문학석사 학위논문

The Interlocutor Proficiency Effect on Test-taker Performance in a Paired Oral Assessment

짝 형식 말하기평가에서 대화상대자의 영어능력이 수험자의 말하기 수행에 미치는 영향

2013년 2월

서울대학교 대학원

영어영문학과 어학전공

손 영 아

The Interlocutor Proficiency Effect on Test-taker Performance in a Paired Oral Assessment

지도교수 이 용 원

이 논문을 문학석사 학위논문으로 제출함

2012년 10월

서울대학교 대학원

영어영문학과 어학 전공

손 영 아

손영아의 문학석사 학위논문을 인준함

2012년 12월

위 원 장 _____ **(인)**

부위원장 _____ **(인)**

위 원 _____ **(인)**

Abstract

Paired speaking tests have increasingly gained popularity in the field of language assessment for their authentic ways of assessing interactive competence. Despite the benefits associated with paired speaking tests, questions have been raised in terms of their fairness, reliability, and construct validity. Given that paired assessments involve the interaction of two or more test-takers, a major concern is the effect the interlocutor's proficiency might have on their partners during the interaction. In this regard, a relatively small body of previous literature has yielded contradicting results.

This study empirically examined the effects of interlocutor proficiency on test-taker performance in paired speaking tests. It used a scheme similar to that of Iwashita (1998) and Davis (2009). Twenty-four Korean university students were assigned to two major proficiency groups (i.e., high and low) and were asked to take two separate paired tests under different conditions: once paired with a partner at a similar proficiency level and once with a partner at a different proficiency level. A mixed-design two-way analyses of variance (ANOVA) were conducted using the pairing conditions as independent variables. First, the analysis of variance examined the statistical significance of the score differences between the two paired test conditions through composite scores (i.e., aggregated scores of the five analytic criteria, namely; grammar; vocabulary; pronunciation; fluency; and discourse management). Second, the scores for each of the analytic criteria were examined separately, again using ANOVA. Finally, each gender group was also analyzed to investigate whether gender was a factor that moderated the

interlocutor proficiency effect.

The analysis of the results indicated that the interlocutor's proficiency had no significant effect on the composite scores of participants. In addition, the interlocutor's proficiency had no statistically significant effect on any of the analytic scores assigned for the rating criteria. Furthermore, even when each gender group was examined separately, the interlocutor's proficiency turned out to have no statistically significant effects on the composite and analytic scores of female as well as male participants. Overall, the interlocutor's proficiency had no statistically significant effect on test-taker performance in the paired speaking tests. The present study has provided evidence in support of the use of paired speaking tests as a tool to measure speaking ability in a valid, reliable, and fair manner.

Keywords: interlocutor proficiency, interaction, paired speaking assessment
Student number: 2010-22943

Table of Contents

List of Tables

List of Figures

CHAPTER I

INTRODUCTION

1.1 Background and Motivation

The ability to speak is one of the most valuable yet difficult language skills to acquire when learning a foreign language (Lado, 1961, Bachman & Alderson in Luoma, 2004). Acquiring the ability to communicate thoughts in real-time is certainly one of the main goals and challenges for language learners. For this reason, speaking has inevitably been a major component in the foreign language teaching curriculum. At the same time, this has generated considerable research interest in speaking assessment in the language testing community.

Ever since language testers began to focus their attention on testing second language speaking, finding reliable and authentic ways to do it has been a major challenge. One of the most daunting tasks has been finding a definition for *what it means to be able to speak* (Fulcher, 2003) and to apply this definition in the design and development of speaking tasks as well as rating scales. In addition, researchers have pointed out that there is an urgent need for tests that can assess the interactive competence of learners (McNamara, 1997). Accordingly, for the last two decades, there has been a growing body of literature on pairwise speaking tests which assess the interactive competence using authentic tasks.

The importance of the study of paired speaking assessments lies mainly in the advantages associated with this particular format of testing. Several

studies (e.g., ffrench, 1999; Saville & Hargreaves, 1999; May, 2000; Együd & Glover, 2001; Taylor, 2001; Brooks, 2009) have presented evidence in support of the use of this type of assessment. One benefit, for example, is its practicality. In paired assessments, two or more test-takers can be assessed at once. Moreover, the tasks in paired tests closely resemble classroom activities encouraging positive washback (Swain, 2001).

Most importantly, some qualities of the paired speaking tests have been demonstrated to have an advantage over the traditional interview format. One of these advantages is that it elicits "a richer and more varied sample of spoken language" (Taylor, 2001:15). This means that students are able to demonstrate their speaking ability in a more authentic manner. The types of interaction observed in paired tests are more symmetrical since test-takers are asked to interact with their peers (Brooks, 2009). Therefore, the speaking performances observed during the testing can reflect their real speaking ability in non-test situations.

Notwithstanding the various benefits of paired speaking tests, questions have been raised as to their fairness, reliability and construct validity (Weir, 1993; McNamara, 1997; Foot, 1999; Fulcher, 2003). One of the biggest concerns is the effect the interlocutors might have on their partners during the interaction. Due to the fact that paired tests involve co-constructed interaction, speaking performance might be affected by a mismatch in pairings of the interlocutors, which could potentially undermine the construct validity of test scores. Although this appears to be a logical reason for preventing the use of paired assessments, more empirical research is needed to investigate these

claims.

Several studies have recognized the need for more concrete evidence and have explored various factors associated with the interlocutor's effect on speaking performances, including personality, acquaintance, gender, and language proficiency (e.g. Berry, 1997; O'Sullivan, 2002; Bonk & Van Moere, 2004; Davis, 2009). Even though this topic has gained popularity among speaking assessment researchers, there is still a need to reach a consensus on how ratings are affected by these conditions.

Only a few empirical studies, both qualitative and quantitative, have focused on the interlocutor proficiency effect, and they have yielded mixed results. Foot (1999), for example, argues that the partner's language ability might be problematic and disadvantageous due to the possible incomprehension between test-takers. Unfortunately, his study – as well as several others – has not provided a quantitative analysis of this negative effect. Empirical studies, such as Iwashita (1998), demonstrated that test-takers belonging to both high- and low-proficiency groups tended to benefit from being paired with a partner from the high-proficiency group. Even though the study took an experimental approach when collecting data, no inferential statistics were used to analyze data; only descriptive statistics were presented as evidence. On the other hand, Davis (2009) analyzed the data using Rasch analysis and reported mixed results. The study indicated that different pairing dyad conditions made little or no difference in the test-takers' speaking scores.

Following these studies, the present investigation intends to further explore this controversial topic by examining the effects of interlocutor

proficiency on test-takers' performances in paired speaking tests in the Korean EFL (English as a Foreign Language) context. It attempts to contribute to a relatively small body of research literature on this particular topic by providing empirical evidence and exploring how this factor might undermine the validity of the test scores in paired speaking tests.

In the first place, it is crucial to examine the construct validity and reliability of paired speaking tests. The underlying assumption is that if there are construct-irrelevant factors (e.g., the interlocutor's proficiency) that affect the test-takers' test scores other than test-takers' English speaking ability, the validity of this test might be threatened (Bachman, 1990; Weir, 2004). The current study used a similar scheme to that of Iwashita (1998) and Davis (2009). In those studies test-takers were assigned to two major proficiency groups (i.e., high and low), and were asked to take two separate paired tests under different conditions: once paired with a partner of similar proficiency and once with a partner of different proficiency. Thereafter, the scores obtained during both paired tests were analyzed to investigate whether there was a statistically significant difference in scores between both pairing conditions.

Secondly, the present study explores how the interlocutor's proficiency affects the scores of five different analytic criteria, namely grammar, vocabulary, pronunciation, fluency, and discourse management. A previous qualitative study by Norton (2005) indicated that in several conversations in the Cambridge ESOL First Certificate of English (henceforth, FCE) some test-takers used a method referred to as "appropriation" of grammatical structures

and vocabulary. In other words, some examinees would "appropriate" certain syntactic structures or words used by their partners. The researcher implied that this could be beneficial to lower-proficiency level test-takers when paired with higher-proficiency level partners. This study therefore attempts to investigate whether these claims are supported by empirical evidence.

