TESL-EJ Teaching English as a Second or Foreign Language
The Electronic Journal for English as a Second Language

Instructor- versus Peer-based Participation Scores in EFL Classes: Comparisons and Correlation to Oral Proficiency

November 2021 – Volume 25, Number 3

Ryan Spring
Tohoku University, Japan
<spring.ryan.edward.c4@tohoku.ac.jp>

Abstract

This paper examines the differences correlation between instructor- and peer-based participation scores and oral proficiency in an EFL course focusing on oral communication. It finds that there is very little correlation between instructor-based and peer-based participation scores ($r = .187$, $p = .053$) and that the former is more associated with improvement in fluency ($r = .266$, $p = .005$), but that the latter is strongly associated with pre-class levels of fluency measures such as trimmed speech rate ($r = .364$, $p < .001$). However, neither measure of participation was particularly correlated with post-class levels of fluency. Therefore, some caution is required by instructors when using in-class participation scores. While it is important to encourage students to improve by participating in classes, the methods of assigning participation scores examined in this paper may not be appropriate as a part of the composite final grade in the class as they do not correlate well to final skill level.

Keywords*: participation, student evaluation, oral proficiency, peer-based scores*

Participation in the classroom has often been considered especially important for the development of EFL learners' oral proficiency because it is one of the only chances that many of them will have to practice speaking their L2. Therefore, many EFL teachers include some metric of classroom participation in their grading, especially for oral communication-based classes (Crosthwaite et al., 2015; Rogers, 2011). However, the definitions of participation are not always clear, and Crosthwaite et al. (2015) have suggested that instructor-given participation scores may not be fair or effective due to students having a wide range of learning styles. One idea to provide a fairer and more balanced measurement of participation in class and group work activities that has been implemented outside of EFL is the use of peer-based participation scores (e.g., Heyman & Sailors, 2011; Mello, 2010), such as teamwork scores (e.g., Britton et al., 2017; Brutus & Donia, 2010). However, it is unclear (i) how similar instructor- and peer-based participation scores are to each other, and (ii) if there are any differences between the two, which are more strongly associated with improvement in L2 oral proficiency. This paper seeks to fill in this gap in the research by providing a comparison of peer-based and instructor-based participation in an EFL oral communication class to see how similar they are, and which are more indicative of improvement.

Previous Studies

Participation and Oral Proficiency

In-class participation has been widely regarded as essential for learning. In general, participation is though to represent better engagement, which in turn is assumed will lead to better student outcomes (e.g., Britton et al., 2017; Dancer & Kamvounias, 2005), and in the field of EFL, it is often suggested to be particularly important for the development of oral proficiency (Kang, 2005). This is likely why a large number of works in the literature specifically seek to boost in-class participation (e.g., Morell, 2007; Zhao, 2015), and why willingness to communicate in the target language is considered important for the development of oral proficiency (e.g., MacIntyre & Charos, 1996; Yashima, 2002). However, the relationship between in-class participation and improvement in oral proficiency is largely an assumed one. Obviously, some degree of practice is required to obtain proficiency, and in a foreign language atmosphere, we often assume that students do not or cannot practice speaking outside of the classroom. However, it is also conceivable that students with different learning styles who are less active in classroom group activities might still be learning just as well. Unfortunately, though there are many studies that look at the effects of motivation or the number of learning sessions on specific EFL skills, there are few quantitative studies that specifically examine the link between in-class participation and oral proficiency.

One reason that fewer studies seek to verify the influence of in-class participation scores on oral proficiency, specifically, is that both contain a number of contributing aspects and can thus be difficult to quantify. First, according to Skehan (2009, 2014), there are three aspects of speech that must be considered when discussing L2 oral proficiency: fluency, complexity, and accuracy. To complicate things further, *the trade-off hypothesis* (Skehan & Foster, 1997) suggests that these three aspects rarely increase all at once, and different tasks have different effects on them (e.g., Tavakoli & Skehan, 2005). Therefore, in-class participation in certain activities might influence some of these categories, but not others. Second, there are also many aspects to classroom participation, including whether or not students interact with each other positively, exhibit leadership, cooperate, and actively attempt to communicate (e.g., Britton et al, 2017; Dancer & Kamvounias, 2005). It can be incredibly challenging for a single instructor or researcher to attempt to gauge each of these aspects in a single class or study.

Thus, in order to truly discover how in-class participation affects oral proficiency, one must look at several aspects of both. One attempt to do so was Spring (2020a) who looked at a number of factors in an EFL course to determine which aspects of the class contributed most to increases in oral fluency, complexity, and accuracy. He found that the strongest link was between out-of-class study and spoken complexity, and that there seemed to be a weak link between peer-based in-class participation scores and improvement in oral fluency, although this portion of the study was largely inconclusive.

Rating Participation in the EFL Classroom

Essentially, there are two ways to evaluate participation – either through an instructor-given score, generally based on in-class observation, and peer-given, generally based on student-surveys during groupwork activities. According to Crosthwaite et al. (2015) and Rogers (2011), most EFL instructors tend to give instructor-based participation scores, especially in classes focusing on developing oral proficiency. However, increasing attention is being given to peer-based participation scores, as they have been argued to potentially be fairer and more beneficial to evaluation (e.g., Britton et al, 2017; Dancer & Kamvounias, 2005).

