Pan-Pacific Association of Applied Linguistics 19(1), 195-213

Assessment Behavior and Perceptions of Raters in Paired and Group Oral Interaction

Junko Negishi
Tsurumi University

Negishi, J. (2015). Assessment behavior and perceptions of raters in paired and group oral interaction. *Journal of Pan-Pacific Association of Applied Linguistics*, *19*(1), 195-213.

The study considers the assessment of L2 English learners by trained raters in paired and group oral assessments in comparison to an individual, monologue assessment, to determine 1) the degree to which raters assign pairs/groups shared (the same) scores and the degree to which raters give individual members of pairs/groups higher or lower as opposed to matching individual scores, 2) assessment tendencies by participants' English proficiency and paring/grouping conditions, 3) individual variation among raters, and 4) the types of comments (positive, negative, or mixed) raters give participants in pair/group conditions. It was found that 1) on the whole, pair/group scores tend to be shared and lower than individual scores, 2) raters tended to assign both higher- and lower-level speakers lower scores when they are in interaction with lower-level speakers, but to assign middle-level speakers better scores in such a condition, 3) raters exhibited significant inter-rater variability, and 4) raters tended to give more negative comments when assigning lower scores and more positive comments on higher scores.

Key Words: assessment, oral performance, pair, group, interaction

1 Introduction

Assessing oral performance in communicative situations has been considered one of the most difficult assessment skills (Shohamy, Reves, & Bejarano, 1986); nevertheless, assessing second language (L2) learners' oral performance has become increasingly important in the era of globalization, in which English, for instance, is used as an international language.

Oral performance tests have mainly been administered by employing a task to be completed by a single test-taker, such as a picture description, reading aloud, or speech task. Interviews, in which an interviewer elicits a test-taker's talk, are also regarded as an appropriate test format for a single test-taker. In contrast, test batteries employing paired or group oral interactions (*paired/group orals* hereafter) have not been commonly administered, since interactions between/among test-takers are regarded to have various defects as an approach to language testing. First, uncontrollable

variables between or among test-takers such as differences in levels of language proficiency, familiarity between interlocutors, personality, willingness to speak, age, and gender often have effects on the interaction (cf. Iwashita, 1996; Berry, 2004; Bonk & Van Moere, 2004; Brooks, 2009; Davis, 2009). The quality of interlocutors in paired/group orals (that is, their skill as interaction partners) might be another drawback, in that an interlocutor may disrupt the other(s) as a result of a lack of training in controlling the interaction or eliciting a required utterance, in contrast to interviewers in single-speaker tests, who will generally have this training (Brooks, 2009; Van Moere, 2006). There are also disadvantages that may emerge from the rater's behavior when assessing multiple speakers; for example, raters in this situation generally demonstrate lower inter-rater agreement (Van Moere, 2006) and inconsistent severity/leniency (Bonk & Ockey, 2003). Another deficit is that assessing multiple speakers remains difficult in terms of both fairness and validity (Iwashita, 1996) because of the complexity of the context surrounding the interaction, which creates interlocutor effects between speakers. These variables may have a smaller impact in a single test-taker test format.

Despite these drawbacks, paired orals have been administered as part of the main suites of high-stakes tests such as the Cambridge Assessment by the University of Cambridge Local Examinations Syndicate (UCLES). The Council of Europe (2001) has provided assessment criteria for paired and group orals as well as a single-speaker oral performance assessment within the Common European Framework of Reference (CEFR). Some countries in Asia report that they have used group discussions in this way; for example, in Hong Kong the Hong Kong Use of English Test has been administered, in mainland China the College English Test has been implemented, and in Korea the Educational Testing Service (ETS) has been using group interviews to select scholarship recipients. Thus, paired/group orals have been used increasingly in test batteries in recent years; nevertheless, they are still not common in many countries, including Japan. One reason for choosing not to administer tests in a multiple-speaker format is the presence of various assessment difficulties, which will be described in the next section.

This paper will explore some features of oral performance assessment by looking at raters' rating behavior and perceptions. By doing so, this study is intended to contribute to educational practice at institutions in which teachers wish to implement paired/group orals as an oral performance test.

2 Background

Various concepts and frameworks of communicative competence or (communicative) language ability, advocated by many researchers (cf. Canale & Swain, 1980; Canale, 1983; Bachman & Palmer, 1996), have been introduced in the field of L2 learning and assessment. The common ground

between these concepts and frameworks when performance-based tests are administered is the *individual*. Individuals are generally regarded as being capable of applying their language ability in any situations or contexts. In a multiple-speaker test format, many scholars think it appropriate to rate test-takers as individuals (cf. Shohamy, Reves, & Bejarano, 1896; Fulcher, 1996; Bonk & Ockey, 2003). In contrast, however, McNamara (1996) has pointed out that one weakness of models of communicative competence is that they focus too much on the individual speaker rather than interlocutors in interaction.