Lastly, analyses of score were also conducted to explore whether gender moderated the interlocutor proficiency effect. Since gender effects have been considered to be *complex and unpredictable* (Brown & McNamara, 2004), the present study only examined same-sex dyads. This made it possible for the researcher to observe the differences in patterns of scores for the different pairing conditions for each gender group.

1.2 Research Questions

In this study, the test-takers' overall speaking performances were examined through trait-composite scores, which are the sum of scores on 5 analytic criteria (i.e. grammar, vocabulary, pronunciation, fluency, and discourse management). Having summarized the main points and motivation, this study intends to answer the following research questions:

1. Does the interlocutor's proficiency level have a significant effect on the test-taker's overall performance as represented by composite scores in the paired oral tests?
2. Does the interlocutor's proficiency level have a significant effect on the test-taker's performance in each of the analytic scores used for the paired oral tests?
3. Are there similar patterns of interlocutor proficiency effect observed

across different gender groups when composite as well as analytic scores are used as criterion measures?

1.3 Organization of the Thesis

The present study is organized as follows. Chapter 2 summarizes previous studies on the interlocutor effect in speaking tests. Chapter 3 thoroughly describes the methods and procedures used to collect data. Chapter 4 presents the results of the statistical analyses which were conducted to examine the interlocutor proficiency effect on the participants' speaking scores, i.e., analytic and composite scores. Chapter 5 discusses the major findings of the current study by comparing them with the findings of the previous research studies. Finally, Chapter 6 discusses implications that can be drawn from the findings, reports of some of the caveats of the study, and concludes with suggestions for future studies.

CHAPTER II

LITERATURE REVIEW

This chapter presents a review of previous studies regarding the interlocutor effect in oral tests. It begins with a brief description of a theoretical framework of oral performance testing illustrated by a model proposed in Csépes (2009). Then the next section discusses some of the empirical studies of interlocutor's effect in traditional face-to-face interviews followed by a summary of previous research studies on the comparison between this format and the paired format of speaking assessment. Next, previous studies on the interlocutor's effect in paired speaking tests are summarized. Most importantly, this chapter ends by giving a detailed description of research on the interlocutor proficiency effect in paired speaking tests.

2.1 Theoretical Framework of Speaking Tests

In order to clearly understand the different components of language ability and their relationship to language testing, several models (e.g., Canale & Swain, 1980; Bachman & Palmer, 1996; Chapelle *et al.*, 1997) have been used as reference frameworks. Some models, namely those of Milanovic and Saville (1996) and McNamara (1996), have focused on the interaction between various factors and test-taker performance on the test. According to McNamara (1997), a test-taker's ability is not the only element that affects test performance. Therefore, the last two models mentioned above also include raters, tasks, rating scales, and interlocutors, among other elements in

their model. Most notably, they illustrated the interaction between examinees and examiners. Based on these models, Csépes (2009) proposes a new model (Figure 2.1) that attempts to include the interaction between a candidate and the interlocutor specifying the type of interlocutor as examiner and candidate. In this way the interaction that takes place in a paired speaking assessment is also considered as part of what affects the test-taker's performance.

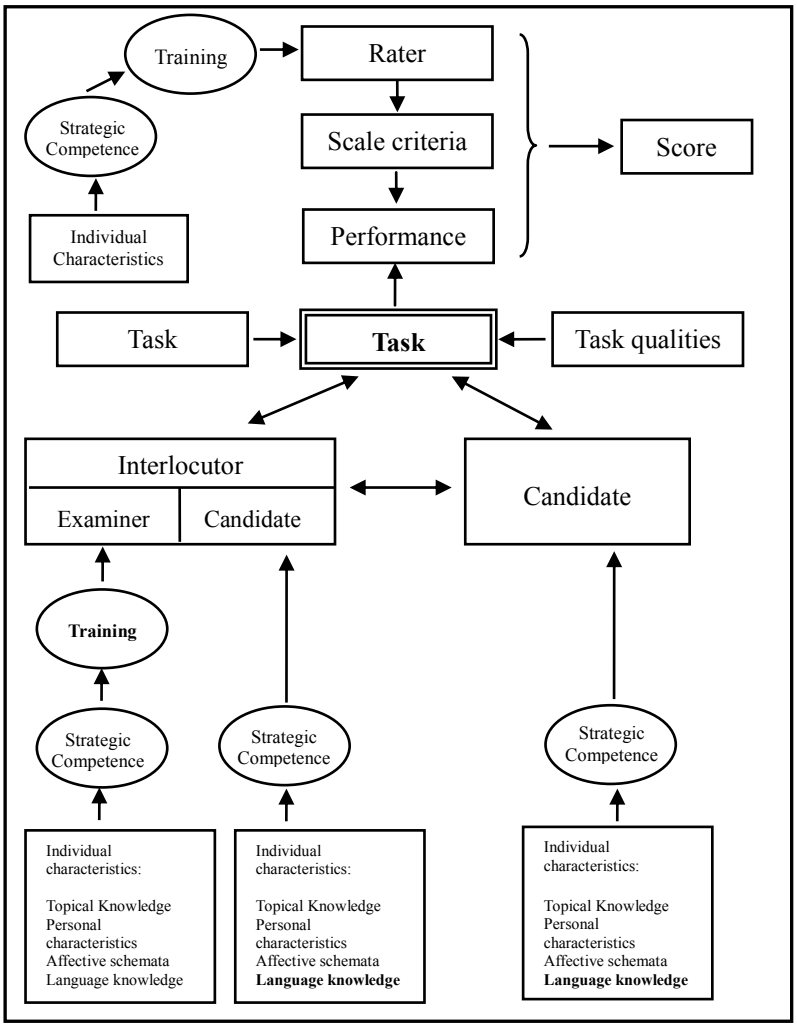


Figure 2.1 Csépes (2009: 31)'s extended model of oral performance testing

Most of the literature on the interlocutor effect in oral tests has concentrated on the interaction between candidates and examiners, without regarding the candidate-candidate component. Nevertheless, over the last two decades, the latter type of interaction has also been under scrutiny. The underlying rationale behind the research for each format of speaking assessment is the same. As demonstrated in Figure 2.1, the oral tests involve an interaction between candidates and interlocutors. The interlocutors can be examiners in the oral interviews or peer candidates in the paired speaking tests. Each of these participants brings their own strategic competence and characteristics into play. Hence, studies regarding this issue have mainly focused on the variability of test-takers' performance due to the interlocutor's characteristics.

2.2 Studies on Traditional Face-to-Face Interviews

When one-to-one interviews were first developed and adopted for speaking assessments, language testers accepted them as having a potentially positive washback effect. As stated in Liskin-Gasparro (2003), Oral Proficiency Interviews (henceforth, OPI), for example, created new insights into how languages were being taught in classrooms. Instructors started to realize the importance of communication between students as opposed to just between teachers and students, which led to pedagogical innovation. Apart from the impact the OPI had on education, it also gained popularity for its face validity (Fulcher, 1997). Test users considered it to be "the best and fairest measure of oral ability" (Van Lier, 1989: 490), since it was one of the earliest tests to

capture the true speaking ability of candidates in a communicative context. Meanwhile, in the United Kingdom the University of Cambridge Local Examinations Syndicate (henceforth, UCLES) was not hindered by concerns about validity and reliability. Its purpose was to develop tests that would encourage communicative teaching; therefore, speaking components were regarded as crucial in all examinations.

