Tabaran Institute of Higher Education
International Journal of Language Testing
Vol. 12, No. 2, October 2022
ISSN 2476-5880

Examining the Interrater Reliability Between Self- and Teacher Assessmnnu uu uuuusss r rr al ee ooommassss

Dexter L. Manzano[1]

Received: January 2022 Accepted: June 2022

Abstract
The increasing popularity of self-assessment prompted several scholars to investigate its effectiveness and accuracy in relation to teacher assessment. However, most of these studies focused only on the consistency estimate perspective. Thus, the current study investigated the interrater reliability between self- and teacher assessment of students' oral performance in Filipino. Specifically, this study used two perspectives (i.e., consistency estimate and consensus estimate perspectives) to see the full picture of interrater reliability between self- and teacher assessment. Fifty (50) college students from various specializations participated in this study. They assessed their respective oral performances using an in-class observation self-assessment with self-viewing. Findings reveal that teacher and students' SA results posted a very strong positive relationship and that their ratings agree with each other. High positive correlations suggest that both the students and the teacher consistently apply the rating scale. These results were attributed to the use of a micro-analytic rating scale, assessment training, and rating procedure used during SA. Implications for classroom assessment and future studies were discussed.

Keywords: assessment for learning; language assessment; performance evaluation

1. Introduction

Since the adoption of constructivism (i.e., an educational theory that advocates for student-focused and process-oriented learning), students have taken an active role in assessment design, criteria, and choices. As such, self-assessment (SA) has become an integral part of any language classroom practice. It involves students rating their own performances and learning. SA has also been considered to promote autonomous learning among students (Ashraf & Mahdinezhad, 2015; Butler & Lee, 2010; Ngo, 2020), help in reaching their learning goals (Goral & Bailey, 2019; Yoon & Lee, 2013), develop their metacognitive knowledge (Black, 2009; Wong & Mak, 2019), and increase their direct involvement and motivation to learn (Brown & Hudson, 1998). Conversely, SA has been reported to be moderated by several factors, such as criteria, instruments, training, and instruments (Guo & Barrot, 2019; Li & Zhang, 2021).

Despite these recognized advantages of SA, students are seldom put in charge of rating their own performances, such as asking students to evaluate their speaking performances

[1] College of Arts and Social Sciences, Tarlac State University, Tarlac, Philippines, Email: dextermanzano56@gmail.com

through a reflection activity or rubric (Luoma & Tarnanen, 2003). One reason is that students have the tendency to overestimate or underestimate their performances relative to teacher assessment or depending on their language ability (Brown & Hudson, 1998; Karnilowicz, 2012; Lew, Alwis, & Schmidt, 2010). According to Evans, McKenna, and Oliver (2005), students inflate their scores because they lack participation in the development of the assessment tool and have a tendency to perform an assessment based on potential rather than actual ability. Some studies even posited that students' SA lacked reliability compared to some external standards, such as teacher and peer assessment (e.g., Falchikov & Buod, 1989; Ross, 1998; Ward, Gruppen, & Regehr, 2002). On this note, it is desirable that teachers train their students in conducting SA and provide them opportunities to accurately and consistently assess their performances for formative purposes (Barrot, 2018). Some ways to help students improve their SA skills are through collaborative development of assessment tool and calibration sessions.

One approach to determine the accuracy and consistency of students' SA is by comparing their SA scores to another external assessor, such as t eiteacher's assessment. A widely used statistics that addresses this objective is correlational analysis, which identifies relationship between two or more related variables (Ross, 1998; Ward et al., 2002). However, it may be insufficient because correlation coefficients only measure the relationship between two variables and not the level of agreement between raters (Miller, 2003). To illustrate, a correlation coefficient may be high (e.g., 0.80); but because the gap between scores is too wide, the interrater agreement would be low. Hence, interrater reliability needs to be examined using two different but complementary angles of analyses. Juxtaposing correlational and Kappa coefficient analyses would lead to a more precise, unified, and holistic estimation and interpretation of interrater reliability.

This study, therefore, attempted to investigate both the interrater consistency (i.e., two raters sharing a common meaning of rating scale using correlation coefficient) and interrater agreement (i.e., the agreement between raters using Kappa coefficient) between teacher assessment and students' SA of their oral performances. This study would shed light on how consistent students' SA is in relation to teacher assessment and who among the group of students are likely to overestimate or underestimate. These findings would be useful for more nuanced assessment training for students.

2. Review of Literature

2.1. Interrater reliability

Interrater reliability refers to the extent to which two or more raters who are using the same rating scale award the same score to performances (Cheung & Tai, 2021; Doosti & Safa, 2021; Graham, Malinowski, & Miller, 2012). It can be measured from either consistency or a consensus estimate perspective (Stemler, 2004). Consistency estimates of interrater reliability (or interrater consistency) measure the relative similarity between two or more sets of scores. For example, rater 1 assigned a score of 1 2, 3, and 4 for each of the performance criteria in a rubric, while rater 2 assigned a score of 3, 4, 5, and 6. With these scores, a perfect relationship is to be expected; that is, if one score from rater 1 increases, so as with rater 2. The most popular statistic to measure interrater consistency is the Pearson correlation coefficient or Pearson *r*. It is a statistic used to measure the strength and degree of relationship between two different

variables. One advantage of using this correlation coefficient is its ability to accommodate continuous data (e.g., 1.50 1.75, 3.20). However, it can only be computed for one pair of raters and one item at a time (Stemler, 2004). It may also be insufficient in determining interrater reliability because correlation coefficients assume that 'data underlying the rating scale are normally distributed' (Osborne, 2008, pp. 39). Hence, results are 'affected by the distribution of observed ratings and can lead to artificially deflated estimates' (Stemler, 2004). It also lacks the ability to discern systematic differences between raters and can be misleading if there is a low variation in the scores across rates (Graham et al., 2012).

