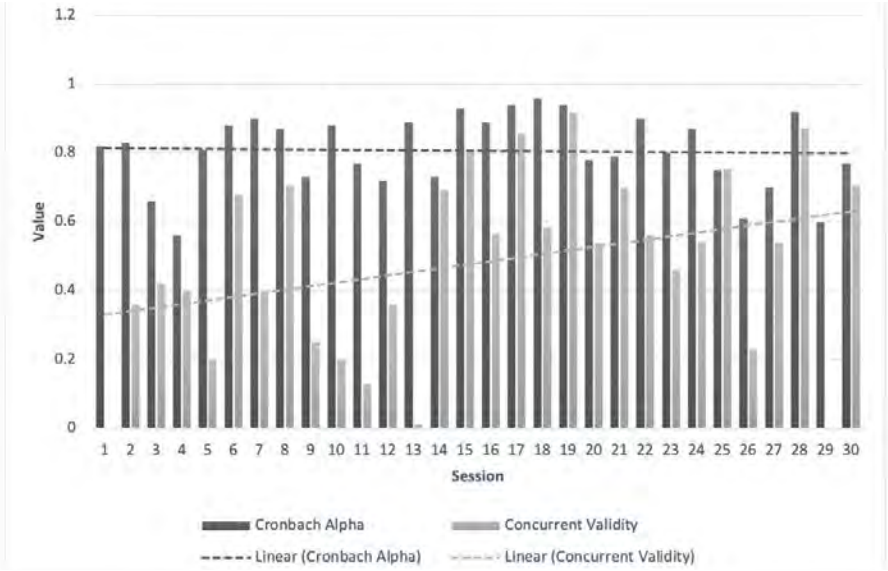

Figure 1: Reliability and validity of oral presentations

Then, descriptive statistics for teacher, peer, and self-assessment were computed as shown in Table 1 below:

assessment type / descriptive statistics	teacher assessment	peer assessment	self assessment
mean	**52.73**	**52.88**	**53.21**
Standard deviation	**7.76**	**5.7**	8.30

Table 1. Descriptive statistics of all assessment types

As can be observed in the table, peer and self- assessments are within one SD of teachers' scores. As a matter of fact, there is no over marking on the part of the presenters. This speaks to the validity of self and peer assessments which are closely aligned with those of the teacher.

Finally, descriptive statistics for every single presentation can be found in Figure 2. There is no clear pattern of improvement over time in descriptive statistics of the presentations.

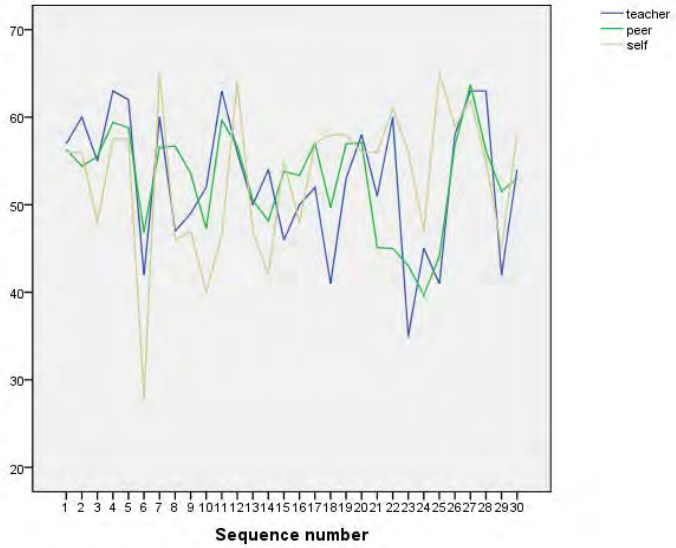

Figure 2: Teacher, peer, and self-assessment over time

Research question 2-What are the effects of the variables of gender, ability level, and the topic of presentations on peer assessment?