Whilst the language elicitation method of the OPI appeared ostensibly authentic, a closer examination of the interaction between examiner and examinee cast doubts on this claim. Van Lier (1989), for example, discussed the similarities and differences between the interactions in the OPIs and real-life conversations. By analyzing several interviews, the study found that interviews showed asymmetrical contingency and/or pseudocontingency, as opposed to reactive and mutual contingency observed in natural conversations. In other words, in the OPI the interviewer had "control over the discourse by asking questions and evaluating answers" (p. 499) and this differed from conversations where both parties had equal rights. As a result, the construct validity of the OPI as a test that assesses the "speaking ability in real-life context" (Education Testing Service, 1982: 13 cited in Van Lier (1989)) was called into question.

Following the same line of reasoning, Johnson and Tyler (1998) compared one interview from Level Two OPI with everyday conversations and reached similar conclusions. The study based the comparison on key features of natural conversations presented in Sacks, Schegloff and Jefferson (1974). The results of the conversational analysis demonstrated that the turn

order, size, and distribution were fixed, the amount of talk was imbalanced between examiners and test-takers, and the turn allocation was mostly limited to the current-speaker-selects-next technique. Furthermore, the role of the interviewer appeared to be fixed and predetermined when managing the topics to be discussed. These are all characteristics not found among the salient features of natural conversations. As a result, the study concludes by questioning the validity of OPIs since they have been described as a tool to measure "speaking ability in a real-life context" (Education Testing Service, 1982:13).

Other studies questioned the validity of face-to-face interviews by addressing issues related to the interlocutor's effect on examinees during the interaction. Ross and Berwick (1992) investigated the interviewers' degree of control over the interview as well as their accommodative behavior when trying to facilitate communication with their interviewees. The study analyzed a total of sixty different OPIs from four different language proficiency levels, in order to explore how the interviewer's perception of the interviewee's proficiency level affects the accommodating behavior. The findings indicated that the features of accommodation and control found during the OPIs were parallel to those found in ordinary conversations between native and non-native speakers. As in the real-life context, the interviewer accommodation was more frequent in interviews with less proficient test-takers. In addition, interviewers had control over the interaction by nominating and reformulating the topics. This was also the case in the Cambridge FCE, where examiners showed more control over the topics discussed (Young & Milanovic, 1992). A

major limitation, and one that threatened the construct validity of this test, was that the interviewers were not aware of their accommodating behavior and were inclined to overuse this technique even when it was not necessary. As a result, the researchers emphasized the need to train interviewers on their discourse behavior.

In this regard, Ross (1992) presented similar evidence that reiterated the importance of interviewer training. This study focused on the occurrence of accommodation in questions asked by the interviewers. The results showed that there were four triggers of accommodation, namely; the interviewee's response to previous questions; structure of the responses; language proficiency level; and the presence of accommodation in previous questions. As in Ross and Berwick (1992), the frequency of these accommodations differed according to the interviewer's perception of the interviewee's language proficiency. Therefore, the study suggested that this frequency of accommodation should also be taken into account in the final ratings in order to include the interviewer's participation in the interaction. At the same time, interviewer training should include the distinction between accommodations that are necessary for simplification from accommodations that are superfluous.

Similar findings were reported in Lazaraton (1996), which analyzed data from 58 interviews of the interview section of the Cambridge Assessment of Spoken English (CASE) for features of interlocutor's support. As in the previously mentioned studies, the results showed that there were certain discourse qualities present in the interviews that can also be found in non-

assessment contexts. Even though this was a positive outcome, since it confirmed that these interviews were assessing the test-takers' ability to engage in a conversation, it still raised questions about the construct validity of the test scores, given that the interviewers' speech behaviors were inconsistent. Analysis of the data demonstrated that there were 8 types of interlocutor support: priming topics; suggesting words; evaluating responses; repeating and correcting responses; overarticulating speech; stating questions that required only confirmation; drawing conclusions; and rephrasing questions. Nevertheless, the frequency with which these techniques were used was not consistent across all interviewers, giving unequal opportunities for candidates to demonstrate their speaking abilities.

Brown (2003) provides more evidence in this regard by examining two IELTS speaking module interviews, involving the same candidate and two different interviewers. Eight ratings were given for each interview. The scores as well as the raters' verbal protocol show that the candidate was perceived to perform better when being interviewed by one interviewer over another. Conversation analysis of the interviews provided evidence of the impact of interviewers over the candidate's speaking performance. Differences in the way both interviewers conducted the interviews, particularly in the "structuring of topical sequence, questioning techniques, feedback and rapport" (p. 17) influenced the way raters perceived the candidate's speaking proficiency. At the same time, the speaking scores were also affected by the interviewers' behavior. However, since this was a qualitative study involving one single student, there were no references to its statistical significance.

Research on traditional one-to-one interviews has raised important questions about interlocutor behavior and its impact on test-takers' speaking performances. Overall, the studies emphasized the need to train interviewers in order to reduce variability from one test to another since this could undermine the construct validity of the speaking scores. Most importantly, these studies set the tone for future research concerning the interlocutor effect in direct speaking assessments, including paired assessments.

2.3 Studies on Paired Speaking Tests

2.3.1 Rationale for Adopting the Paired Speaking Assessment

The United Kingdom has a long experience with direct speaking tests. Taylor (2003) reports that the Cambridge Local Examinations Syndicate (UCLES) has always deemed direct speaking tests to be important. Interest in this type of assessment dates back to the development of the first UCLES English language test, the Certificate of Proficiency of English (henceforth, CPE), in 1913. This test already included a direct speaking test that was composed of a dictation section that lasted half an hour, and a second section for reading aloud together with a conversation that lasted half an hour (Fulcher, 2003). In addition, the Lower Certificate of English, the equivalent of the FCE, was introduced in 1939, and had a face-to-face speaking component. This test was revised numerous times over the years in order to have a positive impact on language education (Saville & Hargreaves, 1999).

According to Taylor (2001), Cambridge ESOL decided to adopt the paired format of speaking assessment mainly for pedagogical reasons as well

as for the advantage it had in eliciting speech samples with greater variety than the traditional interview format offered. Swain (2001) provides evidence in support of the first argument. In this study, the main focus was the interface between second language learning and second language testing. By using a pre-test and post-test in the form of paired interviews, Swain demonstrated that the negotiation of meaning between learners during this task promoted learning. In this matter, Taylor and Wigglesworth (2009) also assert that this test format is advantageous in the learning context since participants are required to use both their *receptive and productive skills*.

In regards to the second argument, several studies (e.g. ffrench, 1999; Taylor, 2001; Brooks, 2009, May, 2000) have focused on the comparison between the one-to-one traditional interviews and the paired speaking assessment. As the aforementioned studies about traditional interviews have shown, the interaction between the examiners and examinees have been characterized as being asymmetrical, since both participants have to follow a fixed role (i.e., interviewer and interviewee). In contrast, in the paired format test-takers must interact not only with the examiners but with other test-takers as well. Thus, there is evidence of "a richer and more varied sample of spoken language" (Taylor 2001:15). In ffrench (1999), qualitative analysis of data from the CPE indicated that the percentages of the distribution of speaking functions in the paired format are significantly different from those of the interview format. In the latter format, the interactional functions as well as the managing of these interaction functions are part of a very small percentage of the total performance. Meanwhile, in the former format, the percentages of

these functions are much larger accounting for almost 50% of the whole performance.

In Brooks (2009), a total of 16 candidates were likewise assessed using both formats of speaking assessment. The tasks consisted of a short text, an idea map, and some discussion questions that would make the test-takers think about the topic. The participants were allowed to veer off the topic and the use of the discussion questions was optional. The same topic and questions were used in both the paired test and interview. Quantitative analysis showed that the test-takers scored higher during the paired format than the interview format. In other words, candidates' performed better when interacting with other test-takers than with an examiner. The qualitative analysis revealed that there was "more interaction, negotiation of meaning, consideration of the interlocutor and more complex output" (p. 341) in the paired format than in the interview format. Furthermore, the results on the number of features of interaction echoed those in ffrench (1999) and Taylor (2001). There were a larger number of features of interaction observed in the paired format compared to the interview format.