Kramsch (1986) thought that there was a considerable degree of disparity between the interactive activities that were being encouraged in language classrooms and the focus on the individual maintained by performance-based tests; on this basis, she advocated the constructive concept of *interactional competence.* Based on the notion of interactional competence, Jacoby and Ochs (1995) described the "co-constructed" nature of interaction and asserted that it is essential for a successful conversation. A co-constructed interaction creates a state of "negotiation of meaning" between/among interlocutors that promotes second language acquisition. On the basis of the concept of interactional competence, researchers have suggested that collaborative interaction between speakers be tested, as their conversation produces rich natural language from which it is easier to elicit various features of interactional competence that may be difficult to draw from interviews (Canagarajah, 2006; May, 2009: Van Moere, 2006). Researchers also raise potential difficulties around separating out each interlocutor in an interaction. For instance, Fulcher argues, "If talk in second language speaking tests is co-constructed… we have to ask many questions, such as how scores can be given to an individual test-taker rather than pairs of test-takers in a paired test format" (2003, p. 46). As a result, interactionists argue that all speakers in multiple-speaker interactions should be assessed together, as the interaction is not an individual performance but a joint achievement of the interlocutors (May, 2009; Swain, 2001; Young, 2000; Weir, 2005).

One researcher who has conducted practical explorations of L2 learners' discourse and interactional patterns and raters' quantitative scores in a paired test is Galaczi (2004, 2008). Based on Storch's dyadic (2002) model, Galaczi analyzed the discourse produced by candidates taking the Cambridge First Certificate in English (FCE) speaking test. The patterns of interaction identified in Galaczi's study were three: *collaborative, parallel*, and *asymmetrical*. According to Galaczi's analysis, collaborative interaction resulted in the highest Interactive Communication (IC) scores on the FCE of the three interaction types, whereas parallel interaction resulted in the lowest. In Galaczi's study, IC scores were given to individual candidates in a pair.

Another practical study was carried out by Negishi (2011) who conducted research with Japanese learners from the junior high school level

to the university level, grouped in threes. Ten English teachers assessed the students using the Common European Framework of Reference[1] (CEFR; Council of Europe, 2001), for five qualitative aspects of their language use: *range*, *accuracy*, *fluency*, *coherence,* and *interaction.* Negishi modified the paired-interaction models of Galaczi (2004, 2008) to be used in group orals. Negishi's participants were novice- to intermediate-level learners whose English speaking levels were much lower than those of Galaczi's participants. As a result, in addition to the collaborative, asymmetric, and parallel interaction patterns, a rudimentary interactional pattern was revealed. In Negishi, assessment scores were given to individuals based on the CEFR criteria. In order to investigate the relationship between *interaction* scores—one of the qualitative aspects of spoken language use in the CEFR—and the interactional patterns, each interactional characteristic was replaced by a number (3 points for collaborative, 2 for parallel and asymmetric, 1 for rudimentary interaction) and the same shared score was given to all three members of the group irrespective of their individual performance. The correlation coefficient (Kendall's *tau*) between the shared score and the individual score for the CEFR *interaction* scores calculated for the group was .586, $p < .01$; that is, the CEFR *interaction* scores were quite significantly associated with the group's shared scores, and therefore their interactional patterns. Negishi's study thus demonstrated the plausibility of assigning shared scores to the interactional performance of a group.

Another type of research relates to the characteristics of scores assigned to multiple speakers by raters. In her paired interaction study, Brooks (2009) found that performance scores were closer to each other when paired than when students performed individually. Brooks inferred that some of her raters might have awarded shared scores to paired candidates unconsciously. Negishi (2011) also reported cases in which two or three members of a group were given shared scores. Thus, the studies of Brooks and Negishi both showed the possibility that raters assigned shared scores to speakers in a pair or group even when rating, and using rating scales designed for, individuals.

Whether or not an interlocutor's proficiency level affects rater scores, Iwashita (1996) found that raters assigned high-proficiency speakers higher scores when the speakers were paired with high-proficiency interlocutors. Raters also gave low-proficiency speakers better scores when they interacted with high-proficiency interlocutors compared to low-proficiency interlocutors. It should be noted, however, there was large variability among individuals. Reporting a similar result to that of Iwashita, Berry (2004) found that speakers' increased production of talk when paired with high-proficiency interlocutors did not lead to significantly different scores. Davis (2009) explored the impact of interlocutor proficiency on the FCE paired test:

[1] For details about the CEFR, refer to Section 3.4 (Assessment).

candidates were divided into higher-proficiency and lower-proficiency level learners, and had one conversation with an interlocutor of similar proficiency and another with an interlocutor of different proficiency. Rasch analysis ability measures did not show any significant effects of interlocutor proficiency; however, lower-proficiency speakers produced more talk when paired with higher-proficiency interlocutors.

Although there is research investigating the assessment behavior of raters and impacts of interlocutor's proficiency level, there is no study, to my knowledge, that has explored rater characteristics among the three test types, that is, a single-speaker test and paired and group orals.

3 The Study

3.1 Purpose of the Study

This study explored raters' assessment behavior and perceptions when scoring the three types of oral test: a single-speaker, monologue test; a paired oral; and a group oral. Four research questions (RQs) are taken up.

RQ1: How do raters' behaviors when assigning scores on a paired or group oral contrast with behaviors for scores given to a single speaker?
RQ2: How do raters assess speakers in terms of proficiency level and pairing/grouping patterns?
RQ3: Is there any inter-rater variability?
RQ4: What types of perceptions do raters have toward participants when assigning scores?