Another way to compute interrater reliability is through a consensus estimate point of view. Consensus estimates of interrater reliability (or interrater agreement) refer to the extent to which raters assign similar scores (Graham et al., 2012; Stemler, 2004). For example, while scores assigned by rater 1 (i.e., 1 2, 3, 4) and rater 2 (i.e., 3, 4, 5, 6) have a very strong positive relationship, their scores have a poor agreement because of the wide gap between two sets of scores. Two of the most popular statistics to measure interrater agreement are percent agreement and Kappa coefficient. Percent agreement is done by adding the number of ratings that obtained similar scores divided by the total number of ratings (Stemler, 2004). The problem with this statistic is that it fails to take into account chance agreement (Viera & Garrett, 2005) which leads to inflated agreement. Unlike percent agreement, the Kappa coefficient takes into account percent agreement and random chance (Stemler, 2004; Viera & Garrett, 2005) and is easy to interpret (Oakleaf, 2009). Furthermore, Kappa coefficient accounts for systematic biases and is well suited to dealing with nominal variables (Stemler, 2004).

2.2. Factors affecting interrater reliability between SA and teacher assessment

Several factors have been considered to affect the reliability of students' SA. They can either be task-related or rater-related factors. The type of language skill (i.e., reading, listening, speaking, and writing) being assessed is one example of a task-related factor. In Ross's (1998) meta-analytic review, he found that the correlations between SA and teacher assessment are strongest in receptive language skills and weakest in productive skills. Ross (1998) further argued that the reason for this is that SA of productive skills is more susceptible to extraneous factors. The other task-related factor is linked to the rating scale. While some scholars (e.g., Chang, Tseng, & Lou, 2012) agree that the rating scale may be a cause of divergence between SA and teacher assessment, others have opposing views on which type of rating scale facilitates higher interrater reliability. For instance, Miller (2003) concluded that as the number of items on the rating scale increases, so as the variance in scores. Gordon (1991) and Jonsson and Svingby (2007), however, argued that scoring consistency improves when raters use analytic rather than holistic rating scales. In fact, Gordon (1991) emphasized that using global SA compromises strong correlations between SA and teacher assessment. The way indicators are worded in the rating scale also affects how students assess themselves. Heilenman (1990), for instance, found that students find it easier to respond to positively stated indicators than to negatively worded indicators.

Aside from rating scale and type of language skills, rater factors (i.e., both students and teachers) affect the agreement and consistency between scores assigned by teachers and students. For instance, Cheng and Warren (1999) indicated that students' practical experience

in rating performances improves their ability to self-assess a similar task. De Grez, Valcke, and Roozen (2012), on the other hand, reported that teachers with rating experience tend to apply various criteria more leniently when assessing the quality of oral presentation. Teachers also tend to be more lenient when grading students' oral performances because students could employ a range of metalinguistic strategies (e.g., eye contact, body gestures) and use interactive visuals to assist in communications with the teacher and audience. This lenient application of criteria leads to variability in scores between teacher assessment and students' SA. The idea here is that the more inexperienced the teacher is as a rater, the greater the gap between teacher and students' SA would be. De Grez and his colleagues (2012) explained that experienced teachers could retrieve from their memory comprehensive models that could guide them in determining whether an oral presentation meets the standards or not.

Assessment training is another factor that influences interrater agreement between SA teacher assessment (AlFallay, 2004; Chang et al., 2012; Chen, 2008; Falchikov & Boud, 1989; Langan et al., 2005; Ross, 2006). In his meta-analytic paper, Ross (2006) reported that adequate consistency and agreement are achieved when learners are trained to self-assess. Similarly, Langan et al. (2005) found that scores given by students who participated in assessment discussions were significantly lower than the scores awarded by students who did not participate in this type of discussion. In the same way, when AlFallay (2004) implemented a three-hour workshop on SA, he found that practice contributed to SA accuracy. Chang et al. (2012) explained that because of practice, students were able to enhance their rating ability, which in turn leads to higher accuracy.

Several studies also confirmed that students' skill and proficiency level impact their ability to self-assess accurately. For instance, Falchikov and Boud (1989) and Karnilowicz (2012) concluded that high-performing students tend to underestimate themselves while low-performing ones tend to overestimate themselves. Similarly, Lew et al. (2010) reported that those who are more academically competent tend to self-assess more accurately than those with lesser ability. This phenomenon was attributed by some scholars to high-achieving students' tendency to be realistic (Falchikov & Boud, 1989). As regards students' language proficiency, Suzuki (2015) found that less experienced speakers tend to overestimate their ability while more experienced ones tend to underestimate theirs.