Though many EFL classes tend to utilize instructor-based given participation scores, a number of studies suggest that these scores may be seriously flawed. For example, Vandrick (2000) suggests that many EFL instructor-based participation scores seem to be given subjectively without a deep understanding of what participation truly consists of. Furthermore, Crosthwaite et al. (2015) argue that instructor-based participation scores do not account for differences in individual learning styles, and thus may not reflect actual effort. Finally, it is conceivable that since it is difficult for a single instructor to properly observe all students equally during in-class sessions, instructor-based scores may result in unbalanced assessments. For these reasons, recent research has suggested that peer-based participation evaluations may be more appropriate.

One common suggestion to remedy the potential unfairness of instructor-based participation scores is to allow students to play a role in evaluating participation in the classes through peer-based participation scores (Heyman & Sailors, 2011; Mello, 2010). Since such scores would not be given by a single instructor, the amount of attention given to each student being score would increase, allowing for the use of a more detailed rubric, which could allow for an assessment of a wider range of aspects related to participation (e.g., Britton et al, 2017; Dancer & Kamvounias, 2005). However, peer-based scores also have a number of drawbacks. For example, most peer-based participation score metrics were not developed with ESL in mind, and thus might measure aspects unrelated to improvement in English ability. For example, Britton et al. (2017) developed a peer-rated teamwork metric that measures participation amongst five categories: contributing to the team project, facilitating contributions of others, planning and management, fostering a team climate, and managing potential conflict. However, it is unclear whether or not each of these areas actually contribute to improvement in language learning.

Research Questions

Thus, though there has been much discussion about how to improve in-class participation in EFL courses and how to evaluate participation, it is still unclear whether or not instructor-based or peer-based scores, if either, are more indicative of improvement. Specifically, there have been no studies comparing how similar instructor-based scores are to peer-based scores, nor have there been any showing which are more closely associated with actual EFL improvement, specifically in oral proficiency. Therefore, this study poses the following research questions:

1. Is there any significant relationship between instructor-based and peer-based participation scores, and if so, in what ways are they similar?
2. Are instructor-based or peer-based participation scores more correlated with improvement in EFL oral proficiency?

Methods

Participants

139 L1 Japanese EFL learners at Tohoku University participated in this study after receiving permission from the ethics review board of the university. All participants signed informed consent forms agreeing to allow their data from the class and speaking tests to be used in this study. The participants were second-year students in one of five identical "Practical English Skills" courses taught by the author and were either 19 or 20 years old. They had been studying English for 7 years and none had spent significant time outside of Japan, resulting in all participants being either CEFR B1 or B2 level. None of the participants were taking any other English courses at the time of the study, so it is reasonable to assume that any gains in oral proficiency were likely due to the course. Their majors varied as the curriculum at the time of the study required students to take two "Practical English Skills" courses for graduation, but students were allowed to choose which

courses they took based on the instructor and course content. Since this study chose to look at the effects of participation on improvement, the data of students who missed groupwork classes were omitted from the data set. Similarly, any students who did not complete the class or the pre- and posttests were removed from the data for this study. These omissions resulted in a data set of 108 students.

Classroom Procedure

Following Spring (2020a; b), a project-based language learning (PBLL) design was used for the course. This course design was used because it was reported as being able to assist students in improving their oral proficiencies, and also because it employed a substantial amount of groupwork, which is necessary for the use of peer-based participation score rubrics. During the course, students attended one 90-minute sessions per week for a total of 15 weeks. The first four sessions consisted of instructor-centered teaching of skills that students would need throughout the rest of the class: conducting meetings, writing meeting minutes, phrases for discussion, short presentations, and academic word parts (i.e., prefixes, suffixes and word roots). In the fourth session, the instructor explained the participation metrics to students, how they would rate each other using it, and what they were expected to do in order to receive high marks on it.

Sessions five through fourteen consisted almost entirely of groupwork during which time students were expected to form groups and conduct meetings in English in order to complete a project – a short film. Students were expected to read chapters in their textbook for homework, the details of which corresponded to the topics they would need to discuss in the following class. Each of these 10 sessions began with a quiz about the homework reading assignment and a short presentation summarizing the previous classes' meeting minutes. At least 60 minutes of each of these sessions was designated as groupwork time, during which students were told that they must speak entirely in English with the exception of asking about words and phrases that they didn't know how to say in their L1. The instructor monitored these groupwork sessions, encouraging them to speak English as much as possible, helping them with their projects, and giving advice when asked. During the final session, students presented their short films to the class.

Students' grades were determined based on their instructor-based participation scores, peer-based participation scores, short presentations, meeting minutes, quiz scores, and an assessment of their groups' final product.

Measuring Participation

Participation scores were taken in two ways. The first was an instructor-given score that followed a simple rubric. Students participated in 10 group-work sessions and the instructor monitored their sessions, assigning a score between 1 and 5 for each session. Students began with a score of 5 and one point was deducted for each inappropriate use of the L1 and for each missed opportunity to participate. I defined inappropriate use of the L1 in the rubric as using the L1 to accomplish groupwork (e.g. communicating for the purposes of accomplishing the daily tasks) and did not include using the L1 to check the meaning of unfamiliar words or get help using their L2 in other ways. Though absences and tardiness also affected students' participation scores, students who were absent or late to class were not included in the data set for this study and therefore these scores do not reflect such point deductions. The scores were summed to create an ordinal score with a range of 10 to 50.