Another argument in favor of paired speaking assessments was explored in May (2000). This study took a different approach and explored the insights of test takers on both formats of speaking assessment. In this study 32 Chinese university students were assessed in two ways: once in a one-to-one interview and once paired with another test-taker. The researcher used questionnaires to compare test-taker reactions to both formats. The results showed that students preferred the paired format over the interview format, finding it "more natural"

and “more relaxing” since there was no power differential. In addition, they considered it a more authentic and effective method of assessing and improving their speaking skills. Együd and Glover (2001) reported similar results with a group of Hungarian secondary school students. When qualitatively comparing the performances of students who took the speaking test in pairs with those who interacted with interviewers, the study found that students seemed to have a wider range of opportunities to show their speaking abilities in the former rather than the latter test. Furthermore, the students’ comments clearly showed a preference for paired assessments since they felt less stress and were able to “give and receive help”. Fulcher (1996) reported similar results from a group of Greek students, who answered questionnaires after completing three oral tasks that included two one-to-one interviews and one group discussion. As with the previously mentioned studies, most of the students preferred the group oral test and found it more enjoyable since they could take the test with friends and “it didn’t feel like a test” (p. 34).

In addition, the literature on paired assessment has also been concerned about test score validity. In terms of content validity, Taylor (2003) states that one of the major strengths of the paired format of speaking assessment is its authenticity in terms of test content and the elicited interaction. Moreover, Galaczi (2008) approaches issues regarding construct validity by focusing on the interactive communication criterion of the Cambridge ESOL FCE. After examining 30 pairs of test takers through conversation analysis, Galaczi found that there were three patterns of interaction: collaborative interaction, parallel interaction and asymmetric interaction (see table 2.1). The first interaction

was characterized as being highly mutual and equal. Thus, there was self- or other-initiated topic expansion as well as a balanced quantity of talk. The second type of interaction was described as representing low mutuality (expansions of self-initiated topics only) but high equality (balanced quantity of talk). The third type of interaction was defined as having moderate mutuality and low equality. This means there was almost no expansion of topics and the quantity of talk was unbalanced. The analyses of various conversations demonstrated that the test scores in the interactive communication criterion represented the ability of test takers to interact. This supported the arguments of the construct validity of the scores in such tests.

Table 2.1 Structure of conversations according to the pattern of interaction (adapted from Galaczi, 2008)

Pattern of interaction	Description
Collaborative	A: Topic initiation B: **Topic extension** A: **Topic extension** OR A: Topic initiation + **Topic extension** B: **Topic extension** + Topic initiation A: **Topic extension**
Parallel	A: Topic initiation + Topic extension B: **Minimal acknowledgement** + Topic initiation + Topic extension A: **Minimal acknowledgement** + Topic initiation + Topic extension
Asymmetric	A: Topic initiation and topic extension **B: Minimal acknowledgement** A: Topic extension **B: Minimal acknowledgement** A: Topic initiation

2.3.2 Overview of the Cambridge ESOL FCE

This section will give a brief overview of the speaking paper of the Cambridge ESOL FCE, as the present study used tasks adapted from this test. Two testers participate in this test, one who plays the role of the interlocutor

and the other of the examiner, and two or sometimes three candidates are involved in the examination process. The interlocutor gives an overall score and guides some parts of the test, while the examiner does not participate in the conversations and just observes and gives analytic scores to each candidate.

This speaking test is composed of four parts, which last approximately 14 to 15 minutes. The first part requires the interlocutor and each of the candidates to engage in a conversation about general topics, such as work, leisure time, and future plans. In the second part candidates are asked to hold a long turn describing photographs and answering a question about them. They must also answer one question from their partners as well as listen to their partners and ask them one question. The third part is a two-way collaborative task that uses picture prompts to elicit speech. Candidates have to engage in a meaningful conversation discussing their opinions about seven pictures, which are related to a certain topic. They are then asked to choose the best two options out of the seven. The fourth part constitutes a three-way discussion led by the interlocutor who asks questions of each candidate and encourages them to talk in depth about a certain topic related to the previous part.

2.3.3 The Interlocutor's Effect in Paired Speaking Tests

Although several studies have provided evidence in support of the paired format of speaking assessment as a valid and authentic way to measure speaking ability, there are still some questions that need to be addressed. Foot

(1999) and Fulcher (2003), among other studies, have raised concerns about some of the interlocutor's characteristics that may have an impact upon the partner's performance. Some of the reasons for skepticism include acquaintance between paired candidates, their personality (i.e., introverts or extroverts), L1 background, gender, and proficiency. The following studies have tried to address some of the issues that may be a threat to the validity and fairness of the scores for this format of speaking assessment.

Following second language acquisition studies (e.g. Plough & Gass, 1993) on the effect of familiarity between students in conversation, several studies have examined the factor of acquaintance in paired assessments. O'Sullivan (2002) examined whether acquaintance was a factor that affected the final rating of test-takers in a paired speaking assessment. The study analyzed data from 24 Japanese university students who were assessed twice: once paired with a friend and a second time with a stranger. The results of a repeated-measures ANOVA indicated that acquaintanceship is a statistically significant factor that affects the scores in a paired speaking assessment. Test-takers within the context of this study performed considerably better when being paired with a friend than with a stranger. This study also demonstrated that the interlocutor acquaintance effect is complex since it also depends on other factors such as gender and the cultural context. In other words, the results indicated that there is an interaction between different variables. In this case, gender seemed to have a greater effect on the performance and accuracy when the pairing was with a stranger. Nevertheless, the conclusion was that the interlocutor's gender effect was "more difficult to explain" since it also

seemed to be related to sociocultural norms. As stated in Brown and McNamara (2004) gender is an *unpredictable factor* in test processes and test outcomes.

Norton (2005) reported similar results after a qualitative analysis of a sample of discourses in Cambridge speaking tests such as the Certificate of Advanced English (henceforth, CAE), and the FCE. A total of 20 candidates were examined: 10 FCE and 10 CAE candidates. The results of the qualitative analysis of the test takers' performances showed that there were effects on the particular pairing of test takers. The study found that in terms of familiarity effect, candidates who were paired with friends performed better than candidates paired with strangers, since they appeared to be more relaxed. In regards to gender, Norton reports that when female Japanese test-takers are paired with males from any other nationality, they tend to assume a supporting role.

Despite solid claims made by both studies about the positive impact of acquaintanceship on ratings, a closer analysis of the data uncovers problems with the generalizability of the conclusions. Fulcher (2003) revealed after a reanalysis of O'Sullivan's (2002) data that the effect size of the acquaintanceship on the test scores is moderate, accounting for only 24 percent of cases where the acquaintanceship is a determining factor of better performance. Since the sample size was small, Fulcher argued that the conclusions are not generalizable in other contexts. Likewise, the findings in Norton (2005) were disputed in Lazaraton (2006) in terms of the small sample size and the lack of quantitative proof to support the claims of the effect of

acquaintanceship on scores. Overall, the study seemed to be based on a few individualized cases (i.e., a pair between a Japanese and Danish candidate) failing to provide a more general view of the problem.

Furthermore, the interlocutor's personality has often been claimed to be a factor that may affect a candidate's speaking performance. In this respect, various studies (Berry, 1995, 1997, 1998 cited in Fulcher, 2003) have focused on the speaking performance of introvert and extrovert students in paired assessments. The studies indicate that both personality types performed better when paired with students with similar personalities. The introverts, however, appeared to be more affected by their partners' extroversion. When the test-takers' performances on paired assessments and one-to-one interviews were compared, both introverts and extroverts turned out to score higher in the former format.

Other studies have investigated the personality effect in oral group tests. Bonk and Van Moere (2004) examined the relationship between the *shyness and outgoingness* and speaking performance on an oral group test of 322 groups, each consisting of three or four test-takers. The results indicated that in terms of personality, shyness was a factor that negatively affected the final scores. More outgoing students tended to score better than introverted ones.