3.2 Participants

The study included a total of 24 participants, who were students at two universities, A and B, in and around the Tokyo area. The 12 participants from university A were recruited from an elective, online cross-cultural distance learning English course designed for intermediate- to advanced-level students, and included three returnees. Their TOEIC scores were relatively high compared with those of university B, between 635 and 960, although none of them were English majors. The other 12 participants, from university B, were all English majors with TOEIC scores between 300 and 690, and one with TOEFL iBT 90. The ratio of male to female students was six to six in university A and seven to five in university B. The participants were recruited in a way intended to assemble a sample that included a wide range of English-speaking abilities. The author, who was also the teacher of these 24 students in the instruction conducted for this experiment, explained the research project to the participants, and they signed written consent forms.

The project was approved by the Ethical Review Board at the author's university (#1121).

The three types of test were administered separately to each group of students (that is, by university) so that test-takers could be matched with an interlocutor whom they know, in an effort to equalize possible assessment variability due to familiarity (Foot, 1999; O'Sullivan, 2002).

3.3 Speaking Prompts

Three types of speaking task were used: an *oral monologue* task, a *paired interaction* consisting of two tasks, and a *group interaction* also with two tasks. In the monologue task, a single participant was presented with a four-panel cartoon taken from the Eiken[2] Grade Pre-1 test, with permission, depicting a protagonist's first week working at a company and the disappointing results. The participant was given one minute to think about the story and was then asked to describe it in two minutes. The result of the monologue task was used for pairing or grouping the participants, dividing them into three levels (higher-proficiency speakers, middle-proficiency speakers, and lower-proficiency speakers) in the subsequent oral performance tests (the paired and group interactions).

In the first paired task, a participant was paired with an interlocutor of a similar proficiency level and they were given a prompt, "family," and asked to talk about it. The second paired task was implemented among paired interlocutors of different proficiency levels, who talked about another prompt, "school." Interlocutors talked about the given topic for about four minutes while being video-recorded; interactions were cut to 200 seconds (100 seconds per person) on DVD for the subsequent ratings and analysis.

In the first grouped task, interlocutors of similar proficiency levels were placed in groups of three by proficiency level to discuss the prompt "dreams." For the second grouped task, one or two student(s) of different proficiency levels (out of the three) were matched to discuss the prompt "English." For example, one higher proficiency level student plus two lower proficiency level students might form a group of three. In these tasks, students were asked to talk for more than five minutes, and the video-taped talk was later cut to 300 seconds (100 seconds per person).

All the participants took part in the monologue (picture description) task; they participated in the paired and group oral tasks partially counterbalanced according to class attendance.

---

[2] The Eiken is a widely-used language assessment test for learners at all proficiency levels (that is, Grades 1 to 5 on the test); it is administered by the Eiken Foundation of Japan.

3.4 Assessment

The oral performance of the participants was assessed utilizing a holistic rating instrument, the CEFR-J[3], developed by Tono in 2013. The CEFR-J is the Japanese version of the CEFR, introduced in Section 2 above (Council of Europe, 2001). The original CEFR was developed to set out learning goals for language learning/teaching, syllabus development, curriculum, examinations, and textbooks, not only within the multi-linguistic and multi-cultural sphere of the European Union but also in other countries, that is, it was meant to be generally adaptable and applicable. Both the CEFR and CEFR-J include "Can Do" descriptors, which explicitly set out language learning goals in terms of the capabilities to be acquired by the learner to fulfill the goal. Japanese junior and senior high schools are being encouraged by the national Ministry of Education, Culture, Sport, Science, and Technology (MEXT) (2013) to compose and implement a "Can Do list" that responds to the school's specific circumstances. While the original CEFR includes "Common Reference Levels" providing a basic framework—A1 and A2 for Basic Users, B1 and B2 for Independent Users, and C1 and C2 for Proficient Users—the CEFR-J divides learners up differently by proficiency in order to fit the state of English learning in Japan, namely, level Pre-A1; levels A1.1, A1.2, A1.3, A2.1, and A2.2; levels B1.1, B1.2, B2.1, and B2.2; and levels C1 and C2. (The letters still correspond to the overall CEFR framework.) The reason the lower levels are more finely divided here is that about 80% of Japanese learners of English can be categorized into level A (Negishi, 2011; Negishi, Takada, & Tono, 2012). The CEFR-J criteria consist of five skills, namely, listening, reading, speaking (interaction), speaking (presentation), and writing. The CEFR-J rating criteria utilized for the study were "speaking: presentation" for the monologue task and "speaking: interaction" for the paired and group orals and participants' performance was assessed holistically. Both of the instruments, the CEFR and the CEFR-J, are designed to assess speakers *individually* even in pair/group situations.

