

Interaction in a paired oral assessment: Revisiting the effect of proficiency

Young A Son
Georgetown University

Paired oral assessments have gained increasing popularity as a method of assessing speaking skills (East, 2015; Galaczi, 2014). Several advantages have been associated with this method, including practicality and authenticity (Taylor, 2003). Nevertheless, concerns have also been raised in terms of the interlocutor effect in paired speaking tests, particularly in regard to the interlocutor's oral proficiency (e.g., Norton, 2005).

The present study reports on an approximate replication of Davis (2009), who looked at the effect of interlocutor proficiency on paired speaking assessments. The current study compared the oral performance of 24 university students in two different pairing conditions: once paired with a partner of the same proficiency level and once with a partner of a different proficiency level. Results of this replication study partially confirmed Davis's (2009) results. There were only minimal differences in test-takers' scores between both conditions. A multi-facet Rasch analysis confirmed these results indicating that the pairing conditions were equivalent in difficulty. There were, however, observable differences in the quantity of talk depending on the interlocutor's proficiency. Unlike Davis (2009), this study found that low-proficiency test-takers produced fewer words when paired with high-proficiency partners. Even though the number of words produced by test takers was affected by their partner's proficiency, their performance scores remained constant.

Keywords: paired speaking assessment, interlocutor proficiency effect, approximate replication

Introduction

Pair or group oral assessments have gained popularity as an alternative tool to assess learners' speaking ability as they elicit rich and authentic speech samples (French, 1999; Eyüd & Glover, 2001; Taylor, 2003; Brooks, 2009), closely resemble communicative classroom activities (Swain, 2001; Philp, Adams, & Iwashita, 2014), and are preferred by teachers (e.g., East, 2015) and students over traditional interviews (e.g., Fulcher, 1996; May, 2000). However, the validity of these types of oral assessments has been questioned, primarily due to the possible effects that test-takers' individual characteristics, namely, personality, acquaintance, gender, and language proficiency, can have on their partner's speaking performance (Foot, 1999; O'Sullivan, 2002; Fulcher, 2003; Norton, 2005).

Most relevant to the present study is the effect of an interlocutor's proficiency on the test-taker's performance. Previous studies on this topic have mainly explored whether pairing test-takers with partners of different proficiency levels could affect the quality of their speaking performances and thus threaten the reliability of the test. These studies have yielded contradictory findings. While some argued that certain test-takers benefited from being paired with different proficiency partners (Norton, 2005), others found that there were no statistically significant differences in the scores, suggesting that scores were not affected by the interlocutor's language proficiency (Davis, 2009). In regard to the difference in amount of talk depending on the pairing conditions, studies have indicated a trend where both high- and low-proficiency test-takers tended to talk more when being paired with high-proficiency test-takers (Davis, 2009; Iwashita, 1998). Another topic of debate has been the influence of interlocutor proficiency on the pattern of interactions, including collaborative, parallel, or asymmetric (Galaczi, 2008) observed among dyads. Davis (2009) found that the pattern of interactions differed according to the test-takers' proficiency and pairing. Meanwhile, Nakatsuhara (2004, 2006) found only slight differences in the styles of conversation due to pairing conditions.

Following these studies, the present investigation conducted an approximate replication of Davis (2009) to explore the extent to which the scores and amount of talk in paired speaking assessments differ according to the pairing conditions, and whether the patterns observed in the quantitative analyses of the scores were also noticeable in the quantity of talk. As described by Porte (2012), an approximate replication is a type of study that closely follows an original study in most respects but includes some changes in non-critical variables in the research design. The main purpose of this type of studies is to examine the generalizability of the original study results to other subjects or contexts.

Accordingly, the present study sought to contribute to the previous literature on the validity of this format of speaking assessment and explore it in the context of Korean EFL learners.

Literature Review

Advantages of the Paired Speaking Assessment

Several benefits have been associated with the use of paired oral tests. Taylor (2001) indicated that the main reason for adopting this format of oral assessment was its advantage over the traditional face-to-face interviews. Recently, East (2015), described a case of assessment reform in New Zealand motivated by the need to move away from “one-sided engagements in somewhat ‘staged’ conversations” (p. 102). Numerous studies (e.g., Brown, 2003; Lazaraton, 1996; Nakatsuhara, 2008; Norton, 2013) have examined the interviewer effect on test-takers in traditional interviews. They have suggested that the interviewer’s behavior or strategies can have an effect on the way interviews are carried out, particularly in terms of rapport, topic sequencing, and backchannelling (e.g., Brown, 2003; Ross, 2007) and that a supportive type of interviewer behavior does not always benefit test-takers (Norton, 2013).

One of the major advantages of this speaking assessment format is its close resemblance to classrooms activities, which are usually carried out with peers in pairs or groups, and thus promote learning as test-takers are able to negotiate meaning through the interaction with other peers (Kasper, 2013; Philp et al., 2014; Swain, 2001). Furthermore, this setting allows for a measurement of interactional competence or interactive communication (Galaczi, 2008, 2014) which can be defined as the ability of test-takers to co-construct conversations in a meaningful way. In addition, there has been evidence demonstrating that peer interaction elicits a wider range of language functions than traditional interviews (e.g., Taylor, 2003). In this regard, studies like French (1999) and Brooks (2009) compared data from the traditional oral proficiency interviews and the paired interview and found that there was a considerably wider variety of language functions, such as informational and interactional functions (e.g., prompting elaboration, referring to partner’s ideas, paraphrasing), elicited in the latter format.

On the matter of test-takers’ preferences, several studies (Együd & Glover, 2001; Fulcher, 1996; May, 2000) used questionnaires to explore the test-takers’ insights on both formats of speaking assessment and found that that test-takers preferred paired assessments to interviews because they felt more relaxed and they considered this format authentic,

effective, and enjoyable. Likewise, language educators have generally expressed positive attitudes towards this format of speaking assessment over the traditional interviews (East, 2015).

Interlocutor Effect in Paired Oral Assessment

Notwithstanding its numerous benefits, paired speaking assessments have also been subject to close scrutiny as concerns about test fairness have been raised in terms of the effect the interlocutor might have on test-takers' performance. Foot (1999) was one of the earliest researchers to point out the possible negative impact that the test-takers' characteristics (i.e., L1 background, proficiency, and personality) may have upon their partner's speaking performance.

In terms of familiarity, O'Sullivan (2002) and Norton (2005) indicated that knowing the interlocutor might be beneficial for the interaction. In these studies, when test-takers were paired with a friend, they performed better than when paired with a stranger. O'Sullivan also pointed out that other factors such as gender and cultural background could play a role in the effect of acquaintance, making this variable more complex. These studies, however, have not been conclusive. Lazaraton (2006), for example, criticized Norton's (2005) study for its lack of generalizability due to a small sample size. Furthermore, Fulcher (2003) reanalyzed O'Sullivan's (2002) data and found that although familiar and unfamiliar group scores were significantly different, the effect sizes were medium with only 24% of cases showing this trend.