2.3. Studies on interrater reliability between teacher assessment and students' SA

In recent years, many studies have investigated interrater reliability between SA and teacher assessment (e.g., AlFallay, 2004; Butler & Lee, 2006; Chang et al., 2012; Karnilowicz, 2012; Ross, 2006) using a correlational coefficient. However, very little research has focused on oral performances. Some of these studies confirmed inconsistencies between SA and teacher assessment (e.g., De Grez et al., 2012; Dlaska & Krekeler, 2008), whereas others found consistency between the two, especially after training (Chen, 2008) and when using on-task SA (e.g., Butler & Lee, 2006).

One study that explored the correlations between teacher assessment and SA is that of Dlaska and Krekeler (2008), who investigated students' ability to accurately self-assess their pronunciation skills in relation to professional raters' assessment. In their study, 46 advanced learners of German assessed their production of speech sounds in comparison with the sounds

articulated by a native speaker. Using holistic assessment, they reported that SA and teacher assessment results were identical in 85 percent of all cases. However, learners were able to identify only half of the total number of speech sounds that professional raters considered as inaccurately produced. They concluded that second language (L2) learners had difficulties performing SA of their pronunciation skills. They speculated that the reasons for these difficulties include native language transfer, previous learning experience, the influence of other prosodies (i.e., the pattern of rhythm and sounds), psychological and individual factors, and sounds that are difficult to assess. De Grez et al. (2012) conducted a similar study that examined the agreement between and among teacher assessment, peer assessment, and self-assessment of oral performances. Findings suggested that teachers and peers remain to interpret rubric indicators in different ways and that self-assessment scores are higher than the scores given by the teachers. Both of these studies used a consistency estimate perspective. More recently, Oren (2018) examined the relationship between self-, peer, and teacher assessment within the teacher education program. Her findings revealed high correlations of self- and peer assessment with teacher assessment and argued that two student-initiated assessments could be useful in evaluating oral performances. This study, however, did not distinguish who among the group of students showed higher consistency with teacher assessment.

While other studies reported inconsistencies between teachers' and students' SA scores, others have differing results. For example, Butler and Lee (2006) examined the Korean students' SA of their oral performances. Specifically, they compared the validity of off-task SA (i.e., assessment in a decontextualized way) to on-task SA (i.e., assessment done immediately after the completion of the task). The findings indicated that students were able to assess themselves more accurately during on-task than during off-task assessment. They further discovered that on-task assessment was less influenced by student attitude or personality.

Unlike Butler and Lee (2006), Chen's (2008) study focused on the effects of training on students' accuracy in SA. Twenty-eight (28) Chinese students participated in the assessment program that included two weeks of training and ten weeks of two-cycle assessment. Results showed that students' SA and teacher assessment differ significantly in the first cycle but more indicated a strong correlation during the second cycle. She also found that students became more critical and independent and learned more after the training. She attributed these results to students' training and personality traits as well as teacher's feedback during the conduct of the assessment program.

As reviewed, there is an evident paucity of studies that compare students' SA and teacher assessment of oral performances, especially from a consensus estimate perspective (i.e., using Kappa coefficient). The present study, therefore, aims to fill this gap by examining not only the interrater consistency but also the interrater agreement between teacher assessment and students' SA of oral performances. Specifically, this study addresses the following research questions: (1) What is the level of students' speaking performance based on students' and teacher assessment? (2) What is the interrater consistency and agreement between teacher assessment and students' SA?

3. Methodology

3.1. Participants and setting

Fifty (50) college students whose first language (L1) is Filipino took part in this correlational study and were selected using random sampling. They were second-year (17 to 19 years old; 23 males and 27 females) students taking up liberal arts and enrolled in a Filipino Course at a state university in the Philippines. They came from various social backgrounds. None of them have experienced using a micro-analytic rating scale for SA purposes. A micro-analytic rating scale is a scoring tool that explicitly details the domains (e.g., content) and their corresponding components (e.g., depth, use of logical appeals).

The teacher of the 50 students also took part in the study. He holds a master's degree in Filipino and a doctoral degree in Education and has been teaching Filipino courses for 23 years. He rated the students' individual oral performances by viewing their recorded speeches using the designated rating scale.

3.2. Instrumentation

The speaking task involved the delivery of a six-minute persuasive speech that talked about a social issue. The speech was written and delivered in Filipino as part of the requirement of the course. Students were given almost two weeks to prepare and plan for their individual speeches prior to the actual delivery in class. The criteria for rating their performances were explained during the presentation of mechanics.

Both the students and the teacher used a micro-analytic rating scale in assessing oral performances. This type of rating scale was used because an analytic type of rating scale tends to obtain higher interrater reliability compared to a holistic rating scale (Johnson, Penny, & Gordon, 2000; Jonsson & Svingby, 2007). Miller (2003) pointed out that increasing the specificity of the rating scale addresses the problem of scoring leniency and range restriction. The rating scale has two versions: for teachers and students. These two versions only differed in the point of view used. While the rating scale for students used first person, the rating scale for the teacher used third-person pronouns. Both rating scales assigned a number to five degrees (5 = *very great extent*, 4 = *great extent*, 3 = *moderate extent*, 2 = *little extent*, 1 = *very little extent*) on 27 components (11 content-related criteria and 16 delivery-related criteria) which were equally weighted (see Appendix). To ensure the validity of the instruments, they were evaluated by two validators who have relevant research experience and at least a master's degree in English language teaching or a related field of study. These scales, as used in this study, reflect a good internal consistency based on its 27 components (α = 0.88).