All participant performances were rated by five Japanese raters, each of whom had been teaching English for more than 10 years and held at least an M.A. degree. None of the participants were students of the raters. Four out of the five raters had previous rating training and experience using the original CEFR criteria to rate participants in group discussions individually. All five raters, including the one with no previous training, trained together by watching the CEFR training video for paired/group orals (North & Hughes, 2003). The author told each rater that they were intended to assess the 24 participants in the three different test types but not to compare the participants within each test format. Then, the raters assessed some other speakers on a trial basis and conducted a vigorous discussion to reach

[3] The CEFR-J is downloadable for free from http://www.cefr-j.org/.

agreement on ratings using the CEFR-J criteria before the subsequent main assessment.

4 Results and Discussion

4.1 Assessment Behavior (RQ1)

First, we will explore the degree to which raters assigned pairs/groups shared scores. Table 1 shows the number of cases in which the raters assigned two or more speakers shared scores in the paired and group tasks, respectively, compared with the number of cases in which the raters gave identical scores when the above scores were replaced by those on the monologue task (numbers in parentheses).

Table 1. Number of Cases and Percentages Assigning Shared Scores

Pair/ Group	Comparison Description	Number of Cases	Percentage
Pair	Number of cases in which the raters assigned paired speakers shared scores	37	15.4%
	(Number of cases in which the raters assigned the above speakers identical scores when their paired oral scores were replaced by their monologue scores)	(17)	(7.1%)
Group[4]	Number of cases in which the raters assigned two or more of a group of speakers shared scores	53	22.1%
	(Number of cases in which the raters assigned the above speakers identical scores when their group oral scores were replaced by their monologue scores)	(32)	(13.3%)

Note: n = 240 (24 ratings x 2 conditions x 5 raters).

The results of the raters' actual assessments show 37 shared scores in the paired oral out of 240 assessments, which accounts for 15.4% of the total. When the monologue task scores were substituted, in contrast, the number of shared cases was 17 and 7.1%; that is to say, the raters assigned speakers in the paired oral shared scores 2.18 times more often than those conducting the

[4] In the group oral, a case was found where the raters assigned at least two of the three participants identical scores. For example, let us assume that Speakers A, B, and C conducted the monologue task separately and that A scored B1, B also scored B1, and C scored A2. Then, Speakers A, B, & C were grouped; in this task, A scored B1 and B & C were given a shared score, B2. In this case, the number of shared scores was 1 (the identical B2 score for Speakers B and C). However, if B and C's shared scores are replaced by their monologue scores, B1 for B and A2 for C, the count drops to 0, since the scores (B1 and A2) are now different.

picture description task. By the same token, while 53 shared scores, or 22.1%, were observed in the group oral, only 32, 13.3%, were observed when the monologue ratings were substituted; thus, shared ratings in groups were 1.66 times more common than in scores for the monologue task.

What we can draw from Table 1 is that in the paired oral, the overall possibility of the rater's assigning identical scores is more than double the baseline possibility in the monologue, while it is more than 1.5 times in the group oral. However, it should be noted that the higher possibility in group orals includes cases where only two of three members were assigned identical scores.

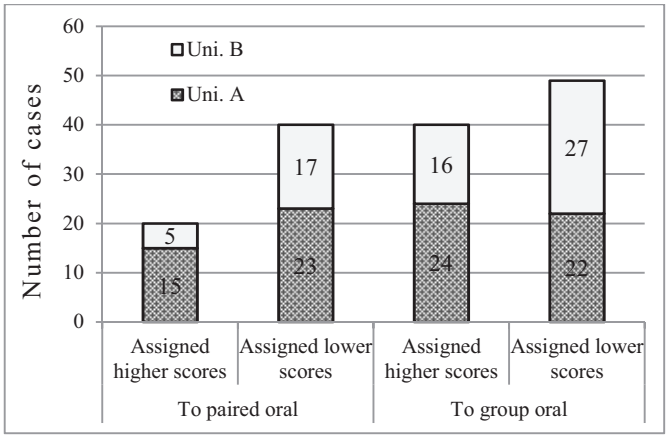

Figure 1. Tendency to assign higher or lower shared scores than individual scores

Second, the degree to which the raters assigned higher or lower scores as opposed to individual scores will be explored. Figure 1 shows the number of cases in which the raters assigned either higher or lower shared scores to pair or group members than when they assessed the monologue.[5] What can be drawn from the data for the paired condition is that the raters assigned lower ratings to more interlocutors compared with the monologue task; to be exact, 40 speakers (23 from university A + 17 from university B) received lower scores. The number of speakers assigned lower scores on the paired interaction than on the monologue was twice the number assigned higher

[5] The number of cases was comparable across pairs and groups, although the number of oral performances was different; 24 times for pairs and 16 for groups. That is because while the number of paired orals was 1.5 times that of group orals, the possibility of assigning identical scores in the group oral was 1.5 times that in the paired oral, because in the group oral, a case was counted as shared even if a rater only assigned identical scores to two speakers out of three.

scores, as cases of the latter kind were only 20 in number (15 from university A + 5 from university B). This phenomenon is especially prominent in university B (3.4 times as many lower as higher scores; 17 vs. 5, as compared to 1.5 times in university A; 23 vs. 15). Thus, the raters were more likely to assign lower scores to participants in the paired oral in university B whose proficiency level was relatively lower.