At the end of the final class, peer-based participation scores were taken using Team-Q (see Appendix 1), following Britton et al. (2017). This metric was used because it was also suggested

by Spring (2000a; b), allows for long-term peer-based rating of participation scores, provides different scores for several different aspects of participation, and has been suggested to be a meaningful measure of in-class participation (Britton et al., 2017). The Team-Q questionnaire given to students consisted of 15 questions translated into Japanese, each with a possible score of between 0 and 4, that students of each group use to rate themselves and all other group members based on how often they exhibited appropriate teamwork (see Appendix 1). I calculated the average score for each group member for each individual question. I then calculated overall teamwork scores as the sum of these average scores for each student. Furthermore, I calculated teamwork scores for the category of "contributes to team project" by summing the average scores for the three questions in Team-Q pertaining to this area. The same methods were used to calculate teamwork scores for the categories of "facilitates contributions of others" (three questions), "planning and management" (three questions), "fosters a team climate" (two questions), and "manages potential conflict" (three questions). The final question in Team-Q asks for a summative evaluation and is therefore not included in any of the subcategories, but is used in the calculation of overall teamwork. These scores were treated as ordinal data.

Measuring Oral Proficiency

Following Spring (2020a; b), at the beginning of the first class and at the end of the last class, all students were given identical speaking tests and their responses were recorded. Though the same tasks were used, i.e., giving three short oral monologues based on practice questions taken from the IELTS test, different questions were used for the pre- and posttests. The questions were shown to students one at a time and they were given one minute to think before being asked to respond. They were then asked to speak about the topic for up to two minutes. Participants were allowed to stop speaking before two minutes, but non-answers were not permitted. The time was monitored with a stopwatch and participants were asked to stop speaking after two minutes if they had not stopped of their own volition beforehand. The recordings were then analyzed for fluency, complexity, and accuracy.

Though fluency, complexity, and accuracy can be rated subjectively, some relatively standard objective measures have been created and utilized in the measurement of EFL speaking skills. Several of these were used in this study to ensure repeatability, for ease of statistical analysis, and to increase comparability between this study and others. However, while this study uses these raw values for further analysis, it differs from previous studies in that it combines individual measures into a single, ordinal variable of improvement for each aspect of oral proficiency. This is done because studies such as Spring (2020a; b) which used only raw data needed to then make subjective judgements as to how well correlated an aspect is in general based on how many of the objective measures correlated to improvement. By combining the measures from the start, the need for a post-hoc subjective judgment is negated, as a single ordinal variable for each aspect allows for direct statistical analysis. The individual measures were combined by counting how many of them showed improvement to create a single, ordinal variable (e.g., if a participant showed any degree of improvement in 4 out of 5 measures of fluency, their ordinal score for fluency would be "4"). The common objective measures to be combined into the single ordinal variable were chosen by selecting representative measures from previous literature that measure the representative aspect in a theoretically different way. This selection method was used because choosing two variables that essentially measure the same thing would result in participants who improved in one variable to improve by two points in the combined ordinal scale. For example, if speech rate as measured in words per minute and speech rate as measured in syllables per second were both used as contributing variables to the single ordinal variable, it is almost guaranteed that a participant who improved in one would also improve in the other because they are measuring the same aspect,

simply in different ways. The same is not true for related variables that do not test the same thing. For example, if speech rate and the number of pauses are both used as contributing variables to the ordinal measure of fluency, it is not guaranteed that a participant who used fewer pauses necessarily improved their speech rate (e.g., they could have had fewer, but much longer pauses which would result in no improvement in speech rate). To ensure that no variables contributing to each ordinal variable were measuring the same thing, I conducted Pearson's correlation tests of the objective measures for each aspect of oral proficiency and allowed only one strong correlation (over 0.8) amongst the measures.

The first aspect of oral proficiency that must be measured is fluency, which generally refers to how quickly and without pause a speaker can iterate ideas. This has been measured in a number of different ways, the first of which is speech rate, i.e. the number of utterances as a function of time. One of the most intuitive ways to measure this is to simply look at the number of syllables per second, which is referred to as raw speech rate (RawSR), can be calculated automatically (i.e., de Jong & Wempe, 2009) and been used in a number of studies (e.g., Spring, 2020a; b). Another method of calculating fluency is trimmed speech rate (TrimmedSR), which is the number of utterances minus fillers and restarts per second (i.e., Tavakoli & Skehan, 2005). Many other studies have also offered the number of pauses (NPause), articulation ratio (ArtRat), i.e. the speed at which a participant can create a single utterance, and phonation ratio, (PhonRat) i.e., the percentage of time spent speaking, as measures of fluency (e.g., Lennon, 1990; Spring, 2020b; Wood, 2010). Following Spring (2020a; b), these measures were all calculated with Praat (Boersma & Weenink, 2020) except the trimmed speech rate (TrimmedSR), which cannot be calculated automatically, and was instead done by hand following Lennon (1990). Table 1 shows the only correlation in pretest scores higher than 0.8 is between raw speech rate and phonation ratio. As each of these measures theoretically measures different aspects of fluency and the correlations are not overly strong, these five measures were used to create the ordinal variable of fluency improvement, with a range from 0 to 5.