To examine the role of gender within each presentation, thirty independent t-tests were conducted. Most t-tests were insignificant. In those cases where the difference between males and females was significant, females had given higher scores to male presenters. In the case of presentations 8, 9, 13, 20, 25 and 3, the difference between male and female assessors was significant and the female assessors had given higher scores to the male presenters. Furthermore, the overall difference between males and females in terms of mean scores given by peers was insignificant (df=28, t= .37, *p*=.77). Finally, an independent samples t-test was run to see if there were any differences between males and females in terms of total score they received. The differences were not significant (df=28, p=.544, t=-.01).

Another variable was the ability level of the students. The researchers were interested to see if there were any differences between the ratings of two levels of students. In each oral presentation, the peer assessments given by students of high ability level and those of the low ability were compared with those of the teacher. The results showed that the correlation of the assessment of high ability students with those of the teacher was higher (*r*=.75) than that of low ability students with the teacher (*r*=.61).

The last factor to be investigated was the role of the topic of the presentations. A One-Way ANOVA was run to investigate the effect of three types of topics on the assessment. The results showed that there were significant differences among the three types of topics (F=5.91, df=2, p=.01). Topics dealing with applied sciences were assessed most favorably, followed by non-scientific topics, and finally by purely scientific ones. The fact that the topics that were purely scientific were the least favored ones by the students was interesting because this study was conducted at a technological university.

Research question 3: What is the relationship between self and teacher assessment?

In most presentations, there was an insignificant correlation between self and teacher assessments. For example in one case, the correlation was r=.30. Overmarking was only observed in a few presentations. For instance, in one case, a student gave himself a score of 4.5 out of 5.

Discussion

In this study, the reliability and validity of oral presentations were examined longitudinally. It was found that validity indices did improve over time. However, reliability indices did not improve longitudinally. The validity of oral presentation scores was investigated from three perspectives: descriptive statistics, correlational analysis, and predictive validation. Those scores proved valid from all the three perspectives. The role of demographic variables was also investigated. Female raters gave higher scores to male presenters than they did to female presenters. The role of ability level was also investigated. The scoring of the best students was more analogous to those of the teacher than those of the low-to-average students. Topics of presentations did have a differential effect on the scoring of the presenters. The most favorable scores were given to presentations dealing with applied sciences. Finally, over-scoring was not frequently observed in the study.

As Falchikov (2004) mentions, student familiarity with criteria can enhance the validity of peer assessment. That the peer assessments of the later presentations yielded better validity coefficients can be accounted for by the fact that over a period of four months, the students seem to have grasped the idea of how to rate the presentations. As it was seen in the results section, students' peer evaluations approximated those of the teacher. It seems to be the case that training improves students' awareness of the rating criteria (Pope, 2005). The results support the ideas of Falchikov (2004) who is of the opinion that students should be involved in the process of assessment. As Adachi et al. (2018) mention, some of the task of assessment can be allocated to students. The reason for the high reliability indices might be well-defined criteria of the rubric (Chang & Wu, 2012).

The reasons for oddities observed in reliability and validity data might be as follows for five different presentations. Presenter 6 was one of the high achievers who was highly proficient. The title of her presentation was about social robots. It seems to be the case that her presentation was quite a success in that it could engage the attention of the audience. That is why she received a high assessment on her performance. Out of 65, she got 56.49. Presenter 8 was the other way around. He was a low achiever and got 47.30 out of 65 from a combination of himself, the teacher and the peers. Peer assessors seem to have consensus concerning low and high achievers even at the beginning of presentations. There were three more odd cases towards the end of the term. These are presentations 23, 26 and 29. Presenter 23 was a

low achiever. He over-scored himself. He gave himself 55 out of 65. The teacher's evaluation of him was 46 out of 65. The peers' evaluation of him was 53.83. The teacher has given a lower mark than the student and the peers. This was one of the students who missed classes for no good reason. It might be the case that the teacher of the class rated the student a little bit subjectively. In such cases, it is best to have two or more raters instead of one to ensure fair and valid assessment. The discrepancy in 26 might be explained by the topic of the presentation which was about "child labor" which is not related to science and as such it might have led the peers to give her a low score of 45 out of 65 as opposed to the teacher who gave her a score of 60 out of 65. The low alignment of peers and teachers in number 29 might also be due to the effect of the topic of the presentation which was "great men". It is a deviation from scientific topics. The presenter talked about great men who defended the country. The study supports the findings of Lin et al. (2001) who found that peer assessment proved to be more valid than self-assessment. This is understandable on the grounds that we believe in peer assessment the sources of errors are cancelled out. In other words, aggregation of scores will give a better representation of reality than single scores.