In the group oral, the disparity between higher and lower scores was much smaller: 1.2 times (assigned lower scores 22 in university A + 27 in university B = 49 and higher scores 24 in university A + 16 in university B = 40). In university A, the number of speakers assigned lower scores and higher scores was nearly equal (22 vs. 24), while in university B, the number of speakers given lower scores was 1.7 times that for higher scores (27 vs. 16).

On the whole, with regard to RQ1, the raters showed a tendency to assign participants shared scores and lower rather than higher scores in pairs and in groups compared to on the monologic picture description test.

4.2 Participants' Proficiency Level and Pairing/Grouping (RQ2)

The participants were divided into three levels in terms of proficiency, namely, higher-proficiency speakers, middle-proficiency speakers, and lower-proficiency speakers. As described in Section 3.3, the participants were paired/grouped once with (an) interlocutor(s) of a similar proficiency level and once with a different proficiency level. In this section, how raters assess speakers in terms of their proficiency level and paring/grouping conditions will be described, and assessment tendencies, first between the pairs and monologues and second between the groups and monologues, will be explained.

Table 2. Raters' Assessment Tendencies toward Participants' Proficiency Level and Pairing Patterns in Contrast to the Monologue Task

	Assigned *higher* scores		Assigned *lower* scores		Same scores
	Same proficiency	Different proficiency	Same proficiency	Different proficiency	
Higher-proficiency speakers	12.5%	0.0%	25.0%	56.2%	6.3%
Middle-proficiency speakers	25.0%	49.9%	6.3%	12.5%	6.3%
Lower-proficiency speakers	6.3%	18.8%	37.4%	31.2%	6.3%

Table 2 shows the scores, expressed in percentages, that the raters assigned to the participants in the paired oral in contrast to those in the monologue task within paring conditions, by the participants' proficiency level. When looking at higher-proficiency speakers, the raters showed a tendency to give lower scores when they were matched with different proficiency speakers; that is, 56.2% of the scores assigned to the paired speakers were lower compared with single-speaker scores when they interacted with middle- or lower-level interlocutors. In other words, higher-proficiency speakers were more likely to receive higher scores on the monologue task than on the paired tasks. From this, it can be inferred that advanced speakers are more experienced at producing language alone and are more comfortable with a solo task format. It can be inferred from this that, in Japan, learners tend to practice speaking tasks alone for the purpose of taking English speaking examinations, since multiple-speaker tasks are rarely given to test candidates—that is, higher-level learners might have practiced solo task formats more than lower-level learners. This seems to indicate that the higher-level participants may have "played down" to their interlocutors to accommodate their lower level.

Next, the raters were likely to give middle-level speakers higher scores in the paired format, especially when they interacted with speakers of different proficiency levels. Overall, three-fourths (25.0% + 49.9%) of speakers received higher scores compared with a monologue task. Different-proficiency pairs mostly meant pairings between middle-level and higher-level speakers. From this it can be inferred that middle-level speakers' proficiency level was elevated by the help of higher-level interlocutors.

Third, the raters tended to assign lower-level speakers lower scores. This might stem from these speakers' low proficiency, which might have caused breakdown between interlocutors. They may also have been more unfamiliar with this type of format compared with other-level speakers. They may therefore need more communicative, interactive activities in their classrooms.

Table 3 shows the grouping patterns, which indicates a similar result to that in Table 2. That is, the raters exhibited a tendency to assign lower scores to higher- and lower-proficiency participants and to assign higher scores to middle-level participants. Specifically, scores were salient when higher-level participants interacted with other-level interlocutors; 62.4% of the participants were given lower scores when matched with lower-level interlocutors. In contrast, 87.5% (18.8% + 68.7%) of the middle-level participants received higher scores compared with the monologue task irrespective of their group partners' proficiency level. Lower-level speakers did not indicate any prominent differences between conditions.

Table 3. Raters' Assessment Tendencies toward Participants' Proficiency Level and Grouping Patterns in Contrast to the Monologue Task

	Assigned *higher* scores		Assigned *lower* scores		Same scores
	Same proficiency	Different proficiency	Same proficiency	Different proficiency	
Higher-proficiency speakers	6.3%	12.5%	12.5%	62.4%	6.3%
Middle-proficiency speakers	18.8%	68.7%	0.0%	12.5%	0.0%
Lower-proficiency speakers	12.5%	25.0%	25.0%	37.5%	0.0%

Overall, in terms of assessment tendencies by participants' proficiency, the raters exhibited a tendency to rate middle-level speakers as having higher speaking ability and higher- and lower-level speakers with lower speaking ability compared with the monologue task.

4.3 Inter-Rater Variability (RQ3)

This section will look at inter-rater variability in terms of the scores the raters assigned to participants.