Regarding the interlocutor's personality effect, Vivien Berry conducted several studies (see Berry, 2007) that focused on the effect of the interlocutor's introversion and extroversion on test-takers' performance. Her findings suggested that test-takers performed better when they interacted with a partner with the same type of personality. Other studies have focused on personality in group oral tests and have investigated the benefits or disadvantages of certain personality types. Bonk and Van Moere (2004), for example, found that while shyness negatively influenced test-takers' scores, outgoingness benefited them. In addition, Ockey (2009) indicated that assertive test-takers scored significantly higher than non-assertive ones. Nakatsuhara (2011) examined the interaction between variables, such as personality, proficiency, and number of participants in a group. The results showed that extraversion levels had a greater effect on test-takers' performance in the groups composed of three participants. She also found that proficiency level affected both groups, but that there were larger effect sizes in groups of three than groups of four.

Interlocutor's Proficiency Effect in Paired Oral Assessment

The effect of the interlocutor's language proficiency has also attracted a great deal of attention in research on paired speaking tests. Norton (2005) analyzed real test data collected from the Cambridge ESOL FCE and CAE. She observed that low-proficiency participants tended to score higher when being paired with high-proficiency partners. When analyzing the speech samples, she also found that low-proficiency test-takers benefited from being paired with high-proficiency partners as they would appropriate grammatical and lexical items from their partners. East (2015) examined teachers' perceptions on the implementation and use of similar paired speaking assessments and found that there is some concern among language educators that the language ability of a partner could affect test-takers' performance in the tests.

On the other hand, several studies have specifically designed procedures to collect data in more controlled settings. Csépes (2009), for example, analyzed the differences in the scores of 30 Hungarian secondary school students when being paired with students from three different language proficiency groups. Each participant was tested three times: once with a partner of higher proficiency, another time with a partner of the same proficiency and a third time with a partner of lower proficiency. The results indicated that participants were not affected negatively or positively by the interlocutor's language proficiency.

Similarly, an early study by Iwashita (1998) examined the performance of 20 Japanese learners in two separate paired tests, once with a partner of the same proficiency and once with a partner of different proficiency. The analysis of the raw scores indicated that both high- and low-proficiency test-takers benefited slightly from being paired with high-proficiency partners. In addition, the amount of talk was greater in pairings with high-proficiency partners.

Following a similar methodology, Davis (2009) analyzed the performance scores of 20 Chinese EFL learners. The study sought to determine whether the interlocutor's proficiency had an effect on average rating scores and amount of talk, whether the amount of talk was associated with average scores, and whether there was a relationship between the interlocutor's proficiency and the type of interaction produced. Certain aspects of the research methodology in Iwashita (1998) were improved. For example, the test-takers' proficiency was determined by several factors, such as institutional status (i.e. Software Design vs. English major), self-reported standardized tests, as well as performance scores on a monologic task. Furthermore, the performance scores of the paired speaking assessment were analyzed through a multi-facet Rasch analysis. The

analysis of scores indicated that the interlocutor's proficiency had no effects on the test-takers' speaking scores. Both pairing conditions (when paired with a same proficiency partner or with a different proficiency partner) were associated with the same difficulty level. In other words, the pairing conditions did not make the test more difficult or easy. As for the amount of talk, the results echoed Iwashita's (1998) findings, showing that all test-takers tended to talk more when paired with high-proficiency partners. In particular, low-proficiency test-takers were found to talk significantly more when paired with high-proficiency test-takers.

In sum, the literature on this domain has investigated how the interlocutor's proficiency might (or might not) affect test-takers' performance in paired speaking assessments. Nevertheless, the considerable methodological variability across studies has challenged the comparability of results and the interpretation of broader implications on the impact of interlocutor proficiency on speaking performances. For example, there has been considerable variability in the criteria for grouping participants into different proficiency groups (i.e., higher or lower proficiency). While some studies took into consideration test-takers' performance in a monologic task as part of the criteria (e.g., Davis, 2009), others considered their performance in an oral interview (e.g., Iwashita, 1998) or ratings in a speaking self-assessment questionnaire (Csépes, 2009). In other words, what for some studies was considered higher- or lower-level participants might have had a different meaning in other studies. In addition, task types used for the paired assessments varied across studies. Davis (2009), on the one hand, used decision-making tasks whereas Iwashita (1998) utilized information gap tasks. Task types might have been a moderating factor that affected the way interlocutor proficiency influenced test-takers' speaking performances. All in all, it is unclear how results of previous studies are related to one another. For this reason, more replication studies are necessary in order to shed light on the validity of the interpretation of results and examine their generalizability in different contexts and learner populations. The present study, then, sought to conduct an approximate replication to Davis (2009) in order to further explore the effect of interlocutor proficiency in a paired speaking assessment by verifying the results of the original study on a new population and setting, while also improving on some of its methodological limitations. The study was guided by the following research questions:

To what extent does the interlocutor's proficiency level have an effect on the test-taker's overall and analytic scores of their speaking performance?

To what extent does being paired with a different- or a same-proficiency partner affect the quantitative dominance in the interactions between test-takers?

To what extent is there a relationship between the amount of words produced and the overall performance scores?

Methodology

Participants

Thirty-five Korean EFL university students from various majors were recruited to participate in this study. Initially they were separated into high- and low-proficiency groups according to their self-reported scores on the Test of English developed in Seoul National University (TEPS, henceforth) taken within 2 years prior to start of the study. Participants with a score 850 or above were categorized as high-proficiency, whereas those with scores ranging from 650 to 750 were categorized as low-proficiency². Due to the absence of a speaking component in TEPS, a non-interactive test (NI-test, see next section for details), similar to the one used in the original study by Davis (2009), was administered before the paired tests in order to more accurately classify participants into proficiency groups based on speaking abilities. Five participants were reassigned to different proficiency groups according to the scores of the non-interactive test.

Furthermore, after the non-interactive test three participants dropped out of the study, reducing the initial pool to 32. The remaining participants were categorized into three proficiency groups: high, intermediate, and low, according to their NI-test and TEPS scores. As the study aimed at examining whether there was an interlocutor proficiency effect, only the data of the high- and low-proficiency test-takers were analyzed, resulting in a final pool of 24 participants (a slight increase in sample size than the original study) of which half were female and half were male. Table 1 presents data from the scores of TEPS and the NI-Test. An independent samples t-test was conducted to compare test-takers' performances on the non-interactive test. The results showed a statistically significant difference between low- and high-proficiency test-takers ($t(22) = 13.51, p = .000, d = 5.05$), suggesting that participants from both groups had significantly different proficiency levels as measured by the non-interactive test.

² According to TEPS (2009), the scores 850 and 750 are equivalent to a TOEFL scores of 111 and 103, respectively.