Table 1. Correlation of pretest fluency measures

RawSR	.638**			
NPause	-.538**	-.618**		
ArtRat	.429**	.636**	-.216**	
PhonRat	.581**	.848**	-.675**	.153
	TrimmedSR	RawSR	NPause	ArtRat

**significance at p<.01, *at p<.05

Two varieties of spoken complexity are generally measured for L2 research: syntactical and lexical. Syntactical complexity (SyntaxC) is often measured as the number of clauses per analysis of speech (AS) unit (e.g., Foster et al., 2000; Spring, 2020b), and this is was calculated for this study as well. Several measures of lexical complexity are offered by a number of different studies, but this study follows Spring (2000a; b) and utilizes Lu's (2012) free tool for calculating various measures of lexical complexity that he found to be highly correlated with better L2 speaking ability. However, since there is considerable overlap in some the nine measures suggest by Lu (2012), this study follows (Spring 2000a; b) and employs the highest correlated measure for each aspect of lexical complexity: the number of different words (NDW), the number of different words in random samples of 50 words (ES50), a corrected type-token ratio (CTTR), and a corrected measure of verb variance (CVV1). Table 2 shows the amount of correlation between these five measures. The only correlation over 0.8 observed in the pretest scores was between the number of different words and

corrected type-token ratio. Because these variables all look at rather different aspects of complexity and there was not too much correlation between them, they were all incorporated into the ordinal variable of complexity improvement, with a range from 0 to 5.

Table 2. Correlation of pretest complexity measures

NDW	.269**			
ES50	.009	.113		
CTTR	.174	.882**	.429**	
CVV1	-.040	-.397**	.548**	.014
	SyntaxC	**NDW**	**ES50**	**CTTR**

**significance at p<.01, *at p<.05

Finally, though accuracy is generally only measured syntactically, Spring (2020a; b) also uses an objective measure of pronunciation accuracy, repeated here. Syntactical accuracy is often measured as the percentage of error-free clauses (e.g., Tavakoli & Skehan, 2005), and pronunciation accuracy has been measured as the percentage of words that could be correctly transcribed by automatic speech recognition (Spring 200a; b). Following Spring (2020a; b), two research assistants counted the number of clauses with grammatical errors and I made the decision when there were any discrepancies between them. The total number of clauses was then counted and I then calculated the ratio of error free clauses. Pronunciation accuracy was calculated semi-automatically using an automatic speech rater and a human to check the number of true differences, according to Spring (2020b). While these two measures address entirely different types of accuracy, Table 3 shows that they exhibit a small amount of correlation in the pretest scores, and not enough to discount either measure. Therefore, these measures were combined into a single ordinal measure of accuracy improvement, with a possible range of 0 to 2.

Table 3. Correlation of pretest accuracy measures

	Pronunciation Accuracy
Syntactic Accuracy	.190*

**significance at p<.01, *at p<.05

Data Analysis

In order to determine the amount of similarity between instructor-based and scores and peer-based participation scores, first a Spearman's rank correlation test was used to observe how well correlated instructor-given participation scores and the overall Team-Q scores were. This was done because the primary purpose of the study was to observe correlation between instructor-based scores and peer-based scores, which are best represented by the overall Team-Q scores, and both sets of data are best considered to be measured ordinally. An ordinal regression analysis was then performed to compare instructor-based scores to the individual category scores within the Team-Q metric to see if the instructor-based scores were more likely based on certain aspects of participation.

In order to answer the second research question, as to whether or not certain participation scores were correlated with more improvement in certain aspects of oral proficiency, individual Spearman's rank correlation tests were used to compare the instructor-given participation scores and the overall Team-Q scores to each of the three ordinal measures of oral proficiency improvement. This was done because each measure of ordinal proficiency is completely separate,

and improvement is unlikely to happen unanimously across these variables according to *the trade-off hypothesis* (Skehan & Foster, 1997). An ordinal regression analysis was not used for looking at overall participation scores because instructor-based and peer-based participation scores supposedly measure the same thing, which violates the assumption of an ordinal regression analysis. However, ordinal regression analyses were then used to compare each ordinal measure of improvement to the individual category scores from the Team-Q metric because these supposedly measure different aspects of participation.

Results

The results of the Spearman's rank correlation tests between instructor-based participation scores and peer-based participation scores (overall Team-Q scores) showed a very weak correlation between the two; r = .187, p = .053. The results of the ordinal regression analysis comparing instructor-based participation scores and the individual categories of the Team-Q metric is shown in Table 4. Overall, the model was significant, but could only explain a small percentage of the variance; r^2 = .169, p = .001. The category scores found to contribute significantly to the model were "contributes" (p = .004), "facilitates contributions of others" (p < .001), and "manages potential conflict" (p = .04), although it should be noted that facilitating the contributions of others actually shared a negative correlation with instructor-given participation scores. Therefore, it seems that the instructor-based scores and the overall Team-Q scores were accounting for rather different aspects of participation, and that the instructor-based scores were likely (i) more focused on contributions to group work and conflict management than other areas, and (ii) associating a negative score to more effort to facilitate the contributions of others.