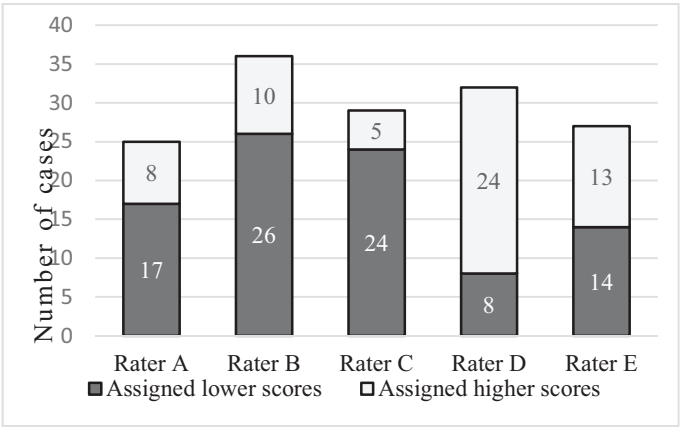

Figure 2. Number of cases assigning higher/lower scores on pair/group orals than on monologues

Figure 2 shows that among the five raters (A to E), raters A, B, and C showed a tendency to assign lower scores on paired or group orals compared to those on the monologue. Specifically, Rater A assigned 17 lower scores and 8

higher scores; Rater B, 26 and 10; and Rater C, 24 and 5. These three raters were moderate in terms of severity/leniency as determined by scores calibrated using Multi-Faceted Rasch Measurement (MFRM; Negishi, 2015). Rater D assigned a larger number of higher scores to participants in the paired and group orals than in the single-speaker task, that is, 24 higher scores vs. 8 lower scores. Because of this rating pattern, rater D was regarded as the most lenient rater by the MFRM analysis. Although rater E was determined to be the most severe rater, his number of higher and lower scores was balanced. Overall, rater A assigned the fewest different scores to different formats (17 + 8 = 25) and rater B, the most (26 + 10 = 36).

Despite the fact that all the raters were regarded as intra-rater consistent and no problematic bias was observed in the MFRM analysis, inter-rater variability does exist when looking into individual cases. This agrees with what McNamara (1996) mentioned: that it is difficult to eliminate inter-rater variability.

4.4 Rater Perceptions (RQ4)

Raters were asked to write down short comments while assessing the participants. This section will describe what types of comments raters made about participants when assigning scores. These data are presented only for cases where the rater assigned different scores (higher or lower) between the single-speaker monologue task and the paired/group orals.

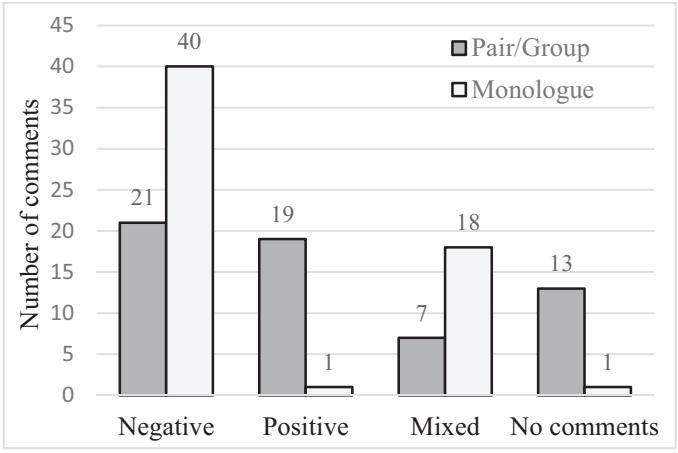

Figure 3. Rater comments when assigning *higher* scores on pair and group work than on monologues

Figure 3 shows the raters' perceptions, as expressed by their comments—negative, positive, or mixed—when assigning *higher*-than-monologue scores to the paired/group oral participants. In the monologic

picture description (monologue task), there was only one positive comment, while there were 40 negative comments. Correspondingly, toward interlocutors who earned higher scores in pairs/groups, the raters wrote a larger number of positive comments (1 → 19) and fewer negative comments (40 → 21). A larger number of mixed comments (positive plus negative comments for the same student) were observed in the monologue task. Raters were likely to be aware of how they rated the participants on the monologue task and might give positive comments unintentionally when assigning higher scores although they were asked not to compare them.

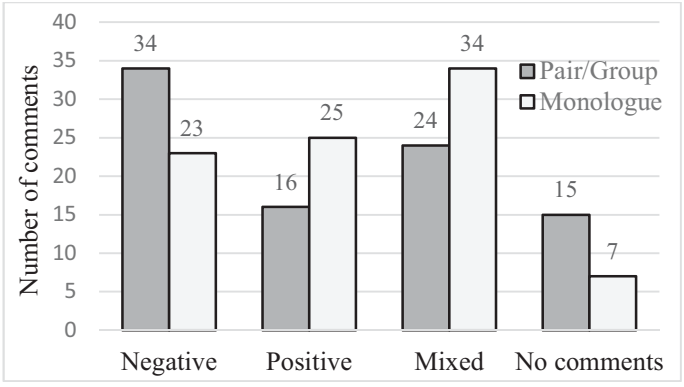

Figure 4. Rater comments when assigning *lower* scores on pair and group work than on the monologue

We can see in Figure 4 that the raters expressed more negative comments about pair/group participants when they assigned *lower* scores in the paired or group oral (21 negative comments in Figure 3 vs. 34 in Figure 4), as would be expected. The disparity in positive comments in pairs/groups between Figures 3 and 4 was not large (19 vs. 16); however, on the monologue task many more speakers received positive comments (1 vs. 25). Furthermore, a larger number of mixed comments, both negative and positive, was observed across groups (18+7 in Figure 3 vs. 34+24 in Figure 4). Despite the tendencies described above, no clear qualitative phenomena were seen. In future research, qualitative analysis of the rater comments should be conducted to further explore the situation