Table 1. Scores on TEPS and Non-Interactive Test ($N = 24$)

| Proficiency | TEPS | | | | Non-interactive Test | | | |
|-------------------|--------|-------|-----|-----|----------------------|------|-----|-----|
| | Mean | SD | Min | Max | Mean | SD | Min | Max |
| High ($N = 12$) | 890.92 | 63.93 | 722 | 984 | 13.00 | 1.04 | 12 | 15 |
| Low ($N = 12$) | 728.08 | 83.97 | 606 | 903 | 8.08 | 0.90 | 6.5 | 10 |

Note. The maximum possible score in the Non-interactive Test was 15

Instruments

Three speaking tests were developed for the purposes of this study. For comparability with the original study, the tasks used in the current study were adapted from retired Cambridge ESOL FCE speaking tests. The first test was a non-interactive test (NI-test, henceforth), which consisted of two picture description tasks and one short open-ended question about those pictures. Each task was scored on a five-point holistic scale which represented an overall assessment of grammar, vocabulary, pronunciation, fluency, and coherence. In addition, two separate paired speaking tests were administered for pairing test-takers once with a partner of same proficiency and once with a partner of different proficiency. In these tests, test-takers were asked to talk to each other about seven pictures related to one topic (e.g., Coffee shop; see Appendix A) and respond to a question related to those pictures for 5 minutes. In relation to the rubric, Davis (2009) used a five-point scale measuring four criteria, namely grammar and vocabulary, pronunciation, fluency, and discourse management. In the present study, however, grammar and vocabulary were assessed using separate criteria (See Appendix B for the rubric). The scoring rubric was adapted from the Cambridge FCE rating scale (Cambridge ESOL, n.d.) and the scale developed by Nakatsuhara (2007) for assessing English speaking in group oral activities. Both of these rubrics specified descriptors for the discourse management criterion, which made it appropriate for the present study. Interaction studies have indicated that learners engage in and resolve lexical language-related episodes to differing degrees depending on who they are interacting with (e.g., Fernández-Dobao, 2012, 2014). Therefore, it was important to keep vocabulary as a separate criterion in order to see whether use of lexical items could be affected by the interlocutor's proficiency.

Pearson correlation coefficients for all tests, namely the NI-test, paired tests, and TEPS were conducted to examine the relationship among the test items, test conditions, as well as concurrent validity with high-stakes exams (i.e., in this case TEPS). As observed in Table 2, the correlation coefficients were high, especially within the test types. In other words, all three items in the NI-test and all prompts and pairing conditions in the paired tests showed positively strong correlation coefficients. In addition, it should be noted that the correlations between the oral tests (NI-test and paired tests) and TEPS were lower

than other correlation coefficients, ranging from 0.76 to 0.81. Several studies (Sawaki, Stricker, & Oranje, 2008; Powers, 2010) have presented evidence to suggest that different components such as reading, listening, speaking, and writing, though strongly correlated, measure different language skills. As TEPS measures receptive skills (i.e., reading and listening) rather than productive skills (i.e., speaking and writing), participants who scored high on TEPS did not always score high on the NI-test and paired tests. Nevertheless, the coefficients were still significant, $p < 0.01$.

Table 2. Correlation coefficients between tests

| Test Scores | NI-Test | Paired Test | | | | TEPS |
|-------------|---------|-------------|------|------|------|------|
| | | Café | Film | PSP | PDP | |
| NI-Test | 1 | | | | | |
| Café | .85 | 1 | | | | |
| Film | .87 | .94 | 1 | | | |
| PSP | .88 | .97 | 0.97 | 1 | | |
| PDP | .86 | .97 | 0.97 | 0.94 | 1 | |
| TEPS | .76 | .76 | 0.81 | 0.78 | 0.77 | 1 |

Notes. NI-Test = Non-interactive test; Café and Film = Prompts; PSP= Paired with same proficiency; PDP = Paired with different proficiency

Data Collection Procedure

The data collection procedure followed a similar scheme to the original study. However, for more accurate pairing and clear-cut difference between low- and high-proficiency test-takers, data collection was carried out in two phases. In Davis's study (2009), test-takers were initially paired according to their academic majors but relabeled under a different proficiency group during data analysis. This resulted in two dyads being eliminated from the study because they were paired with the same type of partner in both of the paired tasks. Thus, in the present study, during the first phase, students were assessed individually through the NI-test. According to the scores in the NI-test, they were classified into three groups: high-, intermediate- and low-proficiency. As mentioned earlier, the intermediate group was omitted from the study. In the second phase, participants were divided into groups of four and called back to perform the paired speaking tests. In each group, two test-takers were from high- and two were from low-proficiency levels. Each pair was tested in two separate classrooms simultaneously. When both pairs were finished taking the first paired test, one student from each dyad exchanged places and took the second paired test. Therefore, all participants were tested twice; once with a same-proficiency partner and once with a different-proficiency partner. The ordering of the task and pairing conditions was counterbalanced. In order to control possible gender effects, there were no gender-mixed groups.

Subsequently, all audio recordings were sent to two raters. The audio files were arranged in randomized order so that speech samples from the same test-taker would not appear consecutively. The raters were native English speakers and had extensive experience in the field of EFL assessment while working at TEPS. They were trained to use the specific scoring rubrics created for this study. Raters were instructed to provide an analytic rating based on the scoring scale provided in Appendix B. These ratings were then averaged by the researcher to obtain the overall scores for each test-taker. Raters' performance was examined through agreement rates and Rasch measurements. In terms of rater agreement, the two raters had an exact agreement rate of 49% in paired oral tests. However, this rate increased to 97% when considering adjacent agreement rates (i.e., scores within one point of difference). In addition, Rasch measurements (Table 3) suggested that raters had similar levels of severity with a span of 0.34 logits between the most severe and the most lenient rater. The reliability of separation index was low (0.41), indicating that the raters were relatively equally severe (i.e., the Rasch model could not distinguish between them very reliably in terms of their severity). Moreover, they were internally consistent as observed through the measures of model fit, with infit and outfit mean square values within the acceptable ranges of 0.7 to 1.3 (McNamara, 1996).

Table 3. Rasch measurement for rater performance

| Raters | Observed | Severity | | Model Fit | |
|---------|----------|----------|------|-----------|--------|
| | Average | Measure | S.E. | Infit | Outfit |
| Rater 1 | 3.78 | -0.17 | 0.13 | 1.02 | 1.05 |
| Rater 2 | 3.69 | 0.17 | 0.13 | 0.92 | 0.92 |

Data Analysis

As in the original study, a multi-faceted Rasch analysis using FACETS (Linacre, 2013) was conducted to examine four facets: examinee, rater, pairing condition, and analytic scoring criteria. The Rating Scale Model (RSM) was used to analyze the data. This model assumes that all items share the same rating scale (Wright, 2000). Due to the small sample size, it was not appropriate to conduct a Partial Credit Model (PCM). The Rasch measurements helped to explore the interaction between all facets (i.e., examinee ability estimates, rater severity, and difficulty of the condition) and the possible misfitting elements of both paired tests.

In order to address the first research question, which asked whether the pairing conditions had an effect on the overall and analytic scores of the test, descriptive statistics were examined for differences in scores. To answer the second and third research questions, which asked whether the quantitative dominance between test-takers differed

according to the pairing conditions and whether this affected the scores, all conversations were transcribed verbatim following Conversational Analysis (CA) conventions. Although the amount of detail that CA provides was not necessary for the present study, this specific convention was chosen because it provided a more comprehensive description of the interactions (e.g., false starts, repetition, pauses, hesitations), ensuring the accuracy of the word count. After transcribing the conversations, the words were counted to examine the difference in amount of talk depending on the pairing conditions. This is in line with the original study which indicated that the number of words is a robust measurement of the amount of talk by each test-taker in the conversation. In this replication study, however, the proportion of words was also taken into consideration for exploring test-takers' contribution to their interactions in relation to their partners. A balanced conversation was considered to be one that contained a similar percentage of words contributed by both test-takers.

Results

RQ 1 - Interlocutor Proficiency Effect on Composite and Analytic Scores

The results of the composite scores confirmed findings in Davis's study (2009), indicating no interlocutor proficiency effect on the overall oral performance of both high- and low-proficiency test-takers. The average raw scores suggested that low-proficiency test-takers showed a slightly higher average raw score when being paired with high-proficiency partners, whereas high-proficiency test-takers performed consistently in both pairing conditions.