3.3. Procedure

Two weeks prior to students' actual delivery, they were asked to write a persuasive speech in Filipino and to prepare for their speech. During their scheduled delivery, each of the students delivered their persuasive speech in class via MS (Microsoft) Teams and was video recorded by the teacher. The students were instructed to view their recorded speeches via MS Teams videos immediately after their presentation.

After all speeches had been delivered, a one-hour online training session for SA was conducted so that students would assess their respective performances consistently and in

accordance with the rating scale. During the training session, they were informed about the purpose of the study (i.e., comparing their self-assessment scores to teacher assessment scores). Thereafter, copies of the rating scale were posted through MS Teams forms. They were allowed to familiarize themselves with its content for 15 minutes. Afterward, the rating procedure was elaborated by explaining to them how they would use and interpret the criteria and apply them consistently to promote a certain level of objectivity (Stemler, 2004). Note that students' oral performances were assessed prior to the training session. The teacher-rater took at least 15 minutes to rate each of the performances.

After the training session, another session was allotted for an in-class observation SA; that is, students viewing their own performance (Brown & Hudson, 1998; 2002). Using their individual gadgets (e.g., smartphone and laptop computer), students were instructed to view their own performance individually at least twice prior to actual SA. Then, a copy of the rating scale was distributed for them to complete in one hour. Students were asked to choose the score that corresponds to their performance. They were also prohibited from conferring with one another and were not aware of the scores given by their teacher so as not to influence the results of the evaluation. After the students completed the SA phase, the accomplished rating scales were turned in for tallying and analysis using MS Teams. The ethical research protocol was observed prior to, during, and after collecting data.

3.4 Data analysis

Descriptive and inferential statistics through SPSS version 20 were used to analyze data. Descriptively, the mean scores and standard deviations of students' SA and teacher assessment results were computed. These scores were then subjected to Pearson product correlations to determine interrater consistency. Values greater than or equal to 0.70 are deemed to reflect strong correlations.

To determine the interrater agreement between teacher assessment and SA, Cohen's Kappa was computed for each of the performances. In other words, 50 separate Kappa coefficients were computed based on each item. This statistic was used because it takes into account chance agreement (Stemler, 2004; Viera & Garrett, 2005). Since Kappa coefficient does not accept continuous data, the overall Kappa was computed by getting the average Kappa coefficient of all cases. The level of acceptability was set at 0.41 (Yen et al., 2013).

4. Results and Discussion

The present study sought to determine the level of interrater consistency and agreement between students' SA and teacher assessment results, as well as the factors that might account for the obtained results. Findings reveal that moderate interrater agreement and very strong positive correlations exist between the scores assigned by the teacher and students on the latter's oral performance. These observations confirmed and extended previous studies (e.g., AlFallay, 2004; Butler & Lee, 2006; Chang et al., 2012; Chen, 2008; Falchikov, 2013; Goral & Bailey, 2019; Karnilowicz, 2012; Li & Zhang, 2021) that students have the ability to assess themselves in the way teachers apply the criteria and that their way of assessing themselves are shaped by their proficiency.

 ISSN 2476-5880

Table 1

Descriptive statistics of students' SA and teacher assessment based on the performance criteria.

Criteria		Students' Rating		Teacher's Rating	
Category	No.	*Mean Score*	*SD*	*Mean Score*	*SD*
Content-related criteria	1	4.82	0.67	4.50	0.48
	2	4.16	0.64	3.44	0.50
	3	4.42	0.75	3.90	0.58
	4	4.40	0.63	4.04	0.59
	5	4.34	0.65	3.94	0.71
	6	4.06	0.79	3.58	0.70
	7	4.44	0.75	4.08	0.56
	8	4.50	0.67	4.06	0.69
	9	4.34	0.71	3.98	0.77
	10	4.48	0.64	4.16	0.95
	11	4.18	0.82	3.54	0.82
Delivery-related criteria	12	4.18	0.91	3.86	1.00
	13	4.3	0.83	3.94	0.80
	14	4.00	0.92	3.68	0.92
	15	4.32	0.73	4.08	0.75
	16	4.16	0.73	3.92	0.79
	17	4.12	0.86	3.80	0.82
	18	4.10	0.73	3.90	0.92
	19	4.08	0.89	3.72	0.71
	20	4.36	0.69	3.88	0.69
	21	4.12	0.68	4.24	0.77
	22	4.12	1.12	4.12	0.71
	23	2.48	1.47	2.36	0.71
	24	2.36	1.41	1.96	0.76
	25	3.96	0.72	3.80	0.65
	26	4.64	0.59	4.28	0.60
	27	4.22	0.64	4.22	0.95
		4.14	0.80	3.81	0.74

Table 1 presents the mean scores and standard deviations for each performance criterion. Overall, the students posted a higher overall mean score ($\bar{x}$=4.14) compared to the teacher's rating ($\bar{x}$=3.81), and that their difference was significant (p=0.002). A greater variability in the scores assigned by the students (SD=0.80) was also observed compared to that of the scores assigned by the teacher (SD=0.74). More specifically, results revealed that the mean score difference between teacher assessment and students' SA is relatively higher in content-related criteria (0.45) than in delivery-related criteria (0.24).