In addition to Bonk and Van Moere's study, Ockey (2009) investigated the test-taker assertiveness effect on group oral tests. The study included 225 Japanese university students. After administering a NEO Personality Inventory test which examines personality characteristics such as neuroticism, extraversion, openness, conscientiousness, and agreeableness, students were

grouped according to levels of assertiveness and non-assertiveness. An assertive test-taker was assessed in three different environments: a group of only assertive students; mixed assertive and non-assertive students; and one of only non-assertive students. The same type of grouping was formed when assessing non-assertive test-takers. The results showed that while assertive test-takers scored significantly higher when assessed in groups of majority non-assertive test-takers, the non-assertive test-takers showed no significant differences in scores in all three environments.

Another study, Nakatsuhara (2011), explored the interaction between factors such as personality, proficiency, and number of participants in a group oral test. The study involved 269 Japanese students who took a test in groups of three or four students. Following proficiency tests as well as personality questionnaires, all participants were tested on three tasks. The results demonstrated that the extraversion-levels had a greater effect on the speaking performance in groups of four than groups of three. In terms of the proficiency level, it had an effect on both types of group. However the influence was greater in groups of three than in groups of four. As a result, this study provides further evidence that the test-takers' characteristics, as well as those of their interlocutor, impact speaking performance.

2.3.4 Interlocutor's Proficiency Effect in Paired Speaking Tests

The potential effect of interlocutor proficiency has been examined in several studies through different approaches. The aforementioned study by

Norton, for example, observed the interlocutor's proficiency through the analysis of real data collected from performances of the FCE and CAE. When examining the received scores and analyzing the discourse of the conversations between candidates, Norton noted that candidates with lower linguistic ability appeared to score higher when paired with candidates with higher linguistic ability. Additionally, several features of interaction (i.e., appropriation of linguistic structures and lexical items) were observed in this type of pairing as well. In other words, candidates with lower language proficiency seemed to benefit from being paired with a partner with higher language proficiency. The mean utterance length was calculated together with the total percentage of words and total percentage of turns for each pair. Apart from the limitations pointed out in Lazaraton (2006) mentioned above, namely the small sample size as well as the lack of quantitative data, another drawback of her study is that other factors that might have affected the speaking performance of participants were not controlled.

In contrast, Nakatsuhara (2004) examined the interlocutor proficiency effect on the discourse of 24 test-takers in paired speaking assessments using a more experimentally-based approach. Participants were divided into high- and low- proficiency groups and were asked to complete tasks similar to Part 3 of the CAE speaking test. The conversations between different dyads were analyzed for discourse features outlined in Young and Milanovic (1992), being interactional contingency, goal orientation, and quantitative dominance. Contrary to Norton (2005), the results showed no statistically significant differences between different dyad types regarding these features of discourse.

A further qualitative analysis demonstrated that even though there were some differences in the discourse between higher- and lower-proficiency test-takers when paired together, the interlocutor's proficiency level had little effect on the discourse. The researcher suggested that in mixed-proficiency pairs, the conversations achieved a balance through supportive behavior (i.e., accommodative behavior) that test-takers showed to their peers.

A recent study by Bennet (2011) explored the perceptions and the effects of the interlocutor's proficiency on the scores of paired speaking assessments in a group of 1 Chinese and 11 Italian learners of English. Unlike the aforementioned study by Nakatsuhara, in this study all participants were required to take two tests. Bennet divided test-takers into 5 dyads consisting of same-proficiency level students, and 7 dyads consisting of different-level students. With the use of pre and post-test questionnaires, perceptions about the factors that are likely to affect test-takers' speaking performance were gathered. The analysis of the pre-test questionnaires showed that while most of the participants thought their scores would be affected by the proficiency of their partners, in the post-test questionnaires they thought differently. No evidence was found which indicated that the interlocutor's level of linguistic ability affected the test-takers' speaking performances. The biggest limitation of this study was the small sample size. The analysis of the data showed that 3 out of the 12 candidates showed variation according to whom they were paired with. Even though the researcher provides possible reasons, beyond the interlocutor proficiency to justify the variation of scores, drawing conclusions from 9 cases lacks generalizablity.

Csépes (2002) examined the same factor in the data of a larger sample size. A total of 120 Hungarian secondary school students was recruited to participate in a study designed to investigate whether the interlocutor proficiency effect had an effect on the ratings of test-takers when being assessed in pairs. Thirty students, referred to as 'core students' were paired with three different language proficiency level students, namely 'top students' (higher-proficiency level than that of the core students), 'bottom students' (lower-proficiency level) and 'middle students' (same-proficiency level). Thus, the 30 participants were assessed three times on distinct but parallel oral tests. The analysis of the results showed that there was no statistically significant difference in ratings when assessed under all three conditions. Participants were not affected negatively or positively by their partner's language proficiency.

In another empirical study, Iwashita (1998) examined the interlocutor proficiency effect on scores and discourse in paired speaking assessments. Twenty learners of Japanese, from high (n=10) and low (n=10) language proficiency levels, were tested twice, once with a partner of the same proficiency and once with a partner of different proficiency. Each time, they were asked to complete three different tasks which included two one-way tasks and one two-way task. The task types differed in that, in the one-way tasks, the test-takers held all the information and delivered it to their partners, while in the two-way task, both participants were given part of the information that needed to be delivered. Performances were scored by two experienced raters. Analyses of the scores indicated that both high- and low-

proficiency level test-takers scored higher when being paired with high-proficiency partners. Likewise, when examining the amount of talk, which was measured in c-units and turns, both high- and low-proficiency students talked more when being paired with high-proficiency partners. However, there were individual differences between participants; not all test-takers followed the same trend. Another important finding was that the amount of talk did not necessarily correlate with the scores.

Although Iwashita (1998) was one of the earliest studies to shed light into the interlocutor proficiency effect in a control setting, the conclusions were mostly drawn from raw scores. Accordingly, Davis (2009) further expanded the research on this topic through an inferential statistics approach by analyzing data using multi-faceted Rasch analysis. The collection of data used a similar method to that of Iwashita's. The study collected data from 20 Chinese first-year university students. Students belonged to two different majors: English and Software majors. Participants from the English major had a higher proficiency level than those from the Software major. Pairings were randomly assigned so that each student took the test once with a partner from the same-proficiency group and once with a partner from a different-proficiency group. In contrast to Iwashita (1998), the results indicated that the interlocutor's proficiency had no statistically significant effect on the scores of the test-takers. Thus, the proficiency level of the interlocutor had almost no impact upon the speaking performances of the participants. There were, however, individual differences; that is, the interlocutor effect influenced some participants but not others. The researcher explains that this variability

might be due to other variables in the testing process. As for the amount of words, low-proficiency students talked significantly more when paired with high-proficiency students, but did not necessarily score higher because of it. No meaningful differences in the amount of talk were found among high-proficiency test-takers.

Overall, these previous research studies have provided valuable insights into the effect of the interlocutor on the test-takers' speaking performance. Whilst some studies focused on the interactions between the interviewers and test-takers, others concentrated on the interactions between two or more test-takers. There are still, however, many more questions that remain unanswered, particularly regarding the interlocutor proficiency effect on speaking performance at the analytic score levels and the moderating effect of gender.

CHAPTER III

METHOD

This chapter deals with the methodology used to collect and analyze the data. It will begin with the description of the participants, raters, and examiners, followed by instruments and methods of data analysis. The collection of data was done in two different stages: Stage 1, which involved a non-interactive test (henceforth, NI Test), and Stage 2, which involved two paired tests.

3.1 Participants

Thirty-five Korean EFL learners from Seoul National University (SNU) participated in the initial part of the present study. All participants were native speakers of Korean with little (i.e., less than one year) or no experience living in English-speaking countries. Their areas of specialization varied from humanities, education, and law, to social sciences and engineering. At the time of the data collection, their ages ranged from 18 to 29, with more than 75% belonging to the age range of 20 to 25. All personal information was collected through a questionnaire (Appendix A).