Table 4. Composite scores for low- and high-proficiency test-takers

| | Low-proficiency (<i>N</i> = 12) | | | High-proficiency (<i>N</i> = 12) | | |
|-----|----------------------------------|-----------|------|-----------------------------------|-----------|------|
| | Mean | <i>SD</i> | SEM | Mean | <i>SD</i> | SEM |
| PSP | 14.88 | 2.65 | 0.76 | 22.54 | 1.89 | 0.55 |
| PDP | 15.17 | 2.51 | 0.72 | 22.54 | 2.27 | 0.66 |

Note. PSP = Paired with Same Proficiency; PDP = Paired with Different Proficiency; Scale of 0-25.

Similarly, Rasch measurements (Table 5) indicated that the level of difficulty of the pairing conditions ranged from 0.07 to -0.07 logits (i.e., a difficulty span of 0.14). These results reflect that the pairing conditions were of equivalent difficulty. Thus, the interlocutor proficiency had little effect on both high- and low-proficiency test-takers scoring higher or lower on the paired tests.

Table 5. Difficulty measurement for pairing conditions based on Rasch model analysis

| Pairing Condition | Observed Average | Difficulty | | Model Fit | |
|-------------------|------------------|------------|------|--------------|--------------|
| | | Measure | S.E. | Infit (MnSQ) | Oufit (MnSQ) |
| PSP | 3.72 | 0.07 | 0.13 | 0.95 | 1.01 |
| PDP | 3.75 | -0.07 | 0.13 | 1.00 | 0.95 |

Note. PSP = Paired with Same Proficiency, PDP= Paired with Different Proficiency

The analysis of the average subscale scores also indicated that there was no interlocutor proficiency effect on test-taker performance in any analytic scoring category, although the raw scores suggested a slight difference in test-taker performance in some of the analytic subscales (Table 6). For example, the criterion that showed the most difference across pairing conditions for high- and low-proficiency test-takers was pronunciation, followed by vocabulary for the low-proficiency, and grammar for the high-proficiency test-takers. Subscale scores differed more within groups, indicating that criteria such as grammar and vocabulary are perceived differently from each other and thus should be scored separately.

Table 6. Analytic scores for low- and high-proficiency test-takers

| | Low-proficiency | | | | | | High-proficiency | | | | | |
|----------------------|-----------------|------|------|------|------|------|------------------|------|------|------|------|------|
| | PSP | | | PDP | | | PSP | | | PDP | | |
| | Mean | SD | SEM | Mean | SD | SEM | Mean | SD | SEM | Mean | SD | SEM |
| Grammar | 2.75 | 0.66 | 0.19 | 2.71 | 0.45 | 0.13 | 4.38 | 0.53 | 0.15 | 4.50 | 0.56 | 0.16 |
| Vocabulary | 3.04 | 0.45 | 0.13 | 3.13 | 0.53 | 0.15 | 4.54 | 0.45 | 0.13 | 4.54 | 0.58 | 0.17 |
| Pronunciation | 2.92 | 0.85 | 0.24 | 3.17 | 0.86 | 0.25 | 4.42 | 0.29 | 0.08 | 4.25 | 0.66 | 0.19 |
| Fluency | 2.79 | 0.54 | 0.16 | 2.79 | 0.33 | 0.10 | 4.58 | 0.56 | 0.16 | 4.54 | 0.45 | 0.13 |
| Discourse Management | 3.38 | 0.61 | 0.18 | 3.38 | 0.57 | 0.16 | 4.63 | 0.43 | 0.13 | 4.71 | 0.45 | 0.13 |

Note. PSP = Paired with Same Proficiency; PDP = Paired with Different Proficiency

Rasch model analysis of the five scoring criteria showed that not all of the categories were of equivalent difficulty, with a small separation value of 2.69 ($\chi = 37.5$, $df = 4$, $p < 0.00$). These findings indicate that the difficulty levels among the criteria could be distinguished to some degree by the Rasch measurement approach. The most leniently scored criterion was discourse management and the most harshly scored was grammar. The difficulty span between these two subscales was 1.68 logits (Table 7). In addition, the mean square values for infit and outfit demonstrate that there were no misfitting criteria. In other

words, the scores of each category were internally consistent. Overall, when considering the observed average scores, raters appear to have scored leniently in each analytic category. They did not use the full range of scores in the five-point scoring scale, which is a known effect when using ordinal scales (Knoch, 2009). Most of the test-takers had scores higher than 3 on all scoring criteria and no participant was given a score of 1 (see Appendix C for vertical map). This indicates either that the scale needs to be revised or that the pool of participants was of a rather similar overall proficiency level.

Table 7. Difficulty measurements for the five scoring criteria

| Scoring Criteria | Observed Average | Difficulty | | Model Fit | |
|----------------------|------------------|------------|------|--------------|--------------|
| | | Measure | S.E. | Infit (MnSQ) | Oufit (MnSQ) |
| Grammar | 3.56 | 0.65 | 0.19 | 0.90 | 0.87 |
| Fluency | 3.64 | 0.39 | 0.20 | 0.93 | 0.90 |
| Pronunciation | 3.69 | 0.19 | 0.20 | 1.07 | 1.04 |
| Vocabulary | 3.79 | -0.20 | 0.20 | 0.86 | 1.09 |
| Discourse Management | 4.00 | -1.03 | 0.23 | 1.13 | 1.02 |

RQ2 - Interlocutor Proficiency Effect on Quantitative Dominance

Quantitative dominance was measured through the raw amount of words produced by test-takers as observed in the original study. However, to explore the test-takers' contribution of talk in the interaction in relation to their partners, it was deemed important to also explore the proportion of the number of words uttered. These measurements provided a general overview of the extent to which test-takers held the floor and the amount of words they contributed to the conversation (Itakura, 2001; O'Sullivan & Nakatsuhara, 2011). As in Davis (2009), the present study also found that the amount of talk was affected by interlocutor proficiency. The total amount of words uttered in interactions between high-proficiency test-takers was almost double the amount of words in interactions between low-proficiency test-takers. Moreover, the mixed proficiency pairs (i.e., high-low) tended to produce a higher total amount of words than low-proficiency pairs. Table 8 presents the average words uttered by each pair categorized by high-high, low-low, and mixed proficiency pairs together with the amount of words uttered by the pair with the least and most amount of words.

Table 8. Average amount of words uttered by different types of dyads

| Pair Type | Mean | SD | Min | Max |
|-----------------------|--------|-------|-----|-----|
| High-High ($N = 6$) | 608.17 | 84.88 | 516 | 740 |
| Low-Low ($N = 6$) | 366 | 32.34 | 339 | 428 |
| High-Low ($N = 12$) | 472.5 | 59.75 | 399 | 620 |

Note. N represents the number of observations (i.e., number of dyads).

In regard to the difference in the amount of words produced by test-takers, only minimal differences were found between both conditions (Table 9). High-proficiency test-takers tended to talk slightly more, while low-proficiency test-takers talked slightly less in the mixed pairing condition when compared to then same-proficiency pairing condition (i.e., high-high and low-low pair type). Unlike Davis (2009), the present study found that these differences were minimal in both conditions.