Results also show that the students overestimate their performance relative to teacher assessment. Overestimation, in the context of this study, refers to the tendency of students to assess their performance more positively relative to teacher assessment. Specifically, almost all students rated themselves higher compared to scores given to them by their teacher except students 15, 20, 31, and 40. While students 31 and 40 rated themselves equally compared to the teacher's rating, students 15 and 20 rated themselves lower. Findings also revealed that almost all high-performing students (upper quartile) except student 15 assigned similar or lower scores compared to the teacher's rating.

These results followed earlier SA studies (e.g., Falchikov & Boud, 1989; Karnilowicz, 2012; Lew et al., 2010; Suzuki, 2015) that low-achieving students tend to overestimate while high-achieving ones tend to underestimate their performance relative to teacher assessment. One possible reason for this is that high-achieving students are more critical of themselves. A pattern was also obtained as to the difference between SA and teacher assessment scores. Findings show that students with high rating performances from the teacher do not overrate themselves. This lent support to Boud and Falchikov (1989) when they contended that the weaker the students are, the greater the degree of overrating.

Table 2 shows the interrater reliability between students' SA and teacher assessment. Kappa coefficient shows that the students and the teacher reached a considerable agreement (= 0.45) about the students' oral performances. The coefficients between the two scores ranged from poor agreement (= 0.01) to almost perfect agreement (= 0.84). Interestingly, 81 percent (13 out of 16) of delivery-related criteria posted an acceptable level of agreement (i.e., 0.70). This is much higher compared to the 27 percent (3 of 11) of content-related criteria that posted an acceptable level. One possible explanation for the higher level of agreement in delivery-related criteria is that these criteria are more discrete and directly observable. Among these observable aspects of speaking performance are eye contact, gestures and body movement, facial expression, mannerism, and posture. Another important finding that needs to be highlighted is the very poor interrater agreement and consistency in certain areas, particularly item 2. This issue may have emerged from the fact that item 2 (i.e., my arguments and insights were well researched) is not readily observable to the teacher. The teacher may have based his assessment on students' performance alone, whereas students based their assessment on what they actually did during the writing process.

As regards interrater consistency, the overall correlation between the SA and the teacher assessment was very strong (r=0.70). This finding suggests that when the teacher gives higher scores in oral performance, the students will also give higher SA scores and vice versa. As to the individual correlation coefficients, all were positive and ranged from weak (r=0.24) to high (r=0.93). Moreover, the correlations between the teacher assessment and students' SA are strong in all delivery-related criteria but not in content-related criteria. Not surprisingly, almost all of those criteria that obtained an acceptable level of the agreement also posted strong to very strong correlations.

 ISSN 2476-5880

Table 2
Interrater reliability between students' SA and teacher assessment.

Criteria			*r*
Category	No.		
Content-related criteria	1	0.43	0.52
	2	0.01	0.24
	3	0.26	0.61
	4	0.28	0.60
	5	0.38	0.53
	6	0.37	0.67
	7	0.37	0.77
	8	0.42	0.70
	9	0.48	0.64
	10	0.48	0.79
	11	0.29	0.56
Delivery-related criteria	12	0.54	0.84
	13	0.55	0.79
	14	0.45	0.76
	15	0.48	0.68
	16	0.33	0.72
	17	0.47	0.73
	18	0.60	0.87
	19	0.42	0.72
	20	0.34	0.68
	21	0.68	0.89
	22	0.84	0.93
	23	0.63	0.82
	24	0.48	0.74
	25	0.59	0.73
	26	0.36	0.63
	27	0.59	0.88
		0.45	0.70

Furthermore, the present study yielded a higher *r* value compared to previous studies. The overall correlation between the teacher assessment and students' SA of all oral performances (r=0.70) was higher than the value indicated by Ross (1998). In his study, Ross (1998) reported that the average correlation between SA of speaking and the criterion variables was 0.55 which he attributed to the use of analytic scoring. The present results also differed from previous findings (e.g., De Grez et al., 2012; Dlaska & Krekeler, 2008; Oren, 2018) that the teacher assessment and students' SA lack consistency. This can be attributed to the limited specificity of the assessment tool used in these studies. Furthermore, results showed that variability of scores is higher in SA compared to teacher assessment.

From these findings, it can be hypothesized that a micro-analytic rating scale could have improved interrater consistency and agreement. As Blanche and Merino (1989) explained, high consistency between SA and external standards could be obtained when the skills to be assessed are clear and detailed. However, this hypothesis is still subject to further studies. It should be taken that the rating scale used in the present study used positively stated indicators, as in the case of previous studies (e.g., De Grez et al., 2012; Heilenman, 1990). This feature may have also contributed to obtaining good interrater consistency. As Heilenman (1990) conjectured, respondents have a tendency to agree with positively constructed items to conform to the perceived social values and norms.