Table 4. Correlation between instructor based-participation scores and Team-Q individual category scores

Category	***B***	**Wald**	***p***
Contributes	.791	8.410	.004*
Facilitates	-1.339	17.315	< .001*
Plans	.093	.157	.692
Fosters	.294	.751	.386
Manages	.525	4.229	.040*

*significance at p<.05

The results of the Spearman's rank correlation tests between the two types of participation scores (i.e., instructor-based and peer-based, as measured by overall Team-Q scores) and improvement in the three aspects of L2 oral proficiency (i.e., fluency, complexity, and accuracy) are shown in Table 5. According to these results, instructor-based participation scores are correlated with improvement in fluency (r = .266, p = .005), but overall peer-based participation scores were not positively correlated with increase in any measure of oral proficiency, and were actually negatively correlated with improvement in fluency (r = -.210, p = .029).

Table 5. Correlation between participation scores and improvement in L2 oral proficiency

Participation Score	Fluency		Complexity		Accuracy	
	r	*p*	*r*	*p*	*r*	*p*
Instructor-based	.266	.005**	-.141	.147	.020	.839
Peer-based	-.210	.029*	-.156	.106	.043	.659

**significance at p<.01, *at p<.05

The ordinal regression analysis comparing the peer-based scores of individual categories of participation, as measured by Team-Q, showed that the model for fluency improvement was significant with very weak correlation (r^2 = .169, p = .001) but that the models for complexity (r^2 = .081, p = .248) and accuracy (r^2 = .026, p = .723) were not. The results for the contribution of individual category scores to the respective models are shown in Table 6. According to these results, the individual categories of teamwork seem to have little correlation to improvement in any of the areas of oral proficiency, with the exception of "facilitates contributions of others," which shows a strong negative correlation that likely explains the majority of the negative correlation between the overall peer-based scores and improvement in fluency.

Table 6. Correlation between categorical participation scores and improvement in L2 oral proficiency

Category	**Fluency**			**Complexity**			**Accuracy**		
	B	***Wald***	***p***	***B***	**Wald**	***p***	***B***	***Wald***	***p***
Contributes	.322	1.514	.219	-.078	.089	.765	-.115	.163	.687
Facilitates	-.702	5.428	.020*	-.276	.872	.351	.392	1.478	.224
Plans	.138	.347	.556	-.412	3.003	.083	-.016	.004	.951
Fosters	-.080	.057	.811	-.109	.108	.743	-.547	2.210	.137
Manages	-.005	.000	.985	.493	3.808	.051	.012	.002	.964

**significance at p<.01, *at p<.05

Discussion and Conclusion

According to the results of this study, instructor-based participation scores in an EFL class were quite different from peer-based scores, as taken via the Team-Q metric. The correlation analysis between the individual category scores in the Team-Q metric and the instructor-based scores suggest that one reason for this is that the instructor only considered certain aspects of participation or weighed them more heavily, specifically contributing to the groupwork and managing conflicts within the group. This could be true, as these skills are more likely to result in speaking turns which would then be noticed by the instructor. Specifically, the planning category could be done outside of class time or performed via writing, and fostering a team climate has more to do with attitude, which may not result in more speaking turns. Finally, facilitating the contributions of others may in fact be misconstrued by an instructor as an act of not taking a turn in group work discussions, and could partially explain the strong negative correlation between this category and instructor-based participation scores.

However, there are also a number of potential reasons for the differences in the instructor-based and peer-based participation scores in this study. First, the Team-Q metric was not specifically designed with EFL in mind, and simply asks whether students were active in the groupwork, not whether or not they did it in their target language. Therefore, students who participated actively in the group in the L1 may have received high scores from their peers, but not from their instructor. Furthermore, since the instructor was continually monitoring several groups at once, it is very possible that he simply could not accurately judge the level of participation by each student. Finally, though both the instructor-based scores and peer-based Team-Q scores were based on subjective judgements, the instructor-based scores were all assigned by the same single person, whereas the peer-based Team-Q scores might have varied from group to group if some groups were simply more critical overall than others.

Next, it seems that enhanced in-class participation in groupwork, as measured by instructor-based scores, seems to be somewhat correlated with improvement in fluency, but not complexity or accuracy, which is only somewhat congruent with Spring (2020a). Though Spring (2020a) seems to suggest that instructor-based participation scores are not as informative as peer-based measures in predicting improvement in fluency, he determined that these results were largely inconclusive. This is probably due to the fact that his results were largely skewed due to the inclusion of various other variables in his models as participation scores were not the focus of his study, and the fact that he includes data from students who were sometimes absent and tardy. This study looks more specifically at in-class participation by excluding the data of students who may have been affected by missing class, and also by looking exclusively at participation scores in finer detail.

It is also possible that since instructor-based participation scores and peer-based Team-Q scores are measuring different aspects of participation they correlate to oral proficiency in different ways. For example, since any student can theoretically attempt to participate in the groupwork in their L2, the instructor-based score may be looking more at student efforts to communicate whereas the peer-based Team-Q scores are more closely related to success in communication. If this is true, peer-based Team-Q scores may be associated with high levels of fluency, but would probably result in students with higher Team-Q scores also having higher beginning levels of fluency. Due to the nature of diminishing returns, students with higher beginning levels are less likely to exhibit improvement, which would affect the correlation between Team-Q scores and improvement. To test this theory, I performed a Spearman's rank correlation test on the five original measures of fluency in participants' pretest scores and the various measures of participation, the results of which are summarized in Table 7.