Table 9. Difference in the amount of words by different conditions

| Proficiency Level | Pairing condition | | | | Difference | |
|-------------------|-------------------|-------|-----------|-------|------------|----|
| | Same | | Different | | Words | % |
| | Mean | SD | Mean | SD | | |
| High | 304.08 | 47.82 | 315.33 | 73.74 | -11.25 | -2 |
| Low | 183 | 47.86 | 157.17 | 32.57 | 25.83 | 8 |

In terms of quantitative dominance, the current study considered the analysis of the proportion of the words contributed by each test-taker under the two different conditions to be particularly important. Analysis of the percentage of words contributed indicated that the dyads composed of two high-proficiency test-takers were more balanced than those composed of two low-proficiency test-takers or by high- and low-proficiency test-takers (Table 10). In other words, when high-proficiency learners were paired with partners of the same proficiency, they contributed a similar amount of talk to the conversation. The proportion of the contribution was measured in terms of percentages of amount of talk and the difference between those percentages was computed in order to get a sense of how balanced or imbalanced the conversations were. For example, if test-taker A contributed 282 words and test-taker B 263, then the proportion was calculated as 51.74% and 48.26% respectively. In this particular case, the percentage point difference was 3.48%. Thus, the dyads with a lower percentage point difference were considered more balanced than the ones with a higher percentage point difference.

As observed in the Table 10, the high-high pair type presented a lower percentage point difference than the low-low or high-low pair types. This indicates that the high-high pair

type showed a more balanced proportion of talk when compared to the low-low and high-low pair types. Furthermore, the high-high pair types showed a lower degree of variability between pairs. All six pairs were comparable in terms of balanced talk. On the other hand, the low-low pairs showed more variability with percentage point differences ranging from a pair with only 1.40 to a pair with 41.10. Similarly, the high-low pairs appeared to be unpredictable with percentage point differences ranging from 15.32 to 69.68. In all 12 mixed-proficiency dyads the high-proficiency test-takers dominated the conversation, contributing more in terms of words. However, as observed through the range, the degree of dominance was different for each pair. In some pairs, the high-proficiency test-taker talked much more, accounting for up to 84.84% of the number of words in an interaction with a low-proficiency partner whereas in other pairs their contributions were more balanced with 57.66% of the words.

Table 10. Difference in proportion of talk by different types of dyads

| Pair Type | Mean | Min | Max | Range | SD |
|-----------------------|-------|-------|-------|-------|-------|
| High-High ($N = 6$) | 6.53 | 2.16 | 11.89 | 9.74 | 4.15 |
| Low-Low ($N = 6$) | 17.97 | 1.40 | 41.10 | 39.70 | 15.75 |
| High-Low ($N = 12$) | 32.26 | 15.32 | 69.68 | 54.36 | 15.68 |

RQ3. Relationship between amount of words produced and performance scores

Following the original study, a correlation analysis was conducted between the raw number of words and overall performance scores to examine the relationship between quantitative dominance and scores. Davis (2009), however, reported a single correlation coefficient which did not account for the differences in pairing conditions. The present study investigated the relationship separately for each pairing condition to account for the possible effect that pairing conditions might have on the number of words produced.

For the same proficiency pairing condition, a moderate positive correlation coefficient of $r = 0.66$ was found between the number of words and overall performance scores. Likewise, a moderate positive correlation coefficient was found in the different proficiency pairing condition with $r = 0.68$. The similarity in the correlation coefficients indicated that regardless of the pairing conditions, the relationship between the number of words and scores was similar. This is evident in the overlap in word count between pairing conditions observed in Figure 1. Thus, the data in this study yielded no evidence that the amount of words is more important in one condition than the other.

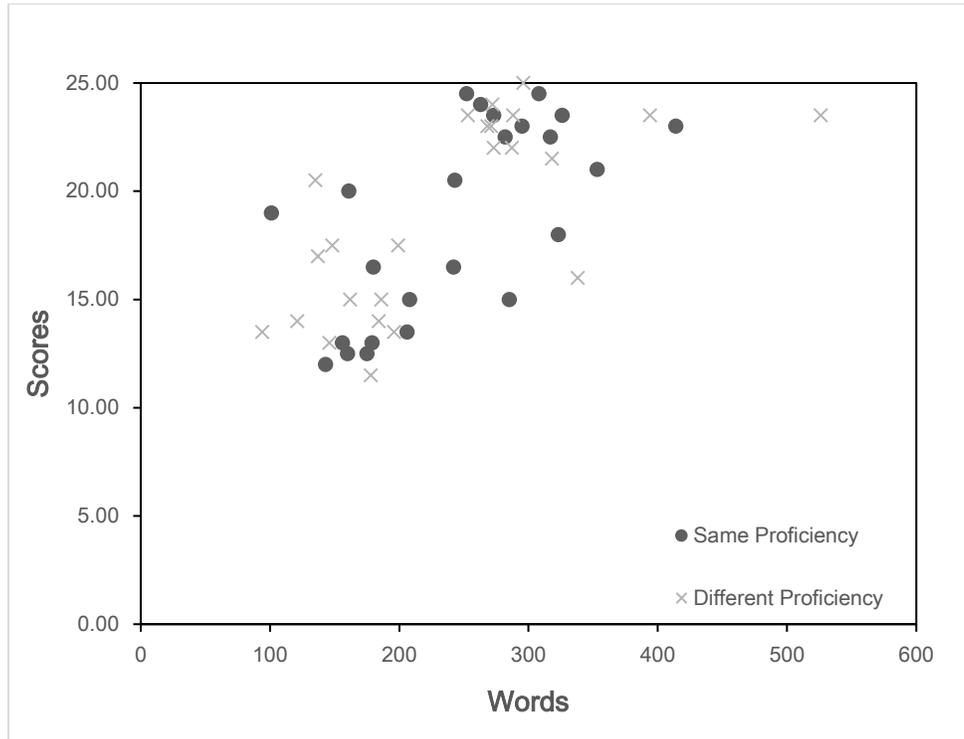


Figure 1. Scatter plot of the number of words by overall scores

In addition, when examining the differences in scores and proportion of words between the two conditions, there was no apparent pattern. Talking more did not yield higher scores or vice versa. While there were some test-takers whose scores and amount of talk changed depending on who they were paired with, the difference in scores varied positively or negatively without any evident pattern.

Discussion

The present study found that the interlocutor proficiency had little effect on test-taker composite and analytic scores in a paired speaking task. The analysis of the raw scores indicated that there was only a minimal increase in mean composite scores of low-proficiency test-takers when paired with high-proficiency test-takers. These results partially agree with Iwashita's (1998) findings, which demonstrated that both high- and low-proficiency test-takers performed better when paired with high-proficiency test-takers. In this study only the low-proficiency group had a minimal increase in mean scores while the high-proficiency group showed no differences. Nevertheless, the increase in scores was very minimal compared to the maximum possible score of 25. This is in line with the study by Davis.

Similarly, the analytic criterion scores showed small differences across pairing conditions. As in Davis (2009), overall, the mean rating scores suggested that interlocutor proficiency had no effects on the test-taker scores. The current study also found individual variation in scores. Thus, while some test-takers scored higher when being paired with a different proficiency partner, others scored lower in the same pairing condition. There was, however, considerably less variation, ranging from -35% to 11% when compared to Davis's (2009) data which reported a range of -50% to +21%.