On top of the type of rating scale used, the training session may have also contributed to a moderate agreement and high consistency. These results, therefore, support the findings of previous studies (e.g., AlFallay, 2004; Chang et al., 2012; Chen, 2008; Falchikov & Boud, 1989; Langan et al., 2005; Ross, 2006) that training improves consistency.

Note that students viewed their respective performances immediately before they performed SA and were informed that their accomplished SA sheets would be reviewed by the teacher. These two features of the rating procedure employed in the present study may have augmented interrater reliability results. This would then provide partial support to Butler and Lee's (2006) finding that on-task SA could improve assessment accuracy.

While the present study reported an acceptable interrater agreement, a higher interrater agreement (i.e., 0.61 to 0.81) would have been ideal. However, there are some factors that may have constrained this. One of which is the lack of rating experience of student raters. Because of this, it is likely that they also lack the benchmarking skills needed to assess their performances accurately. And because experienced raters have already gained confidence that makes them more critical (Barkaoui, 2010), it is also likely that the teacher became more severe in his assessment. To illustrate, a score of 3 for teachers may be 5 for students.

5. Conclusions

The aim of the current study was to investigate both the interrater consistency and agreement between teacher assessment and students' SA of their oral performance, as well as the factors that might account for the obtained results. Findings reveal that teacher and students' SA results posted a very strong positive relationship and that their ratings agree with each other. High positive correlations suggest that both the students and the teacher consistently apply the rating scale. It can be speculated that the use of a micro-analytic rating scale, assessment training, and rating procedure used during SA had something to do with these results. However, these three may not be sufficient to ensure substantial or almost perfect agreement due to other intervening factors (e.g., raters' characteristics). Given the findings of this study, it is highly encouraged that students engage in SA regardless of their proficiency level.

While the current study reported some interesting insights, several limitations should be noted. Firstly, a single teacher rating of oral performance and a limited number of students may not be sufficient for more conclusive results. Another factor that may have influenced the results is the lack of involvement of students in the process of designing assessment criteria. For these reasons, the findings reported in the present study should be interpreted with caution and treated as tentative.

Despite the limitations, the present results have implications for language assessment practices. First, the findings suggest that sufficient training and a well-defined rating procedure allow students to assess themselves more accurately and reliably. Doosti and Safa (2021) were correct in emphasizing the value of rater training to promote interrater reliability. Moreover, raters need to consider the specificity of criteria when assessing students' performances and engaging them in SA. It does not only increase reliability (and construct validity), but it can also provide students with relevant information regarding their weaknesses and strengths in any language task. However, caution should be made as to how comprehensive the criteria are. As pointed out by Price and O'Donovan (2006), too comprehensive rating scale may be counterproductive.

Since there are too many factors that would put a limit on obtaining a higher interrater agreement (e.g., almost perfect agreement), teachers as assessors should not worry themselves about getting a high interrater agreement as it may just be a matter of perspective. The success of students' SA should not also be judged solely based on how similar or parallel their rating is to their teacher's rating. Instead, SA should be treated as a tool to reinforce formative assessment and facilitate learning in any language classrooms. It is because even teachers (e.g., experienced vs. inexperienced) do not sometimes agree among themselves. Instead, these differences should be used as a tool to have a more holistic view of students' abilities and as a tool to better understand the intricacies of assessment. It is likely that their SA results could actually be giving us a glimpse of students' language ability beyond what they exhibit during their actual performances.

Finally, given the paucity of studies that investigated the interrater agreement between SA and teacher assessment of oral performances, future research should be performed using a larger number of participants in various teaching contexts. Qualitative approaches (e.g., think-aloud protocol, interview, comment sheets) could also be performed to supplement quantitative approach to obtain different types of data for a more meaningful and conclusive interpretation.

Acknowledgments

The researcher is heartfully indebted to Tarlac State University for funding this research. Many thanks to the participants who shared their precious time to provide the pertinent data. He is also thankful to all anonymous reviewers whose constructive comments and insightful suggestions helped improve this paper and to the editors of IJLT for the publication of this paper.

Declaration of Conflicting Interests

No conflicting interests

Funding

This research was funded by Tarlac State University, Philippines.

References

AlFallay, I. (2004). The role of some selected psychological and personality traits of the rater in the accuracy of self- and peer-assessment. *System, 32*(3), 407–425. https://doi.org/10.1016/j.system.2004.04.006

Ashraf, H., & Mahdinezhad, M. (2015). The role of peer-assessment versus self-assessment in promoting autonomy in language use: A case of EFL learners. *International Journal of Language Testing*, *5*(2), 110–120.

Barkaoui, K. (2010). Do ESL essay raters' evaluation criteria change with experience? A mixed methods, cross sectional study. *TESOL Quarterly*, *44*(1), 31–57. https://doi.org/10.5054/tq.2010.214047

Barrot, J. S. (2018). Using the sociocognitive-transformative approach in writing classrooms: Effects on L2 learners' writing performance. *Reading & Writing Quarterly*, *34*(2), 187–201. https://doi.org/10.1080/10573569.2017.1387631

Black, P. (2009). Formative assessment issues across the curriculum: The theory and the practice. *TESOL Quarterly, 43*(3), 519–524. https://doi.org/10.1002/j.1545-7249.2009.tb00248.x

Blanche, P., & Merino, B. J. (1989). Self-assessment of foreign-language skills: Implications for teachers and researchers. *Language Learning*, *39*(3), 313–340. https://doi.org/10.1111/j.1467-1770.1989.tb00595.x

Brown, J., & Hudson, T. (1998). The alternatives in language assessment. *TESOL Quarterly, 32*(4), 653–675. https://doi.org/10.2307/3587999

Brown, J., & Hudson, T. (2002). *Criterion-referenced language testing*. Cambridge: Cambridge University Press.