5 Conclusion

Oral performance tests with multiple speakers, that is, paired or group orals, have become more popular irrespective of their drawbacks because collaborative interactions between speakers create negotiation of meaning, which in turn promotes language learning. One important issue involved in implementing these interactional oral performance tests is that raters'

behavior and perceptions have not yet been well investigated across different test formats. This study was carried to explore assessment behavior and perceptions of raters when assigning scores on three different types of test: a single-speaker monologic task (picture description), a paired oral interaction, and a group oral interaction. Twenty-four students whose English speaking abilities varied from lower- to higher-level participated in the study. In the paired and group orals, the participants were paired/grouped once with interlocutors of a similar proficiency level and once with interlocutors of a different level. Five Japanese raters received training in appropriate rating methods for this task structure and assessed the participants.

With regard to RQ1 (How do raters' behaviors when assigning scores on a paired or group oral contrast with behaviors for scores given to a single speaker?), the participants were assigned shared scores at twice the rate of individual scores in the paired oral and more than 1.5 times the rate in the group oral. Double the number of participants received lower scores than received higher scores in the paired oral as compared to the monologue. More specifically, lower-proficiency speakers tended to be assigned lower scores in pairs; in the group oral, the difference between higher- and lower-scoring speakers was not prominent. As the results indicate, even when the criteria required assessment of individual speakers, raters often ended up assigning shared scores. This result indicates that it is not always necessary to use assessment criteria in order to assign shared scores, because raters unintentionally give shared scores in individual assessment situations. If there was a great disparity in speaking ability between/among interlocutors, they might feel it was unfair to receive a shared score.

With respect to RQ2 (How do raters assess speakers in terms of participants' proficiency level and pairing/grouping patterns?), higher-level speakers showed a tendency to obtain higher scores in the single-speaker task than in pairs or groups. This might be because they are used to and comfortable with the monologue test format, or because in pairs/groups they have to accommodate their proficiency to lower-level interlocutors. However, when lower-level speakers interacted with other lower-level speakers, they also tended to get lower scores, which may mean that they were incapable of maintaining an effective L2 communicative interaction by themselves. As mentioned in Section 3.4, many Japanese learners of English are categorized as lower level, and so they might find performing well in a multiple-speaker test format a little difficult. In contrast, the raters were likely to assign middle-level participants higher scores, a finding that was more prominent in the group oral. This result might suggest that such multiple-speaker tasks work well for this level of speakers. Regarding this result, classroom teachers should conduct pair and group activities more in order for all learners of all levels to fully exhibit their interactional competence.

As for RQ3 (Is there any inter-rater variability?), individual raters did demonstrate some different characteristics. For example, three raters tended

to assign lower scores to interlocutors in paired or group orals compared with those in the monologue. One rater assigned more higher scores to interlocutors in multiple-speaker conditions. Another rater's number of higher and lower scores was balanced across conditions. Rater difference is said to be the most variable factor in an assessment, because raters may be more severe or lenient towards a particular candidate, and they may in a general sense interpret the rating scale differently or inconsistently. For these reasons, it is essential to conduct rater training so as to eliminate any conceivable disparity. Nonetheless, studies show that rater differences are still evident even after comprehensive training is conducted (Lumley, 2002; Lumley & McNamara, 1995; McNamara, 1996). For this reason, multiple raters should be used for this kind of assessment to ensure fairness and validity.

Last, with regard to RQ4 (What types of perceptions do raters have toward participants when assigning scores?), the raters showed a greater tendency to write positive comments on the multiple-speaker task than on the single-speaker task when assigning higher scores. Although the raters were asked not to compare the participants' scores, they showed a tendency to write positive comments when rating them better and negative comments when rating them worse. The exact nature and meaning of this superficially unsurprising finding may be fleshed out further by looking at their comments, which will be done in future research.

A limitation of the study is that the number of participants was not large enough. The 24 participants were recruited in consideration of the raters' skill and work burden. In future research, a larger number of participants may give us more generalizable results. Another limitation was that the study was carried out quantitatively. Future research should use a qualitative method to explore the discourse and interaction patterns elicited by each of the three types of test, such as how meaning is negotiated, or how interactional competence is exhibited.

Acknowledgement

This work was supported by the Japan Society for the Promotion of Science (JSPS) KAKENHI, Grant Number 25884073.

References

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice: Designing and developing useful language tests.* Oxford: Oxford University Press.

Berry, V. (2004). *A study of the interaction between individual personality differences and oral performance test facets.* Unpublished doctoral dissertation. King's College, University of London.

Bonk, W. J. & Ockey, G. J. (2003). A many-facet Rasch analysis of the second language group oral discussion task. *Language Testing*, *20*(1), 89–110.

Bonk, W. J., & Van Moere, A. (2004, March). *L2 group oral testing: The influence of shyness/outgoingness, match of interlocutors' proficiency level, and gender on individual scores.* Paper presented at the Language Testing Research Colloquium, Temecula, CA.