In addition, results of the Rasch analysis suggested that out of the five scoring categories, vocabulary and discourse management were the most leniently scored. Interestingly, in a post-rating questionnaire, raters pointed out that vocabulary and discourse management were the hardest criteria to rate. This could be an indication that raters gave higher scores in categories they found most difficult to score. The Rasch analysis also suggests that raters perceived grammar and vocabulary as separate skills and thus these criteria should be scored separately. Furthermore, the Rasch measurements also indicated that raters were not using the full range of the rating scale. These findings underscored the importance of understanding raters' perception of the scoring criteria and scales.

In contrast to the results in scores, the quantitative analysis of the discourse between test-takers suggested that the interlocutor's language proficiency had an effect on the test-taker performance in terms of amount of talk. When comparing the total number of words produced per dyad, high-high dyads produced almost double the amount of words than the low-low dyads. This could be a representation of how more proficient test-takers had the language resources to express themselves better when compared to less proficient test-takers.

In addition to the analysis of raw amounts of talk, the present study also explored the proportion of contributions and found that when high-proficiency test-takers were paired with a partner of similar proficiency, they constructed a more balanced conversation. Great variation was observed in low-low pair types in terms of balanced conversations. Some dyads showed a balanced contribution in the amount of talk whereas others were imbalanced. The same was observed with mixed-proficiency dyads. Nevertheless, in all of the cases the high-proficiency test-takers quantitatively dominated their low-proficiency peers. These findings are consistent with those of Nakatsuhara (2006). The variation in the proportion of words across different pairs in low-low and high-low dyads suggest that interlocutor proficiency has little influence on the quantitative balance of conversations. Thus, the proportion of words contributed to a conversation might be attributed to factors other than language proficiency. Watanabe

(2008), for example, found that successful collaboration in peer interaction was not dependent on the interlocutor's proficiency but rather on the willingness of participants to share ideas. Along similar lines, in a study on peer interaction in group work, Choi and Iwashita (2016) revealed that learners reported perceiving interaction as being successful when their interlocutors showed more engagement and encouraged collaboration from other members of the group.

In contrast with Davis's (2009) findings, which showed an increase in the amount of talk when test-takers were paired with high-proficiency partners, the present study found minimal differences in the amount of talk between both conditions. More importantly, low-proficiency test-takers produced fewer words when paired with high-proficiency partners than when paired with same-proficiency partners. In terms of the proportion of total words, high-proficiency test-takers always produced a higher proportion of total words when paired with low-proficiency test-takers than when paired with same proficiency partners. On the other hand, except for two cases, low-proficiency test-takers produced a lower proportion of total words when paired with high-proficiency partners than when paired with same proficiency partners. Equivalent results were observed in the difference in the average number of words between pairing conditions for the low-proficiency test-takers. This proficiency group produced fewer words when paired with high-proficiency partners. The difference in these findings from those of Davis's (2009) study could be due to the differences in the L1 background of test-takers. It might be the case that low-proficiency Korean EFL learners are especially shy when it comes to speaking with another learner with a higher language proficiency level.

Finally, a moderate positive correlation was found between the number of words and the overall scores for both pairing conditions. Nevertheless, as Davis (2009) indicated, these results are not a representation of causality of the number of words to performance scores but rather an indication that both of these factors are jointly related with other variables that might help explain the variance in scores. In other words, talking more does not result in a higher score. The difference in the proportion of the words contributed between both pairing conditions also helps illustrate the same idea. The individual performance scores for each test-taker showed random differences in scores according to pairing condition. Some test-takers scored better or worse when speaking more or less with no systematic pattern.

Conclusion

The possibility of an interlocutor proficiency effect in paired oral assessments was explored by examining the test-takers' scores as well as their performance. On the one hand, it was found that the interlocutor's language proficiency did not affect the test-takers' scores in the paired tests, confirming the results in Davis (2009). On the other hand, the analysis of quantitative dominance between test-takers suggested that there were differences in the interactions due to the interlocutor's proficiency. While high-high dyads produced more balanced contributions to the interaction, low-low and mixed-proficiency dyads showed great variability in the degree of quantitative dominance. Particularly, in mixed-proficiency dyads, the high-proficiency test-takers tended to talk more than the low-proficiency test-takers resulting in more quantitatively imbalanced conversations.

Nevertheless, the differences in quantitative dominance did not seem to affect test-takers scores. Even though low-proficiency test-takers spoke less when paired with high-proficiency partners, their scores remained almost constant with very minimal differences. This might be explained by the raters' perception of what features are considered important to score, and their interpretations of the scoring rubric. Although the scoring rubric referred to characteristics of collaborative interaction as crucial for a higher score in discourse management, this might have been interpreted differently. As several studies (e.g., Ducasse & Brown, 2009; Galaczi; 2011, 2014; May, 2009, 2011) have suggested, rater training and the development of accurate scale and descriptors is crucial to ensure a fair, reliable and valid measure of oral proficiency through paired oral tests.

Finally, there were several limitations in this study. As with the original study, the findings should be interpreted with caution due to the small sample size. Particularly, the Rasch measurements require a larger sample size to provide robust results. The analysis of a larger sample size can provide stronger evidence for the claims made as well as detect smaller differences. In addition, a greater difference in proficiency level between low- and high-proficiency groups could have yielded different results. Moreover, a single task was used in each paired test which might have limited the amount of conversational strategies the test-takers could use. Other types of tasks should also be included in order to assess a wider range of conversational features and have a more precise measurement of score reliability.

Furthermore, for future studies, a qualitative analysis following a micro-analytic approach could shed some light on the differences in turn-taking strategies as well as degree of involvement between same- and mixed-proficiency pairs. A preliminary

analysis indicated a tendency for high-high pair type to shift topics in a more natural and stepwise manner, unlike low-low and some mixed-proficiency dyads. This is in line with the findings in Galaczi (2014). A preliminary analysis of topic management in terms of the number of topic initiation turns and topic continuation turns was not insightful because these counts tended to obscure the manner in which turn-taking was taking place. Thus, for future studies this type of analysis should follow the CA conventions and offer finer-grained analyses of topic management. In addition, stimulated recalls from raters and examinees might be used as a method to examine the perceptions of raters when they score certain features of interaction, explaining the match or mismatch between scores and variation of conversational features.

Acknowledgements

I would like to thank Dr. John Norris for his constant support and guidance in improving this project. I am also indebted to Dr. Larry Davis who provided me with insightful comments on this article. My thanks also go to Bart Deygers for helpful comments and encouragement to publish this manuscript. This article was a further development of my master's thesis, and I am also grateful to my MA advisor Dr. Yong-Wong Lee who helped me at different stages in the development of this study.

References

- Berry, V. (2007). *Personality differences and oral test performance*. Frankfurt: Peter Lang.
- Bonk, W.J. & Van Moere, A. (2004, March). *L2 group oral testing: The influence of shyness/outgoingness, match of interlocutor's proficiency level, and gender on individual scores*. Paper presented at the Language Testing Research Colloquium.
- Brooks, L. (2009). Interacting in pairs in a test of oral proficiency: Co-constructing a better performance. *Language Testing*, 20(1), 89–110.
- Brown, A. (2003). Interviewer variation and the co-construction of speaking proficiency. *Language Testing*, 20(1), 1–25.
- Choi, H. & Iwashita, N. (2016). Interactional behaviors of low-proficiency learners in small group work. In M. Sato and S. Ballinger (Eds.), *Peer interaction and second language learning: Pedagogical potential and research agenda* (pp. 113–134). Amsterdam and Philadelphia: John Benjamins.