In order to pair students for the second stage of the experiment, participants were first recruited according to their gender as well as their self-reported scores of the Test of English Proficiency developed by Seoul National University (henceforth, TEPS). In terms of gender, among the 35 participants, 17 were females and 18 were males. In addition, participants

were further subcategorized into groups of higher and lower proficiency. According to the Language Education Institute, scores of 850 or above are equivalent to a level of 1 or 1+, while scores of 600 to 750, to a level of 2 or 2+. These scores are equivalent to TOEFL scores of 111 or above and 86 to 103 (TEPS, 2009). Students within the former score range were classified as the higher proficiency group and students in the latter range as the lower proficiency group. The final result was four groups: 11 female students with higher proficiency (FH), 6 female students with lower proficiency (FL), 10 male students with higher proficiency (MH), and 8 male students with lower proficiency (ML). Table 3.1 summarizes the demographic data of the participants for the first part of the experiment.

Table 3.1. Demographic and TEPS score data of the Korean EFL participants

Subjects			TEPS Score	
Gender	Proficiency	N	Mean	SD
Female	Higher	11	909.72	30.03
	Lower	6	696.17	41.19
Male	Higher	10	899.10	35.26
	Lower	8	692.50	41.83
Total		35	820.43	110.37

TEPS scores served as an initial classification criterion to divide the students into different proficiency groups. All students in the university are obligated to take this test before admission and as a requisite for graduation. Therefore all participants in this study were able to report scores from the same standardized test. In addition, TEPS assesses communicative language

skills through four components, listening, grammar, vocabulary, and reading, which can provide a rough estimate of each student's English proficiency. Nevertheless, since this test lacks a speaking component, all students were additionally tested with a speaking test that consisted of three individual tasks that could measure their speaking abilities (see below for more detailed explanation).

According to the results of the NI test, the number of participants in each proficiency group was modified in order to reflect their speaking ability scores. Figure 3.1 indicates some incidences where TEPS scores did not directly correlate with the individual test scores. These cases were especially present in the female group. Several studies (Bachman & Palmer, 1982; Carroll, 1983; Bachman et al., 1995; Sasaki, 1996; Shin, 2005) have demonstrated that language ability is multicomponential. Following this line of reasoning, recent studies (Sawaki, Stricker, & Oranje, 2008; Powers, 2010) have presented ample evidence to suggest that different components such as reading, listening, speaking, and writing, though strongly correlated, measure different language skills. The same can be applied to TEPS scores, which focus more on receptive skills (reading and listening) rather than productive skills (speaking and writing). Students who scored high on the standardized test did not always score high on the non-interactive oral test, since the latter assessed oral productive skills as opposed to only receptive skills.

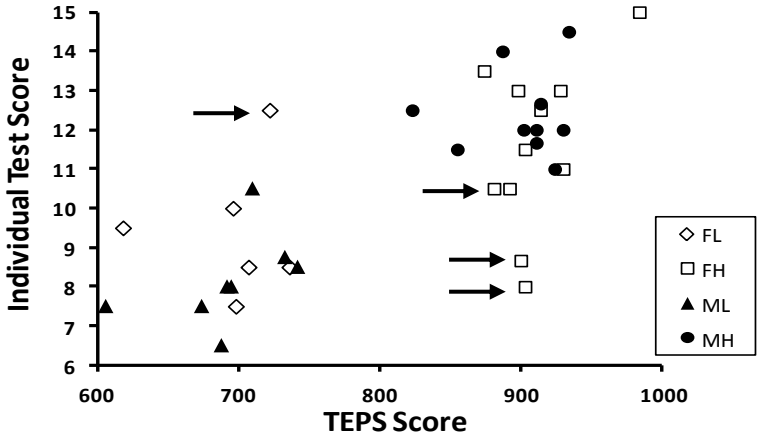


Figure 3.1. Scatterplot showing the relationship between TEPS scores and the non-interactive speaking scores.

Furthermore, of 35 participants, 3 (2 males and 1 female) dropped out of the study which reduced the initial subject pool (n = 35) to 32 participants (16 males and 16 females). According to the score in the NI test, these were classified into 3 proficiency groups: high, intermediate, and low. Although all participants took the paired tests during Stage 2, this study will mostly focus on the high and low proficiency groups. Data from the intermediate group was excluded from most of the analyses since the focus of the present study was to examine the behavior of the high- and low-proficiency test-takers. See Table 3.2 for data describing test-takers who participated in the second stage of the experiment.

Table 3.2 Data from the final pool of participants

			Non-interactive Test Score[a]	
		Number	Mean	SD
Female	Lower	6	8.53	0.84
	Intermediate	4	10.88	0.48
	Higher	6	13.25	0.94
	Total	16	10.89	2.24[b]
Male	Lower	6	7.67	0.68
	Intermediate	4	11.17	0.53
	Higher	6	13.06	1.06
	Total	16	10.52	2.51[c]

[a] The SD between the mean scores of female lower and higher was of 3.34
[b] The SD between the mean scores of male lower and higher was of 3.73

3.2 Raters

Two native English speakers from North America rated the speaking performances of all participants on both the individual task (Stage 1) and the paired assessment task (Stage 2). Both raters had extensive experiences in the fields of ESL/EFL assessment, having worked at TEPS for about 4 years writing items as well as scoring some components of TEPS and the Test of Oral Proficiency (TOP) also developed by Seoul National University.

Since the scoring rubrics (described below) for non-interactive and paired tests were created specifically for this study, the raters were trained by the researcher on the specific criteria of each type of rubric. A holistic rating scale was used to score the speech samples for the NI-Test in Stage 1. The rubric was first given to raters to get them familiarized with the different descriptors. Then, the raters received several practice audio files of real speech samples in testing situations taken from the TOEIC speaking sample test (YBM, n.d.) and a previous pilot test done by the researcher. The

researcher met with the raters to discuss, the rating scale as well as the audio files, and to resolve possible difficulties when interpreting the rubric. Furthermore, unlike in Stage 1, an analytic rating scale was used in Stage 2. Raters also received training for this rubric in a manner similar to the way they did for the holistic scale. They were first given audio files converted from videos of real speech samples from the Cambridge FCE speaking sample test (Cambridge ESOL, 2009) and from the previously mentioned pilot test to practice rating with the analytic scale. Both raters' scoring for each of the evaluation criterion was discussed until they reached a consensus about how to interpret all descriptors and scales.

A third rater was brought in for cases where raters disagreed in their scoring of individual tasks by two points or more. For the discrepant scores in the non-interactive test, all three scores (i.e., from all three raters) were averaged (Bejar, 1985). However, in the case of the paired tests, the third rater's scores were used to adjudicate between the two inconsistent scores. To be more precise, the two closest pair of scores were selected and averaged.

3.3 Examiners

Two Korean examiners participated in this study to guide students through different tasks. One was the researcher and the other was a graduate student from the same department. In the present study, the examiners played roles that were very different from those in previous studies. They were in charge of reading the instructions aloud, handing out the prompt cards, and audio and video recording the speaking performances but did not participate in the

interactions between test-takers. They used a script (Appendix B) to avoid any differences in the examiner's wording and behavior during the administration of the tests. For reasons of practicality, the researcher deemed it necessary to have two examiners instead of one administering the tests. This was particularly important in the case of Stage 2, when test-takers had to take two different paired speaking tests with different partners. In order to have participants perform one task immediately after the other, two examiners were needed, each in two different rooms (more detailed explanation is found below in section 3.5).

3.4 Instruments

This section deals with various instruments used for data collection. Two of the instruments used in this study were the different speaking tests: the non-interactive (NI) Test and paired test. The former resembled a semi-direct test used in the TOEFL iBT and the latter was a direct test. All tasks were adapted from the second and third part of the Cambridge FCE speaking paper (or section). The tasks were slightly modified according to feedback given by several university students who tried out the tests beforehand. Their comments helped to make instructions and descriptions more comprehensible and to pick the best topics that would elicit a sustained conversation. In addition, the speaking performances of each participant were scored using different scoring rubrics: a holistic rubric for the non-interactive test and an analytic rubric for the paired test.