Butler, Y.G., & Lee, J. (2006). On-task versus off-task self-assessments among Korean elementary school students studying English. *The Modern Language Journal*, *90*(4), 506–518. https://doi.org/10.1111/j.1540-4781.2006.00463.x

Butler, Y.G., & Lee, J. (2010). The effect of self-assessment among young learners of English. *Language Testing*, *27*(1), 5–31. https://doi.org/10.1177/0265532209346370

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 and Education*, *58*(1), 303–320. https://doi.org/10.1016/j.compedu.2011.08.005

Chen, Y. (2008). Learning to self-assess oral performance in English: A longitudinal case study. *Language Teaching Research, 12*(2) 235–262. https://doi.org/10.1177/1362168807086293

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

Cheung, K. K. C., & Tai, K. W. (2021). The use of intercoder reliability in qualitative interview data analysis in science education. *Research in Science & Technological Education*, 1–21. https://doi.org/10.1080/02635143.2021.1993179

De Grez, L., Valcke, M., & Roozen, I. (2012). How effective are self-and peer assessment of oral presentation skills compared with teachers' assessments?. *Active Learning in Higher Education, 13*(2) 129–142. https://doi.org/10.1177/1469787412441284

Dlaska, A., & Krekeler, C. (2008). Self-assessment of pronunciation. *System*, *36*(4), 506–516.

Doosti, M., & Safa, M. A. (2021). Fairness in oral language assessment: Training raters and considering examinees' expectations, *International Journal of Language Testing*, *11*(2) 64–90.

Evans, A.W., McKenna, C., & Oliver, M. (2005). Trainees' perspectives on the assessment and self-assessment of surgical skills. *Assessment and Evaluation in Higher Education, 30*(2) 163–174. https://doi.org/10.1080/0260293042000264253

Falchikov, N. (2013). *Improving assessment through student involvement: Practical solutions for aiding learning in higher and further education*. New York: Routledge Falmer.

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

Goral, D. P., & Bailey, A. L. (2019). Student self-assessment of oral explanations: Use of language learning progressions. *Language Testing*, *36*(3), 391–417. https://doi.org/10.1177/0265532219826330

Graham, M., Milanowski, A., & Miller, J. (2012). *Measuring and promoting inter-rater agreement of teacher and principal performance ratings*. Washington, DC: Center for Educator Compensation Reform.

Guo, Q., & Barrot, J. S. (2019). Effects of metalinguistic explanation and direct correction on EFL learners' linguistic accuracy. *Reading & Writing Quarterly*, *35*(3), 261–276. https://doi.org/10.1080/10573569.2018.1540320

Heilenman, L. K. (1990). Self-assessment of second language ability: The role of response effects. *Language Testing, 7*(2) 174–201. https://doi.org/10.1177/026553229000700204

Johnson, R .L., Penny, J., & Gordon, B. (2000). The relation between score resolution methods and interrater reliability: An empirical study of an analytic scoring rubric. *Applied Measurement in Education, 13,* 121–138. https://doi.org/10.1207/S15324818AME1302_1

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

Karnilowicz, W. (2012). A comparison of self-assessment and tutor assessment of undergraduate psychology students. *Social Behavior and Personality: An International Journal*, *40*(4), 591–604. https://doi.org/10.2224/sbp.2012.40.4.591

Langan, A., Wheater, C., Shaw, E., Haines, B., Cullen, W., Boyle, J, Penney, D. et al. (2005). Peer assessment of oral presentations: effects of student gender, university affiliation and participation in the development of assessment criteria. *Assessment and Evaluation in Higher Education, 30*(1) 21–34. https://doi.org/10.1080/0260293042003243878

Lew, M., Alwis, W., & Schmidt, H.G. (2010). Accuracy of students' self assessment and their beliefs about its utility. *Assessment and Evaluation in Higher Education, 35*(2) 135–156. https://doi.org/10.1080/02602930802687737

Li, M., & Zhang, X. (2021). A meta-analysis of self-assessment and language performance in language testing and assessment. *Language Testing, 38*(2), 189–218. https://doi.org/10.1177/0265532220932481

Luoma, S., & Tarnanen, M. (2003). Creating a self-rating instrument for second language writing: From idea to implementation. *Language Testing, 20*(4), 440–465. https://doi.org/10.1191/0265532203lt267oa

Miller, P. (2003). The effect of scoring criteria specificity on peer and self-assessment. *Assessment and Evaluation in Higher Education, 28*(4), 383–394. https://doi.org/10.1080/0260293032000066218

Ngo, T. T. (2020). Promoting learner autonomy through self-assessment and reflection. *VNU Journal of Foreign Studies*, *35*(6), 146–153.