- Csépes, I. (2009). *Measuring Oral Proficiency through Paired-Task Performance*. Vol. 14. Frankfurt: Peter Lang.
- Davis, L. (2009). The influence of interlocutor proficiency in paired oral assessment. *Language Testing* 26(3), 367–396.
- Ducasse, A. M., & Brown, A. (2009). Assessing paired orals: Raters' orientation to interaction. *Language Testing*, 26(3), 423–443.
- East, M. (2015). Coming to terms with innovative high-stakes assessment practice: Teachers' viewpoints on assessment reform. *Language Testing*, 32(1), 101–120.
- Együd, G. & Glover, P. (2001). Oral testing in pairs – a secondary school perspective. *ELT Journal*, 55, 70–76.
- Fernández Dobao, A. (2012). Collaborative dialogue in learner–learner and learner–native speaker interaction. *Applied Linguistics*, 33(3), 229–256.
- Fernández Dobao, A. (2014). Vocabulary learning in collaborative tasks: A comparison of pair and small group work. *Language Teaching Research*, 18(4), 497–520.
- French, A. (1999). *Study of qualitative differences between CPE individual and paired test formats* (Internal UCLES EFL report). Cambridge, UK: University of Cambridge Local Examinations Syndicate.
- Foot, M.C. (1999). Relaxing in pairs. *ELT Journal*, 53, 36–41.
- Fulcher, G. (1996). Testing tasks: Issues in task design and the group oral. *Language Testing*, 13, 23–51.
- Fulcher, G. (2003). *Testing Second Language Speaking*. London: Pearson Longman.
- Galaczi, E.D. (2008). Peer-peer interaction in a paired speaking test: The case of the First Certificate in English. *Language Assessment Quarterly*, 5(2), 89–119.
- Galaczi, E. D. (2014). Interactional competence across proficiency levels: How do learners manage interaction in paired speaking tests? *Applied Linguistics*, Advanced Access published August 30, 2013, doi:10.1093/applin/amt017.
- Galaczi, E. D., French, A., & Hubbard, C. (2011). Developing assessment scales for large-scale speaking tests: A multiple method approach. *Assessment in Education* 18(3), 217–237.
- Itakura, H. (2001). Describing conversational dominance. *Journal of Pragmatics*, 33, 1859–1880.

- Iwashita, N. (1998). The validity of paired interview format in oral performance assessment. *Melbourne Papers in Language Testing*, 5, 1–65.
- Kasper, G. (2013). Managing task uptake in oral proficiency interviews. In S. Ross and Kasper, G. (Eds.), *Assessing Second Language Pragmatics* (pp. 258–287). New York: Palgrave Macmillan.
- Knoch, U. (2009). Diagnostic assessment of writing: A comparison of two rating scales. *Language Testing*, 26(2), 275–304.
- Lazaraton, A. (1996). Interlocutor support in oral proficiency interviews: The case of CASE. *Language Testing*, 13(2), 151–172.
- Lazaraton, A. (2006). Process and outcome in paired oral assessment. *ELT Journal*, 60, 287–289.
- Linacre, J. M. (2013). *Facets 3.71.3*. Computer program. Chicago: MESA Press.
- May, L. A. (2000). Assessment of oral proficiency in EAP programs: A case for pair interaction. *Language & Communication Review*, 9(1), 13–19.
- May, L.A. (2009). Co-constructed interaction in a paired speaking test: The rater's perspective. *Language Testing*, 26(3), 397–422.
- May, L.A. (2011). Interactional competence in a paired speaking test: features salient to raters. *Language Assessment Quarterly*, 8(2), 127–145.
- McNamara, T. (1996). 'Interaction' in second language performance assessment: Whose performance? *Applied Linguistics*, 16, 159–179.
- Nakatsuhara, F. (2004). *An Investigation into conversation styles in paired speaking tests*. Unpublished master's thesis, University of Essex, Wivenhoe Park, Essex, United Kingdom.
- Nakatsuhara, F. (2006). *The impact of proficiency-level on conversational styles in paired speaking tests*. University of Cambridge ESOL Examinations Research Notes 25, 15–20.
- Nakatsuhara, F. (2008). Inter-interviewer variation in oral interview tests. *ELT Journal* 62(3), 266–274.
- Nakatsuhara, F. (2011). Effects of test-taker characteristics and the number of participants in group oral tests. *Language Testing* 28(4), 483–508.
- Norton, J. (2005). The paired format in Cambridge Speaking Tests. *ELT Journal*, 59, 287–297.

- Norton, J. (2013). Performing identities in speaking tests: Co-construction revisited. *Language Assessment Quarterly*, 10(3), 309–330.
- Ockey, G. J. (2009). The effects of group members' personalities on a test taker's L2 group oral discussion test scores. *Language Testing*, 26(2), 161–186.
- O'Sullivan, B. (2002). Learner acquaintanceship and oral proficiency test pair-task performance. *Language Testing*, 19(3), 277–295.
- O'Sullivan, B. & Nakatsuhara, F. (2011). Quantifying conversational styles in group oral test discourse. In B. O'Sullivan (Ed.), *Language Testing: Theories and Practices*. New York: Palgrave MacMillan, 164–185.
- Philp, J., Adams, R., & Iwashita, N. (2014). *Peer interaction and second language learning*. New York: Routledge.
- Porte, G. (2012). *Replication research in applied linguistics*. Cambridge: Cambridge University Press.
- Powers, D. (2010). *The Case for a Comprehensive, Four-Skills Assessment of English-Language Proficiency R & D Connections, No. 14*. Princeton, NJ: Educational Testing Service.
- Ross, S. (2007). A comparative task-in-interaction analysis of OPI backsliding. *Journal of Pragmatics*, 39, 2017–2044.
- Sawaki, Y., Stricker, L., & Oranje, A. (2008). *Factor structure of the TOEFL® internetbased test (iBT): Exploration in a field trial sample* (ETS Research Rep. No. RR-08-09). Princeton, NJ: Educational Testing Service.
- Swain, M. (2001). Examining dialogue: Another approach to content specification and to validating inferences drawn from test scores. *Language Testing*, 18(3), 275–302.
- Taylor, L. (2001, November). *The paired speaking test format: recent studies*. University of Cambridge Local Examinations Syndicate Research Notes 6, 15–17.
- Taylor, L. (2003, August). *The Cambridge approach to speaking assessment*. University of Cambridge Local Examinations Syndicate Research Notes, 2–4.
- Watanabe, Y. (2008). Peer-peer interaction between L2 learners of different proficiency levels: Their interactions and reflection. *The Canadian Modern Language Review*, 64, 605–635.
- Wright, B. D. (2000). *Rating Scale Model (RSM) or Partial Credit Model (PCM)?* Retrieved January 5, 2015. <http://www.rasch.org/rmt/rmt1231.htm>