Brooks, L. (2009). Interacting in pairs in a test of oral proficiency: Co-constructing a better performance. *Language Testing*, *26*(3), 341–366.

Canale, M. (1983). From communicative competence to communicative language pedagogy. In J. Richards, & J. R. Schmidt (Eds.), *Language and communication* (pp. 1–27). London: Longman.

Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics*, *1*(1), 1–47.

Canagarajah, S. (2006). Changing communicative needs, revised assessment objectives: Testing English as an international language. *Language Assessment Quarterly*, *3*, 229–242.

Council of Europe (2001). *Common European framework of reference for languages: Learning, teaching, assessment.* Cambridge: Cambridge University Press.

Davis, L. (2009). The influence of interlocutor proficiency in a paired oral assessment. *Language Testing, 26*(3), 367–396.

Foot, M. C. (1999). Relaxing in pairs. *ELT Journal*, *53*(1), 36–41.

Fulcher, G. (1996). Testing tasks: Issues in task design and the group oral. *Language Testing 13*(1), 23–51.

Fulcher, G. (2003). *Testing second language speaking.* Harlow: Pearson.

Galaczi, E.D. (2004). *Peer–peer interaction in a paired speaking test: The case of the First Certificate in English.* Unpublished doctoral dissertation, Teachers College, Columbia University.

Galaczi, E.D. (2008). Peer-peer interaction in a speaking test: The case of the First Certificate in English examination. *Language Assessment Quarterly*, *5*(2), 89–119.

Iwashita, N. (1996). The validity of the paired interview format in oral performance assessment. *Melbourne Papers in Language Testing, 5*, 51–66.

Jacoby, S. & Ochs, E. (1995). Co-construction: An introduction. *Research on Language and Social Interaction*, *28*(3), 171–183.

Kramsch, C. (1986). From language proficiency to interactional competence. *The Modern Language Journal*, *70*(4), 366–372.

Lumley, T. (2002). Assessment criteria in a large-scale writing test: What do they really mean to the raters? *Language Testing, 19*(3), 246–276.

Lumley, T., & McNamara, T. F. (1995). Rater characteristics and rater bias: Implications for training. *Language Testing 12*(1), 54–71.

May, L. (2009). Co-constructed interaction in a paired speaking test: The rater's perspective. *Language Testing*, *26*(3), 397–421.

McNamara, T. F. (1996). *Measuring second language performance*. London: Longman.

Ministry of Education, Culture, Sports, Science and Technology (Japan). (2013). *Kaku chu-, koto-gakko no gaikokugo kyoiku ni okeru "CAN-DO list" no katachi deno gakushu totatsudo mokuhyo settei no tame no tebiki* [*A guidebook for setting learning goals using a "CAN-DO list" in foreign language teaching at junior high schools and high schools*]. Tokyo: Taishukan.

Negishi, J. (2011). *Characteristics of group oral interactions performed by Japanese learners of English.* Unpublished doctoral dissertation. Waseda University.

Negishi, J. (2015). Effects of test types and interlocutors' proficiency on oral performance assessment. *ARELE 26*, 333–348.

Negishi, M., Takada, T., & Tono, Y. (2012). A progress report on the development of the CEFR-J. *Studies in Language Testing*, *36*, 137–157.

North, B., & Hughes, G. (2003). *CEF illustrative performance samples: For relating language examinations to the Common European framework of reference for languages: Learning, teaching, assessment (CEF).* Zürich: Eurocentres and Migros Club Schools.

O'Sullivan, B. (2002). Learner acquaintanceship and oral proficiency test pair-task performance. *Language Testing*, *19*(3), 277–295.

Shohamy, E., Reves, T. and Bejarano, Y. (1986). Introducing a new comprehensive test of oral proficiency. *English Language Teaching Journal 40*(3), 212–220.

Storch, N. (2002). Patterns of interaction in ESL pair work. *Language Learning*, *52*(1), 119–158.

Swain, M. (2001). Examining dialogue: Another approach to content specification and to validating inferences drawn from test scores. *Language Testing*, *18*(3), 275–302.

Tono, Y. (2013). *CAN-DO list sakusei katsuyo, eigo totatsu-do shihyo [A CEFR-J guidebook. CAN-DO list development and utilization; Reference for English achievement*]. Tokyo: Taishukan.

Van Moere, A. (2006). Validity evidence in a university group oral test. *Language Testing*, *23*(4), 411–440.

Weir, C. J. (2005). Limitations of the Common European Framework for developing comparable examinations and tests. *Language Testing*, *22*(3), 281–300.

Young, R. (2000, March). *Interactional competence: Challenges for validity*. Paper presented at the Annual Meeting of the American Association for Applied Linguistics, Vancouver.

Junko Negishi
Tsurumi University
Bld. 6, Tsurumi University
2-1-5, Tsurumi, Tsurumi-ku
Yokohama, 230-0063 JAPAN
Phone & Fax: +81-29-252-4866
Cell phone: +80-90-9688-7310
Email: negishi-j@tsurumi-u.ac.jp

Received: April 1, 2015
Revised: June 1, 2015
Accepted: July 1, 2015