Table 7. Correlation between participation scores and participants' pretest measures of fluency

	Trimmed SR		Raw SR		No Pauses		Art Rat		Phon Rat	
	r	*p*	*r*	*p*	*r*	*p*	*r*	*p*	*r*	*p*
Instructor-based	-.007	.942	-.159	.100	.125	.197	-.045	.647	-.150	.123
Overall peer-based	.364	.000**	.148	.126	.057	.555	.121	.213	.142	.141
Contributes	.311	.001**	.065	.507	.077	.428	.096	.325	.058	.553
Facilitates	.333	.000**	.179	.063	.020	.835	.140	.147	.166	.086
Plans	.287	.003**	.091	.350	.142	.142	.078	.422	.090	.355
Fosters	.409	.000**	.263	.006**	-.098	.313	.246	.01*	.202	.036*
Manages	.326	.001**	.115	.237	.040	.681	.054	.578	.138	.156

**significance at p<.01, *at p<.05

The results of Table 6 suggest that having a high trimmed speech rate, i.e., being able to produce more meaningful utterances per second, before the course is highly correlated with higher peer-based participation scores in general. Furthermore, it shows that the subcategory "fosters a team climate" is correlated with most pretest measures of fluency. This means that having good L2 fluency likely results in higher peer-based participation scores, but not necessarily in improvement. Conversely, the instructor-based participation scores are not correlated with pre-test scores, and thus the areas that the instructor scored with the rubric, i.e., actively attempting to speak the L2, are not necessarily associated with pretest measures of fluency. However, this begs the question – are these participation scores similarly associated with oral proficiency also after the course. This was tested by repeating the same analysis but with posttest scores, the results of which are given in Table 8.

Table 8. Correlation between participation scores and participants' posttest measures of fluency

	Trimmed SR		**Raw SR**		**No Pauses**		**Art Rat**		**Phon Rat**	
	r	***p***	***r***	***p***	***r***	***p***	***r***	***p***	***r***	***p***
Instructor-based	.069	.475	.021	.827	-.013	.894	.204	.034*	-.094	.332
Overall peer-based	.119	.220	.041	.677	.182	.059	.064	.512	.021	.829
Contributes	.051	.603	.010	.921	.203	.035*	.065	.505	-.039	.688
Facilitates	.087	.371	.019	.846	.162	.094	.046	.639	.008	.936
Plans	.100	.302	.034	.727	.179	.064	-.007	.943	.046	.638
Fosters	.223	.020*	.124	.202	.061	.529	.152	.116	.074	.444
Manages	.120	.215	.016	.866	.163	.092	.089	.362	.006	.949

**significance at p<.01, *at p<.05

The results of Table 8 differ vastly from those of Table 7. First of all, the overall peer-based scores do not seem to show correlation with any posttest metrics, while the instructor-based scores do show correlation with one metric, namely articulation rate. This result somewhat supports the notion that the instructor-based scores are more indicative of improvement than the peer-based scores. Next, the two sets of results reveal that the same overwhelming trend of correlation between trimmed speech rate and every category of the peer-based scores found in pretest scores is not present in posttest scores. This supports the idea that these correlations are found because students who are naturally fluent before the class begins will be more likely to participate actively. Finally, though the Team-Q category "fosters a team climate" was highly associated with higher fluency before the class, it was not so closely correlated with fluency in the posttest.

Summatively, the results of this study suggest that instructor-based participation scores are more likely to be associated with improvement in oral fluency, but not necessarily with final oral proficiency achievement. Conversely, peer-based participation scores, as measured by Team-Q, are highly associated with pre-treatment measures of fluency, but not necessarily with improvement or post-treatment oral fluency.

Based on the results of this study, it seems that some caution is necessary when using participation scores in the EFL classroom. While instructor-based scores seem to be the current norm, especially in oral communication classes (Crosthwaite et al., 2015; Rogers, 2011), and do seem to have some

correlation to increases in oral fluency, they (i) are not associated with improvement in either oral complexity or accuracy, and (ii) are not very representative of actual skill level. Since there are inherent individual differences in learning styles and even participation styles (e.g., Dörnyei, 2005; Richards, 2014), it is conceivable, based on the results of this study, that some students improve their oral proficiencies through types of participation that are not visible to instructors, such as active listening, or areas that were not correlated with instructor-based scores in this study such as planning, fostering a good climate, or facilitating the contributions of others.

Though some researchers, such as Mello (2010) and Heyman and Sailors (2011) suggest that peer-based participation scores may be more appropriate, the results of this study suggest that caution is necessary here as well. Specifically, the peer-based Team-Q metric seemed be able to identify which students likely had high levels of fluency before the class, but were not very strongly associated with either improvement or fluency levels after the class. Therefore, these scores simply tell us that students with higher proficiency are more likely to be active participants, and do not actually provide a reliable measurement for either improvement or final proficiency levels. This could be in part due to the particular metric used in this study (i.e., Team-Q), and other metrics might provide different results. However, the Team-Q scores and some of the subcategories had clear associations with pre-treatment levels of fluency and clear negative associations with improvement. Therefore, it is not likely that the measurements provided by the Team-Q metric were simply erroneous. Future studies hoping to shine further light on this should take these results into account.