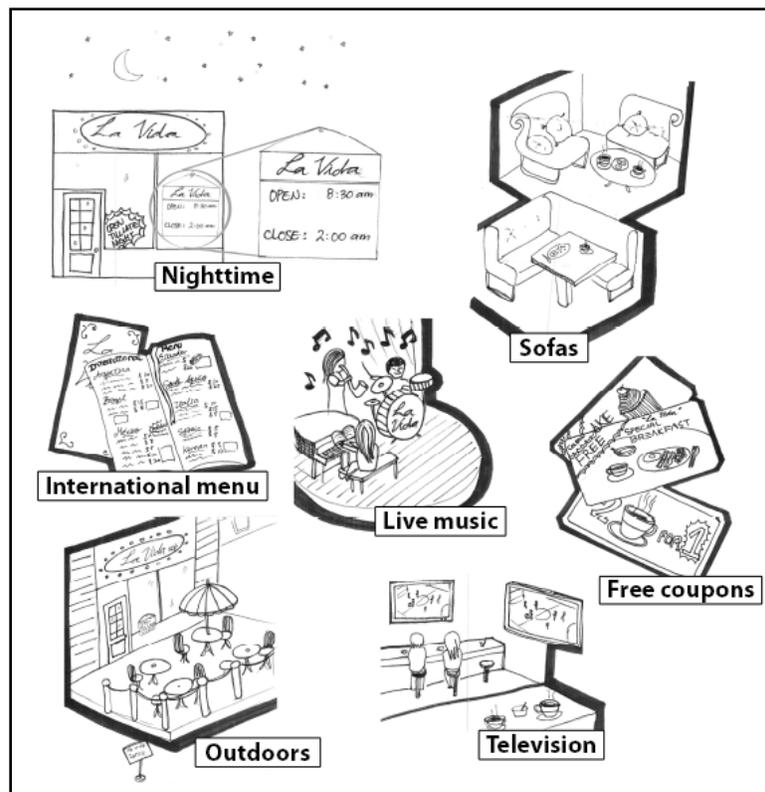
Appendix A

Coffee Shop Prompt

Imagine your friend opened a new café (*La Vida Café*) in Seoul and wants to attract more people. He has given you some suggestions (shown in the picture) and has asked you to help him find the best two options to get more customers. Discuss with your partner to find the best two options for improving your friend's café. Listen to your partner's opinion and also express your opinion.

First, **briefly** talk to each other about **how successful each suggestion might be**.

Then, decide together **which two suggestions would be the best to attract more people**.



Appendix B

Analytic Scale

| | 1 | 2 | 3 | 4 | 5 |
|----------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Grammar | Uses a very limited range of structures inaccurately. Produces mostly incomplete sentences that consist of only basic and inaccurate structures. There are frequent errors that interfere with communication. | Only uses basic structures that are at times inaccurate. All attempts, if any, to use complex forms are inaccurate. Errors are noticeable and may impede communication of the intended message. | Mostly uses basic structures that are occasionally inaccurate. There is some attempt to use complex forms but it is mostly imprecise. Repair techniques are frequently present but not always successful. | Uses a wide range of structures with few inaccuracies. Errors are especially present when attempting to use more complex structures. Can generally use repair techniques. | Uses a full range of complex structures appropriately and accurately. There might be minor errors but they do not interfere with communication. Can effectively use repair techniques. |
| Vocabulary | Uses a very limited range of basic words. The choice of words is inaccurate and repetitive. Errors using expressions are frequently present affecting communication. | Only uses basic vocabulary which is frequently inaccurate and may affect communication. The choice of words is frequently inappropriate and may impede the delivery of the message. | Uses a limited range of vocabulary that is occasionally inaccurate. The choice of expressions is sometimes inappropriate and may interfere with delivery of the intended message. | Uses a wide range of complex words, although they are at times imprecise. The word choice is generally appropriate. There are still a few errors but they do not interfere with delivery of the intended message. | Uses a full range of complex and simple vocabulary effectively and accurately. The choice of expressions is appropriate to successfully deliver of the intended message. |
| Pronunciation | Produces speech that is unintelligible due to very frequent mispronunciations. There are constant stress- and single-sound errors that make the utterances hard to understand. The intonation is unnatural. | Produces speech that is heavily influenced by L1 prosodic features, and which is nearly always difficult to understand. There are various errors that require effort on the part of the listener to understand. The intonation is unnatural. | Produces speech that is marked by L1 prosodic features making it at times difficult to understand. There are some errors in single sounds or stress that may be noticeable in some words. The intonation is at times unnatural. | Produces speech that is influenced by some L1 prosodic features. However, this has minimal effects on intelligibility; thus the speech is easy to understand. There are few errors in single sounds or word stress. | Produces speech that is highly intelligible. The word-stress/ rhythm/ intonation are mostly accurate; thus effortless to understand. There may be a few minimal errors but overall the speech is natural. |
| Fluency | Constant hesitation and long pauses make speech hard to be rated. Speech is full of incomplete sentences that make the delivery very choppy and hard to understand. | Frequent hesitation and long pauses make speech slow and occasionally impedes comprehensibility. There are some abandoned sentences that impede the natural flow of speech. | Hesitation and long pauses are frequent when searching for words or the correct grammatical structure, sometimes impeding the natural flow of speech. | Hesitation and pauses are seldom present when searching for words but they do not impede the natural flow of speech. | Hesitation and pauses are present but they are few and only in content-related situations. Efficiently keeps a natural pace. |
| Discourse Management | Shows almost no willingness to express opinions and expand ideas. Is unable to initiate a topic. Produces only minimal responses to partner's speech. His/her role in the conversation is too passive. | Expresses mostly basic and simple ideas and does not further develop them. There are few attempts to initiate a topic. His/her contributions are limited to minimal acknowledgements; thus his/her role is passive. | Expresses ideas and opinions in a simplistic way. There is a lack of details and further expansion of ideas. Can initiate topics but only expands on his/her own topic with minimal acknowledgment to the partner's utterances. | Generally expresses opinions and arguments in an organized and detailed manner. Can initiate a topic and expand it. However, responses to partner's ideas may sometimes be too simple, failing to elaborate further into a topic. | Effectively expresses ideas and opinions in detail connecting ideas previously mentioned. Shows active engagement in the conversation by initiating a topic and expanding his/her own as well as the partner's topic. |

Appendix C

Vertical Map

| Measr | +examinees | -Rater | -Condition | -Items | Scale |
|-------|------------|--------|----------------------------------------|--------|---------|
| 7 | + | + | + | + | (5) |
| | * | | | | |
| 6 | + | + | + | + | |
| | | | | | |
| 5 | + | + | + | + | |
| | | | | | |
| 4 | *** | + | + | + | |
| | *** | | | | |
| | ** | | | | --- |
| 3 | + | + | + | + | |
| | ** | | | | |
| 2 | + | + | + | + | 4 |
| | * | | | | |
| 1 | + | + | + | + | |
| | * | | | G | |
| | | Penny | | F | |
| | | Roz | | P | |
| * 0 | ** | | * differentproficiency sameproficiency | V | * --- * |
| | | | | | |
| -1 | + | + | + | + | DM |
| | * | | | | |
| -2 | + | + | + | + | 3 |
| | ** | | | | |
| | * | | | | |
| -3 | + | + | + | + | --- |
| | * | | | | |
| | * | | | | |
| -4 | + | + | + | + | (2) |
| | | | | | |
| Measr | * = 1 | -Rater | -Condition | -Items | Scale |