In this study, overmarking was not frequently observed. This lends support to studies like Rian et al. (2015) who found that students in their study underrated themselves. In this study, too, a majority of students underrated themselves. But the averaged means of self-assessment was .48 point larger than that of the teacher. This kind of modesty in self-assessment can be attributed to Persian culture in which modesty is considered a virtue. This goes to show that a few students must have overrated themselves. In each and every class in this university there are students who have won world medals in different kinds of Olympics. These students are often overconfident and tend to overrate themselves in each academic field including language ability. The findings, however, are not in conformity with the findings of Hinkelman (2014) who found that students overrated themselves in presentations. The students in this study did not overrate their classmates' presentations which run counter to the findings of Garcia-Ros (2011). The study does not support the findings of Magin & Helmore (2001) in whose study over-marking was not observed. Actually, in certain presentations, undermarking was observed. Presentations like this might have led to the low degree of alignment between teacher and self-assessments.

The results of the study demonstrate that the ability level of the students made a difference not in differential assessments of students, but in the degree of simliarity to the teacher's assessment. It shows that the assessment of the best students is more valid and trustworthy than those of the low-ability students. One reason can be that the best students are more cognizant of the criteria than the low-ability students are. A case in point is pronunciation. The highly proficient students in these two classes were familiar with all the criteria including pronunciation. It is a truism that if one's pronunciation is not accurate, they cannot make proper judgments about the pronunciation of the others.

The effect of the topic was also explored in this study. There were significant differences among the presentations with reference to the role of topics of presentations. The least favorable topics were purely scientific topics. Perhaps the assessors looked at language classes as a relief from their own classes which are purely scientific. I talked to a small group of the participants as I noticed lower scores were given to presentations dealing with purely scientific topics. They were of the idea that a language class should be a place for the exchange of ideas about topics that are different from their other classes. They see language classes as a welcome change from their strict science classes. Students were generally interested in topics dealing with applied sciences like social robots.

The results about the inconsistency between self and teacher assessments confirm the findings of Knowles et al. (2005). In other words, teachers are better assessors of students' abilities than students themselves. In a way, this casts a show of doubt on self-assessment being the only way of assessment. It has to be accompanied by other types of assessment.

Finally, there was a discrepancy in the two validity perspectives. Specifically, in terms of descriptive statistics, self and peer assessments are within one standard deviation of the teacher whose scoring is deemed valid. In other words, self and peer assessments were said to be valid. It was already established that peer assessment was valid from the perspective of concurrent validity and descriptive statistics. However, self-assessments did not enjoy concurrent validity. The reason for this discrepancy can be explained by the following reasons: first and foremost, descriptive statistics is not an adequate basis for making statistical inferences. Secondly, although self and teacher assessments are similar in terms of descriptive statistics, differential weighting of the criteria can explain the disparity. In other words, a criterion that is important for the teacher might not be important for the presenter.

Conclusion

If students' ratings are valid, then they should be involved in assessment and some burden can be lifted off the shoulders of the teacher (Boud & Soler, 2015). It can be safely concluded that students can be entrusted with the assessment of their peers. In other words, peer assessment can be delegated to students. However, this does not seem to hold true for self-assessment. Students do not seem to have fair assessments of themselves. In this study, the trend for most of them was to underestimate themselves.

Another conclusion that can be drawn from the findings of the current study is that reliability is not necessarily a precondition for validity. As was seen in the presentations, reliabilities as high as .83 do not ensure good validity indices. Thus, authentic or performance assessment seems to depart from traditional assessment in that validity does not seem to be dependent on reliability.