Therefore, before instructors include participation scores in the composite evaluations of their EFL classes, they should perhaps ask themselves what the purpose of the score is in their class. According to the results of this study, it seems that participation scores might not be a good metric for assessing students' final outcomes, as neither instructor-based or peer-based scores in this study seemed to have much association with final levels of oral fluency. If instructors wish to assign class scores based on actual ability, it might be best not to include participation in students' final composite scores, as it would likely pollute the final evaluation of skills or ability with construct irrelevant variance (McNamara, 2000), which could potentially affect the fairness of their final grade (Gipps & Stobart, 2006). On the other hand, teachers need to encourage students to actively try to improve, and the instructor-based scores in this study did seem to have some association with improvement. However, a participation score may not be the best solution here. One reason is that the association found in this study, though statistically significant, was not a very strong one. Furthermore, as pointed out by Crosthwaite et al. (2015), instructor-based scores can conflict with students' particular learning styles, which might cause some students to feel the scores to be unfair or demotivate those with less assertive personalities.

This study was able to bring more light to the issue of participation scores in EFL classes and their connection, or lack thereof, to both skill level and improvement. However, there are several issues beyond the scope of this paper. First, it was unable to discern if there are actual, tangible differences in motivation to participate actively when participation scores are or are not included in a course. Therefore, more research is required to examine the effects of including participation scores (or not) in the final composite grades of on students with a variety of learning styles. Second, the current study was unable to provide a practical and fair solution to the problem of assessing in-class participation. Future studies could perhaps work to refine a participation rubric, based on the results of this study, that would more accurately assess in-class participation in a way that would allow it to be included more fairly in composite class scores.

About the Author

Ryan Spring is an associate professor in the Institute for Excellence in Higher Education at Tohoku University. His research interests include applications of cognitive linguistics to second language acquisition, objectively measuring speaking and writing, and teaching and the use of multimedia in EFL teaching. He currently serves as the vice-president of the East Japan chapter of the Association for Teaching English through Multimedia.

Acknowledgements

I would like to thank the editorial staff of this journal and the two blind reviewers who provided valuable feedback and suggestions for improving this paper.

To cite this article

Spring, R. (2021). Instructor- versus peer-based participation scores in EFL classes: Comparisons and correlation to oral proficiency. *Teaching English as a Second Language Electronic Journal (TESL-EJ), 25*(3). https://tesl-ej.org/pdf/ej99/a15.pdf

References

Boersma, P., & Weenink, D. (2020). Praat: Doing phonetics by computer. http://www.praat.org

Brutus, S., & Donia, M. B., (2010). Improving the effectiveness of students in groups with a centralized peer evaluation system. *Academy of Management Learning & Education, 9*(4), 652–662. https:/doi.org/10.5465/amle.9.4.zqr652

Britton, E., Simper, N., Leger, A., & Stephenson, J. (2017). Assessing teamwork in undergraduate education: A measurement tool to evaluate individual teamwork skills. *Assessment & Evaluation in Higher Education, 42*(3), 378–397. https:/doi.org/10.1080/02602938.2015.1116497

Crosthwaite, P. R., Bailey, D. R., & Meeker, A. (2015). Assessing in-class participation for EFL: Considerations of effectiveness and fairness for different learning styles. *Language Testing in Asia, 5*, 9. https:/doi.org/10.1186/s40468-015-0017-1

Dancer, D, & Kamvounias, P. (2005). Student involvement in assessment: a project designed to assess class participation fairly and reliably. *Assessment & Evaluation in Higher Education, 30*(4), 445–454. https://doi.org/10.1080/02602930500099235

Dörnyei, Z. (2005). *The Psychology of the Language Learner: Individual Differences in Second Language Acquisition.* Lawrence Erlbaum Associates.

de Jong, N., & Wempe, T. (2009). Praat script to detect syllable nuclei and measure speech rate automatically. *Behavior Research Methods, 41*, 385–390. https:/doi.org/10.3758/BRM.41.2.385

Foster, P., Tonkyn, A., & Wigglesworth, G. (2000). Measuring spoken language: A unit for all reasons. *Applied Linguistics, 21*, 354–375. https:/doi.org/10.1093/applin/21.3.354

Gipps, C., & Stobart, G. (2006). Fairness in Assessment. In C. Wyatt-Smith & J. J. Cumming (Eds.), *Educational Assessment in the 21st Century* (pp. 105–118). Springer. https:/doi.org/10.1007/978-1-4020-9964-9_6

Heyman, J., & Sailors, J. (2011). Peer assessment of class participation: Applying peer nomination to overcome rating inflation. *Assessment & Evaluation in Higher Education, 36*(5), 605–618. https:/doi.org/10.1080/02602931003632365

Housen A., & Kuiken, F. (2009). Complexity, accuracy, and fluency in second language

acquisition. *Applied Linguistics, 30*, 461–478. https:/doi.org/10.1093/applin/amp048

Kang, S. (2005). Dynamic emergence of situational willingness to communicate in a second language. *System, 33*(2), 277–292. https:/doi.org/10.1016/j.system.2004.10.004

Lennon, P. (1990). Investigating fluency in EFL: A quantitative approach. *Language Learning, 40*, 387–417. https:/doi.org/10.1111/j.1467-1770.1990.tb00669.x

Lu, X. (2012). The relationship of lexical richness to the quality of ESL learners' oral narratives. *The Modern Language Journal, 96*(2), 190–208. https:/doi.org/10.1111/j.1540-4781.2011.01232.x

MacInteyre, P.D., & Charos, C. (1991). Personality, attitudes, and affect as predictors of second language communication. *Journal of Language and Social Psychology, 15*, 3–26. https:/doi.org/10.12691/education-2-11-8

McNamara, T. (2000). *Language Testing*. Oxford University Press.