Oakleaf, M. (2009). Using rubrics to assess information literacy: An examination of methodology and interrater reliability. *Journal of the American Society for Information Science and Technology*, *60*(5), 969–983. https://doi.org/10.1002/asi.21030

Oren, F. S. (2018). Self, peer and teacher assessments: What is the level of relationship between them?. *European Journal of Education Studies*, *4*(7), 1–19. https://doi.org/10.5281/zenodo.1249959

Osborne, J.W. (Ed.) (2008). *Best practices in quantitative methods*. United Kingdom: Sage.

Price, M., & O'Donovan, B. (2006). Improving performance through enhancing student understanding of criteria and feedback. In C. Bryan and K. Clegg (Eds.), *Innovative assessment in higher education* (pp. 100–109). London: Routledge.

Ross, A. (2006). The reliability, validity, and utility of self-assessment. *Practical Assessment, Research and Evaluation, 11*(10) 1–13. https://doi.org/10.7275/9wph-vv65

Ross, S. (1998). Self-assessment in second language testing: A meta-analysis and analysis of experiential factors. *Language Testing, 15*(1) 1–20. https://doi.org/10.1177/026553229801500101

Stemler, S. (2004). A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability. *Practical Assessment, Research and Evaluation, 9*(4). Retrieved March 23 2015 from http://PAREonline.net/getvn.asp?v=9andn=4

Suzuki, Y. (2015). Self-assessment of Japanese as a second language: The role of experiences in the naturalistic acquisition. *Language Testing*, *32*(1), 63–81. https://doi.org/10.1177/0265532214541885

Viera, A., & Garrett, J. (2005). Understanding interobserver agreement: The kappa statistic. *Family Medicine, 37*(5), 360–363.

Ward, M., Gruppen, L., & Regehr, G. (2002). Measuring self-assessment: Current state of the art. *Advances in Health Sciences Education, 7*(1), 63–80. https://doi.org/10.1023/A:1014585522084

Wong, K. M., & Mak, P. (2019). Self-assessment in the primary L2 writing classroom. *Canadian Modern Language Review*, *75*(2), 183–196. https://doi.org/10.3138/cmlr.2018-0197

Yen, K., Kuppermann, N., Lillis K., Monroe, D., Borgialli, D., Kerrey, B.T., Sokolove, P. et al. (2013). Interobserver agreement in the clinical assessment of children with blunt

abdominal trauma. *Academic Emergency Medicine, 20*(5), 426–432. https://doi.org/10.1111/acem.12132

Yoon, E., & Lee, H. (2013). Do effects of self-assessment differ by L2 language level? A case of Korean learners of English. *Asia-Pacific Education Researcher, 22*(4), 731–739. https://doi.org/10.1007/s40299-013-0111-z

Appendix A

Rating Scale for Students

	5	4	3	2	1
CONTENT-RELATED CRITERIA					
My topic and purpose were relevant and interesting for the audience.					
My arguments and insights were well researched.					
My arguments were supported by examples and analogies.					
My arguments were logically arranged.					
I used logical appeals.					
I used emotional appeals.					
My attention-getter was effective.					
The organizational pattern and transition I used were easy to follow.					
My conclusion emphasized the main points.					
I ended with power.					
I used clear examples and illustrations which supported the main ideas.					
DELIVERY-RELATED CRITERIA					
I maintained eye contact.					
I used volume, pitch and rate varied appropriately.					
I used gestures and body language effectively.					
I showed confidence.					
I pronounced and enunciated the words clearly.					
I avoided verbal and nonverbal mannerisms.					
I maintained the interest of the audience.					
I was enthusiastic and lively.					
I presented myself credibly and professionally.					
I established rapport with the audience.					
My speech was delivered within time limits.					
My presentation aids reinforced the message.					
I handled the presentation aids effectively.					
I handled the audience effectively.					
My language was adjusted to the level of the audience.					
I showed mastery of the piece.					
TOTAL					

ISSN 2476-5880

Appendix B

Rating Scale for Teacher

	5	4	3	2	1
CONTENT-RELATED CRITERIA					
The topic and purpose were relevant and interesting for the audience.					
Arguments and insights were well researched.					
Arguments were supported by examples and analogies.					
Arguments were logically arranged.					
The speaker used logical appeals.					
The speaker used emotional appeals.					
Attention-getter was effective.					
The organizational pattern and transitions were easy to follow.					
The conclusion emphasized the main points.					
The speaker ended with power.					
The speaker used clear examples and illustrations which supported the main ideas.					
DELIVERY-RELATED CRITERIA					
The speaker maintained eye contact.					
The speaker used volume, pitch and rate varied appropriately.					
The speaker used gestures and body language effectively.					
The speaker showed confidence.					
The speaker pronounced and enunciated the words clearly.					
The speaker avoided verbal and nonverbal mannerisms.					
The speaker maintained the interest of the audience.					
The speaker was enthusiastic and lively.					
The speaker presented himself credibly professionally.					
The speaker established rapport with the audience.					
The speech was delivered within time limits.					
Presentation aids reinforced the message.					
The speaker handled the presentation aids effectively.					
The speaker handled the audience effectively.					
The language was adjusted to the level of the audience.					
The speaker showed mastery of the piece.					
TOTAL					