3.4.1 Non-Interactive Speaking Test

The NI Test consisted of three different tasks which were adapted from Part 2 of a retired Cambridge FCE speaking section. It consisted of two tasks based on picture description and one task asking an opinion related to both pictures (Appendix C). Examinees had 20 seconds of planning time for both pictures and 30 seconds to describe each picture separately. Then they were given 10 seconds to prepare a response to the opinion question and 30 seconds to answer it. Hence, the total amount of time allotted to this test was 2 minutes. All speech samples were recorded using the timer record function of the Audacity sound editing software Version 2.0.0 (Audacity, 2012). Each audio file was given random numbers for the purposes of scoring.

Since the non-interactive test scores were used as a criterion to evaluate the score from the subsequent paired tests, it was necessary to make the tasks of the two tests parallel. Accordingly, both tests assessed skills that involved describing pictures as well as expressing their opinions. Moreover, there were two major advantages of having three separate tasks. First, the elicited speech samples were more homogenous and thus more suitable for comparisons. Since the response time in all three tasks was predetermined, the length of the responses of all participants was the same. Second, in terms of scoring, an aggregate of three separate scores was considered more accurate than one single score from the reliability perspective.

3.4.2 Paired Speaking Test

Two paired speaking tasks with different topics (i.e., Topic A: coffee shop and Topic B: film club) were used to assess the speaking performance of

participants under two conditions: when paired with someone with the same proficiency (henceforth, PSP condition) and with different proficiency (henceforth, PDP condition). The dyads were carefully designed so that each participant would be tested twice without repeating the same topic. The test was adapted from the speaking section (paper) of Part 3 of the Cambridge FCE. The tasks consisted of 7 pictures related to a certain topic (see Appendix D). Examinees were paired and asked to engage in a conversation with their partners for about 5 minutes. They were first asked to briefly talk about the 7 pictures, and then jointly decide the best two options.

Due to the lack of familiarity with this format of speaking assessment, the participants were given instructions about the task along with a sample test in Korean one week before they were tested (see Appendix E). In addition, the instructions for each test task were provided on prompt cards which were read aloud by the examiner to make sure the test-takers understood the instructions (see Appendix B). Additionally, a timer clock was displayed in front of the examinees so that they could keep track of time while completing the tasks.

All test-takers signed consent forms prior to the audio- and video-recordings of their speaking performances (see Appendix F). Each audio file was then given a random number before being handed over to the raters for scoring. Raters received the files in randomized order so that speech samples would not appear consecutively.

3.4.3 Scoring Rubric

3.4.3.1 Holistic rubric

As mentioned earlier, a holistic scale was used for the initial part of the

experiment to classify participants into different proficiency groups (see Appendix G). Since Stage 1 had the sole purpose of dividing students into proficiency groups, it was unnecessary to have a separate criterion for each language skill. The scoring rubric was adapted from other scales used for assessing speaking skills, such as the ones used for the TOEIC Speaking and IELTS tests. It also took into consideration the descriptors used in Fulcher (2003)'s fluency scale. It consisted of a five-point holistic scale measuring grammar, vocabulary, pronunciation, fluency, and cohesion. Each of the three tasks in the non-interactive test was assigned a maximum score of 5, and thus the highest aggregated score possible was 15.

3.4.3.2 Analytic rubric

For the second part of the experiment, where participants were tested in pairs, an analytic scale was designed specifically for the present study (see Appendix H). It was adapted from the Cambridge FCE rating scale (Cambridge ESOL, n.d.) as well 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. Moreover, the types of interaction observed in Galaczi (2008, see Table 2.1 above) were also taken into consideration for the scoring of the discourse management criterion.

The analytic scale was finer-grained than the holistic scale, consisting of five bands for five separate criteria: grammar (G), vocabulary (V), pronunciation (P), fluency (F), and discourse management (DM). It was necessary to use 5 analytic criteria, because it enabled us to discern any

possible differences in the speaking performances of participants depending on who they were paired with. Moreover, the final composite scores combined all five criteria, and thus the maximum score was 25.

3.5 Data Collection Procedure

As illustrated in figure 3.2, the data collection was carried out in two stages, which took a total of four days (two days per stage). During Stage 1, students were assessed individually. According to these scores, they were classified into three groups: high-, intermediate- and low-proficiency. This allowed test-takers to be assigned into same- and different-proficiency dyads.

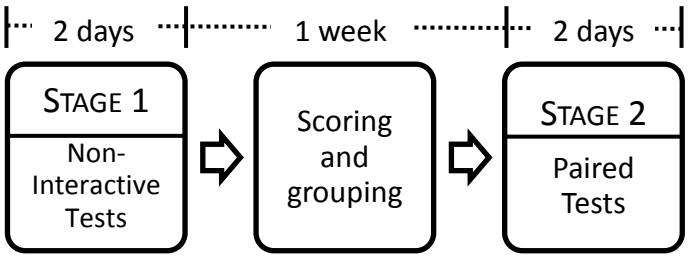


Figure 3.2 – Illustration of the different stages of the experiment

Stage 2 took place one week after Stage 1. During this week, raters scored the audio files for the individual performances. Afterwards, the participants were arranged into groups and scheduled to come a second time to be assessed in pairs. All 32 participants were tested twice, with a different partner each time. They were grouped into 8 groups of 4, each group consisting of 2 higher-proficiency and 2 lower-proficiency students, as shown in Figure 3.2. As mentioned before, there were 8 participants (4 females and 4

males) that belonged to the intermediate-proficiency level; they were grouped together according to their gender. In order to control possible gender effects, there were no gender-mixed groups.

Figure 3.3 describes the test administration procedure for one group of participants. Four students – 2 dyads – were tested at the same time in two different classrooms. When both pairs were finished taking the first paired speaking task, one student from each dyad exchanged places and took the second paired speaking task.

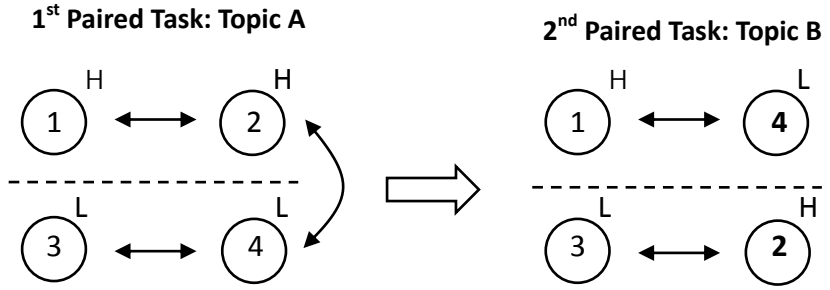


Figure 3.3 – Description of the paired test administration procedure
Notes. H = High-proficiency; L = Low-proficiency

The previous illustration represents a case where the participants were first paired with students with the same proficiency and then with students of different proficiency. However, in order to avoid practice effect, the order of the pairings was counterbalanced. In other words, half of the number of dyads did the first paired task with partners from the same-proficiency group and the other half did it with partners from the different-proficiency group. Likewise, the number of dyads that started with Topic A (i.e., coffee shop) first was counterbalanced with dyads that started with Topic B (i.e., film club). Table

3.3 shows the number of dyads according to the different pairings and topics.

Table 3.3 Number of dyads per topic and proficiency in the first and second tests

	First Paired Task		Second Paired Task	
Same-proficiency partner	Topic A	4	Topic A	2
	Topic B	2	Topic B	4
Different-proficiency partner	Topic A	2	Topic A	4
	Topic B	4	Topic B	2
Intermediate-proficiency group	Topic A	2	Topic A	2
	Topic B	2	Topic B	2
Total		16		16

Notes. Topic A = coffee shop; Topic B = film club

Random numbers were given to all participants during the initial part of the experiment for scoring purposes. Similarly, for the audio files obtained for the paired tasks, other random numbers were assigned, and each participant was labeled A or B depending on who began with the initial turn.