The role of the topic should be investigated in future research on oral presentations in EFL contexts. It might be the case that certain topics are more appealing for students than some others. Further studies need to be conducted regarding the topics of presentations.

The findings of the study underscore not only the importance of using a rubric but also training in the use of it (MacDonald, 2011; Rubin and Turner, 2012). It was shown that validity and reliability indices were enhanced over time, perhaps as the result of students' familiarity with the criteria in the rubrics.

Finally, the findings also showed that it is important to have a convergence of validity perspectives (Ary et al., 2013). In other words, separate lines of validity inquiries should be brought to bear on the results of a single test.

References

Adachi, C., Tai, J. H.-M. T., & Dawson, P. (2018). Academics' perceptions of the benefits and challenges of self and peer assessment in higher education. *Assessment & Evaluation in Higher Education*, *43*(2), 294-306. https://doi.org/10.1080/02602938.2017.1339775

Ary, D., Jacobs, L. C., Irvine, C. K. S., & Walker, D. (2013). *Introduction to research in education.* Cengage.

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford University Press.

Barbera, E. (2009). Mutual feedback in e-portfolio assessment: An approach to the netfolio system. *British Journal of Educational Technology*, *40*(2), 342-357. https://doi.org/10.1111/j.1467-8535.2007.00803.x

Bolivar-Cruz, A., & Verano-Tacoronte, D. (2018). Self-assessment of oral presentation competence: Effects of gender and student's performance. *Studies in Educational Evaluation, 59*, 94-101. https://doi.org/10.1016/j.stueduc.2018.04.001

Boud, D., & Holmes, H. (1995). Self and peer marking in a large technical subject. In D. Boud (Ed.), Enhancing learning through self assessment. Kogan Page .

Boud, D., & Soler, R. (2015). Sustainable assessment revisited. *Assessment & Evaluation in Higher Education,* https://doi.org/10.1080/02602938.2015.1018133

Carnell, B. (2016). Aiming for autonomy: Formative peer assessment in a final-year undergraduate course. *Assessment & Evaluation in Higher Education,* https://doi.org/10.1080/02602938.2015.1077196

Chang, C.-C., Tseng, K.-H., & Lou, S.-J (2012). A comparative analysis of the consistency and difference among teacher-assessment, student self-assessment and peer-assessment in a Web-based portfolio assessment environment for high school students. *Computers & Education, 58*(1), 303-320. https://doi.org/10.1016/j.compedu.2011.08.005

Chang, C.-C., & Wu, B.-H. (2012). Is teacher assessment reliable or valid for high school students under a web-based portfolio environment? *Journal of Educational Technology & Society*, *15*(4), 265-278. https://www.learntechlib.org/p/75474

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), 891-9601. https://psycnet.apa.org/doi/10.1037/0022-0663.98.4.891

De Grez, L., Valcke, M., &. Roozen, I. (2012). How effective are self-and peer assessment of oral presentation skills compared with teacher's assessments? *Active Learning in Higher Education, 13*(2). https://doi.org/10.1177/1469787412441284

Falchikov, N. (1988). Self and peer assessment of a group project designed to promote the skills of capability. *Programmed Learning and Educational Technology*, *25*(4), 327-339. https://doi.org/10.1080/1355800880250409

Falchikov, N. (2004). Involving students in assessment. *Psychology Learning & Teaching, 3*(2), 102-108. https://journals.sagepub.com/doi/pdf/10.2304/plat.2003.3.2.102

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. https://doi.org/10.3102%2F00346543070003287

Garcia-Ros, R. (2011). Analysis and validation of a rubric to assess oral presentation skills in university contexts. *Electronic Journal of Research in Educational Psychology*, *9*(3), 1043-1062. http://www.investigacion-psicopedagogica.org/revista/new/english/ContadorArticulo.php?639

Han, C. (2018). A longitudinal quantitative investigation into the current validity of self and peer assessment applied to English-Chinese bi-directional interpretation in an undergraduate interpreting course. *Studies in Educational Evaluation, 58*, 187-196. https://doi.org/10.1016/j.stueduc.2018.01.001

Hung, Y.-J. (2018). Group peer assessment of oral English performance in a Taiwanese elementary school. *Studies in Educational Evaluation, 59*, 19-28. https://doi.org/10.1016/j.stueduc.2018.02.001

Hinkelman, D. (2014). Rubric-based assessment for video-recorded learner performances: Blending paper rubrics with a Moodle LMS module. *Proceedings of Moodle Moot Japan 2014*, 63-69.