Mello, J. A. (2010). The good, the bad and the controversial: The practicalities and pitfalls of the grading of class participation. *Academy of Educational Leadership Journal, 14*(1), 77–97.

Morell, T. (2007). What enhances EFL students' participation in lecture discourse? Student, lecturer and discourse perspectives. *Journal of English for Academic Purposes, 6*(3), 222–237. https:/doi.org/10.1016/j.jeap.2007.07.002

Richards, J.C. (2014). *Key Issues in Language Teaching.* Cambridge University Press.

Rogers, S.L. (2011). *Grading participation in college courses: Instructor attitudes and practices.* Doctoral dissertation. State University of New York at Albany.

Skehan, P., & Foster, P. (1997). Task type and task processing conditions as influences on foreign language performance. *Language Teaching Research, 1*, 185–211. https:/doi.org/10.1177/136216889700100302

Skehan, P. (2009). Modelling second language performance: Integrating complexity, accuracy, fluency, and lexis. *Applied Linguistics, 30*(4), 510–532. https:/doi.org/10.1093/applin/amp047

Skehan, P. (2014). The context for researching a processing perspective on task performance. In P. Skehan (Ed.), *Planning and Task Performance in a Second Language* (pp. 111–141). John Benjamins.

Spring, R. (2020a). Maximizing the benefits of video-creation PBLL in the EFL classroom: A preliminary analysis of factors associated with improvement in oral proficiency. *STEM Journal, 21*(4), 107–126. https:/doi.org/10.16875/stem.2020.21.4.107

Spring, R. (2020b). Can video-creation project work affect students' oral proficiency? An analysis of fluency, complexity and accuracy. *TESL-EJ, 24*(2), 1–17. http://www.tesl-ej.org/wordpress/issues/volume24/ej94/ej94a1/

Suter, R. (2006). Predictors of pronunciation accuracy in second language learning. *Language Learning, 26*(2), 233–253. https:/doi.org/10.1111/j.1467-1770.1976.tb00275.x

Tavakoli, P., & Skehan, P. (2005). Strategic planning, task structure, and performance testing. In R. Ellis (Ed.), *Planning and Task Performance in a Second Language* (pp.239–273). John Benjamins. https:/doi.org/10.1075/lllt.11.15tav

Vandrick, S. (2000, March 14–18). *Language, culture, class, gender and class participation.* [Paper presentation]. 34th Annual Meeting of Teachers of English to Speakers of Other

Languages, Vancouver, Canada. Available at https://files.eric.ed.gov/fulltext/ED473086.pdf

Wood, D. (2010). *Formulaic language and second language speech fluency: Background, evidence and classroom applications.* Continuum.

Yashima, T. (2002). Willingness to communicate in a second language: The Japanese EFL context. The Modern Language Journal, 86(1), 54–66. https://doi.org/10.1111/1540-4781.00136

Zhou, N. (2015). Oral participation in EFL classroom: Perspective from the administrator, teachers and learners at a Chinese university. *System, 53,* 35–46. https:/doi.org/10.1016/j.system.2015.06.007

Appendix 1: Team-Q Questions

How often does your peer demonstrate the following?

(0 = Never, 1 = Sometimes, 2 = Usually, 3 = Regularly, 4 = Always)*

*In this study, "Always" was defined as peers having done it during 9 or 10 of the groupwork classes, "Regularly" was defined as peers having done it during 7 or 8 of the group work classes, etc.)

Subcategory: Contributes to team project

1. Participates actively and accepts a fair share of the group work
2. Works skillfully on assigned tasks and completes them on time
3. Gives timely, constructive feedback to team members, in the appropriate format

Subcategory: Facilitates contributions of others

4. Communicates actively and constructively
5. Encourages all perspectives be considered and acknowledges contributions of others
6. Constructively builds on contributions of others and integrates own work with work of others

Subcategory: Planning and management

7. Takes on an appropriate role in group
8. Clarifies goals and plans the project
9. Reports to team on progress

Subcategory: Fosters a team climate

10. Ensures consistency between words, tone, facial expression and body language
11. Expresses positivity and optimism about team members and project

Subcategory: Manages potential conflict

12. Displays appropriate assertiveness: neither dominating, submission, nor passive aggressive
13. Contributes appropriately to healthy debate
14. Responds to and manages direct/indirect conflict constructively and effectively

No subcategory:

15. Generally, how would you rate your peer on the effort they put into team tasks, their manner of interacting with others, and the quantity and quality of contributions they make to team discussions?
    0 = Unacceptable, 1 = Poor, 2 = Acceptable, 3 = Good, 4 = Excellent

[back to article]