Jhangiani, R. S. (2016). The impact of participating in a peer assessment activity on subsequent academic performance. *Teaching of Psychology*. https://doi.org/10.1177%2F0098628316649312

Jones, I., *&* Wheadon, C. (2015). Peer assessment using comparative and absolute judgment. *Studies in Educational Evaluation*, *47*, 93-101. https://doi.org/10.1016/j.stueduc.2015.09.004

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

Knowles, M. S., Holton, E. F., Swanson, R. A., & Holton, E. (2005). *The adult learner*: *The definitive classic in adult education and human resource development*. Taylor & Francis Ltd.

Kwan, K. P., *&* Leung, R. W. (1996). Tutor versus group peer assessment of student performance in a simulation training exercise. *Assessment & Evaluation in Higher Education, 21*(3), 205-214. https://doi.org/10.1080/0260293960210301

Langan, A. M., Shuker, D. M., Cullen, W. R., Penney, D., Preziosi, R. F., *&* Wheater, C. P. (2008). Relationships between student characteristics and self-, peer and tutor evaluations of oral presentations. *Assessment and Evaluation in Higher Education*, *33*(2), 179-190. https://doi.org/10.1080/02602930701292498

Langan, A. M., Wheater, C. P., Shaw, E. M., Haines, B. J., Cullen, W. R., Boyle, J. C., Penney, D., Oldekop, J. A., Ashcroft, C., Lockey, L., & Preziosi, R. F. (2005). *Assessment and Evaluation in Higher Education, 30*(1). 21-34. https://doi.org/10.1080/0260293042003243878

Lin, S. S. J., Liu, E. Z. F., *&* Yuan, S. M. (2001). Web-based peer assessment: Feedback for students with various thinking-styles. *Journal of Computer Assisted Learning*, *17*(4), 420-432. https://doi.org/10.1046/j.0266-4909.2001.00198.x

Lindblom-ylänne, S., Pihlajamaki, H., *&* Kotkas, T. (2006). Self-, peer- and teacher-assessment of student essays. *Active Learning in Higher Education*, *7*(1), 51-62. https://doi.org/10.1177%2F1469787406061148

MacDonald, K. (2011). A reflection on the introduction of a self and peer assessment initiative. *Practice and Evidence of Scholarship of Teaching and Learning in Higher Education*, *6*(1), 27-42.

Magin, D. *&* Helmore, P. (2001). Peer and teacher assessments of oral presentation skills: How reliable are they? *Studies in Higher Education*, *26*(3), 287-298. https://eresearch.qmu.ac.uk/handle/20.500.12289/2294

Melvin, K. B., *&* Lord, A. T. (1995). The prof/peer method of evaluating class participation: Interdisciplinary generality. *College Student Journal, 29*, 258-263.

Murillo-Zamorano, L. R., *&* Montanero, M. (2017). Oral presentations in higher education: A comparison of the impact of peer and teacher feedback. *Assessment & Evaluation in Higher Education*, *43*(1). https://doi.org/10.1080/02602938.2017.1303032

Panadero, E., *&* Alonso-Tapia, J. (2013). Self-assessment: Theoretical and practical connotations. When it happens, how is it acquired and what to do to develop it in our students. *Electronic Journal of Research in Educational Psychology, 11*(2), 551-576. https://psycnet.apa.org/doi/10.14204/ejrep.30.12200

Panadero, E., Romero, M., *&* Strijbos, J.-W. (2013). The impact of a rubric and friendship on peer assessment: Effects on construct validity, performance, and perceptions of fairness and comfort. *Studies in Educational Evaluation*, *39*(4) 195-203. https://doi.org/10.1016/j.stueduc.2013.10.005

Peng, J. C. (2010). Peer assessment in an EFL context: Attitudes and correlations. In T. P. Matthew, Y. Watanabe, & S. K. Lee. (Eds.), *Selected proceedings of the 2008 second language research forum* (pp. 89-107). Cascadilla Proceedings Project.

Pope, N. K. Ll. (2005). The impact of stress in self- and peer assessment. *Assessment & Evaluation in Higher Education, 30*(1), 51-63. https://doi.org/10.1080/0260293042003243896

Rezaee, A. A. and Salehi, M. (2008). The construct validity of the University of Tehran English Proficiency Test: A Multitrait Multimethod Approach. *Journal of Teaching English Language and Literature Society of Iran, 2*(8), 93-110.

Richmond, K. (2012). *Inside reading: Academic word list in context*. Oxford University Press.

Rian, J. P., Hinkelman, D., & Cotter, M. (2015). Self-, peer, and teacher assessments of student presentation videos. In P. Clements, A. Krause, & H. Brown (Eds.), *JALT2014 Conference Proceedings.* JALT. https://jalt-publications.org/sites/default/files/pdf-article/jalt2014proc_074.pdf

Rubin, R. F., & Turner, T. (2012). Student performance on and attitudes toward peer assessments on advanced pharmacy practice experience assignments. *Currents in Pharmacy Teaching & Learning, 4*(2), 113–121.

Salehi, M., & Daryabar, B. (2014). Self and peer assessment of oral presentations: Investigating correlations and attitudes. *English For Specific World Purposes World*, 42.

Salehi, M., & Sayyar, Z. (2017). The reliability and the validity of self-, peer-, and teacher assessment. *South African Linguistics and Applied Language Studies,* (5)1, 1-15, https://doi.org/10.2989/16073614.2016.1267577

Sluijsmans, D. M. A., Moerkerke, G., van Merriënboer, J. J. G., *&* Dochy, F. J. R. (2001). Peer assessment in problem based learning. *Studies in Educational Evaluation*, *27*(2), 153-173. https://doi.org/10.1016/S0191-491X(01)00019-0

Tait-McCutcheon, S., *&* Knewstubb, B. (2017). Evaluating the alignment of self, peer and lecture assessment in an Aotearoa New Zealand pre-service teacher education course. *Assessment & Evaluation in Higher Education,* https://doi.org/10.1080/02602938.2017.1408771

Torres-Guijarro, S. *&* Bengoechea, M. (2017). Gender differential in self-assessment: A fact neglected in higher education self and peer assessment techniques. *Higher Education Research and Development*, *36*(5). http://dx.doi.org/10.1080/07294360.2016.1264372

Topping, K. (1998). Peer assessment between students in colleges and universities. *Review of Educational Research, 68*(3), 249-276. https://doi.org/10.3102%2F00346543068003249

Tucker, R. (2014). Sex does not matter: Gender bias and gender differences in peer assessments of contributions to group work. *Assessment & Evaluation in Higher Education*, *39*(3), 293-309. https://doi.org/10.1080/02602938.2013.830282

Vickerman, P. (2009). Student perspectives on formative peer assessment: An attempt to deepen student learning? *Assessment & Evaluation in Higher Education*, *34*(2), 221-230. https://doi.org/10.1080/02602930801955986

Wang, W. (2016). Using rubrics in student self-assessment: Student perceptions in the English as a foreign language writing context. *Assessment and Evaluation in Higher Education*, https://doi.org/10.1080/02602938.2016.1261993

Yucel, R., Bird, F. L., Young, J., *&* Blanksby, T. (2014). The road to self-assessment: Exemplar marking before peer review develops first-year students' capacity to judge the quality of a scientific report. *Assessment & Evaluation in Higher Education*, *39*(8), 971-986. https://doi.org/10.1080/02602938.2014.880400