3.6 Analysis of the Data

All scores submitted by the two raters for the non-interactive and paired tests were keyed into a Microsoft Excel 2007 spreadsheet and then transferred to SPSS Version 17.0 for Windows (SPSS, 2008). This software was used to obtain the descriptive statistics for the composite and analytic scores, correlations among task, test scores and other criterion measures, and

reliability coefficients for ratings. Finally, two-way analysis of variance (ANOVA) was conducted using proficiency and gender groups as independent variables.

To examine the relationship between the non-interactive test, the paired tests and TEPS as a standardized measurement of language proficiency, the Spearman rank-order correlation coefficients were computed. The raw score distribution showed a slight deviation from the normal curve; therefore Pearson product-moment correlations were not the most appropriate for the analysis of this study (Bachman, 2004). However, they were reported in the tables together with the Spearman rank-order correlations for comparison. In addition, these estimates aided in examining the relationship of scores from the non-interactive and paired tasks. At the same time, these measures served as evidence to investigate the construct validity of the scores from the speaking tests developed for the purposes of this study.

Furthermore, the Spearman rank-order correlations, rater agreement indices, and kappa coefficients were calculated to examine the inter-rater reliability. As mentioned earlier, when there was a discrepancy between raters by two bands (or score points) or more in the five band scale, a third rater was acted as an adjudicator. The agreement indices were also computed for the adjudicated scores.

As a preliminary investigation before the ANOVA tests, the raw scores from the two paired assessment tasks were compared by examining the percentage of change of scores between the PSP and PDP conditions. In addition, a 2×2 (Proficiency×Pairing) split-plot ANOVA (also known as

mixed-design analysis of variance) with an alpha level of 0.5 was conducted to test the statistical significance of the differences between the two paired tests conditions. Hence, to answer the first research question, the speaking performances from the high- and low-proficiency groups, when paired with same proficiency partners and when paired with different-proficiency partners, were compared to observe whether there was an interlocutor proficiency effect in the total mean scores. Moreover, the second research question was explored by analyzing the scores of each criterion separately. Finally, the same analyses were repeated separately for each gender group in order to answer the third research question.

The split-plot ANOVA was considered the most appropriate for analyzing the collected data since the experimental design involved participants from different proficiency levels being assessed across two different conditions. In other words, the study dealt with the comparison of two groups on a single variable. The repeated measures ANOVA has the advantage of determining whether there are significant effects within groups across different conditions. In addition, it also provides information regarding the potential interaction effects between the independent variables used in this study.

CHAPTER IV

RESULTS

The following chapter presents the results of the statistical analysis of raw scores, psychometric analysis of items, and analysis of variance. This section first examines the descriptive statistics of the NI Test and Paired Test. Furthermore, it examines the construct and concurrent validities of the different tests scores, as well as other criteria. Then it proceeds to assess the raters' reliability through correlation coefficients. Next, it examines the interlocutor proficiency effect in the paired tests through a split-plot analysis of variance. Finally, it presents the raters' post-rating feedback regarding the paired tests, rubric, and test-takers' performances.

4.1 Descriptive Statistics for NI and Paired Tests

Table 4.1 presents the descriptive statistics of the performances in all three speaking tests for the lower- and higher-proficiency groups. The intermediate group was excluded from the analysis, since this group only worked with same proficiency partners. Thus, the interlocutor proficiency effect could not be observed in this group (see Appendix I for complete data).

Table 4.1 Raw scores of the non-interactive and paired tests for the low- and high-proficiency groups

Tests	Low-proficiency			High-proficiency			Low- and High-proficiencies Combined		
	N	Mean	SD	N	Mean	SD	N	Mean	SD
TEPS	12	728.08	87.71	12	890.92	65.36	24	809.50	112.42
NI Test Total	12	8.10	.86	12	13.15	.96	24	10.63	2.73
Task 1		2.75	.37		4.33	.49		3.54	.91
Task 2		2.61	.39		4.43	.42		3.52	1.01
Task 3		2.74	.35		4.39	.39		3.56	.92
PSP-G	12	2.75	.66	12	4.38	.53	24	3.56	1.01
PSP-V		3.04	.45		4.54	.45		3.79	.89
PSP-P		2.92	.85		4.42	.29		3.67	.99
PSP-F		2.79	.54		4.58	.56		3.69	1.06
PSP-DM		3.38	.61		4.63	.43		4.00	.82
PSP Composite		14.88	2.65		22.54	1.89		18.71	4.52
PDP-G	12	2.71	.45	12	4.50	.56	24	3.60	1.04
PDP-V		3.13	.53		4.54	.58		3.83	.91
PDP-P		3.17	.86		4.25	.66		3.71	.93
PDP-F		2.79	.33		4.54	.45		3.67	.97
PDP-DM		3.38	.57		4.71	.45		4.04	.85
PDP Composite		15.17	2.51		22.54	2.27		18.85	4.43

Notes. NI = non-interactive test; PSP = paired with same-proficiency; PDP = paired with different-proficiency; G = grammar; V = vocabulary; P = pronunciation; F = fluency; DM = discourse marker

As mentioned in the previous chapter, participants were classified into different proficiency groups according to their scores from the NI test and TEPS. There was a difference of 5.06 in terms of the aggregated score differences in the NI test between the low- and high-proficiency groups. This indicates that there was a wide gap between both proficiency groups.

In terms of the speaking performance in the paired tests, the low proficiency group showed a slight increase in the composite scores when being paired with high-proficiency partners. Nevertheless, for the high-proficiency group, the composite scores were the same in both pairing situations. Moreover, even when each analytic score was analyzed separately, both proficiency groups demonstrated no substantial differences in scores.

Table 4.2 reports the differences in the analytic and composite scores between the PSP and the PDP conditions. Together with the raw scores differences, it also presents the percentages of change between the scores in both test situations.

Table 4.2 Differences between the composite scores of the PSP and PDP conditions for low- and high-proficiency test-takers

	Low-proficiency		High-proficiency	
	Difference	%	Difference	%
G	0.04	1%	-0.12	-3%
V	-0.09	-3%	---	0%
P	-0.25	-9%	0.17	4%
F	---	0%	0.04	1%
DM	---	0%	-0.08	-2%
Total	-0.08	-6%	---	0%

For the low-proficiency group, the composite score had an increase of 6% when participants were paired with high-proficiency partners. On the other

hand, as reported above, the high-proficiency group experienced no change in scores. Therefore, the low-proficiency group seemed benefit slightly from being paired with high-proficiency partners while the high proficiency group was not affected. Furthermore, the low-proficiency group showed an increase of 3% and 9% in analytic scores such as vocabulary and pronunciation, respectively, while fluency and discourse management scores remained unchanged. Only the grammar scores were minimally lower, with 1% decrease. In the case of the high-proficiency group, grammar and discourse management scores increased by 3% and 2%, respectively, while the vocabulary scores did not change. In this group there were two criteria – pronunciation and fluency – for which scores decreased by 4% and 1%. Accordingly, the results demonstrated that there was a difference between the low- and high-proficiency groups in terms of the increase and reduction of composite and analytic scores.

Moreover, the paired test scores were further analyzed separately for each gender group. The results are presented in Table 4.3.

Table 4.3 Descriptive statistics for paired test scores classified by proficiency and gender

	Low-proficiency				High-proficiency			
	Female		Male		Female		Male	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
PSP-G	3.08	.80	2.42	.20	4.58	.38	4.17	.61
PSP-V	3.25	.52	2.83	.26	4.75	.27	4.33	.52
PSP-P	3.42	.86	2.42	.49	4.50	.32	4.33	.258
PSP-F	3.08	.49	2.50	.45	4.75	.42	4.42	.67
PSP-DM	3.50	.71	3.25	.52	4.92	.20	4. 33	.41
PSP Total	16.33	2.94	13.42	1.28	23.50	.95	21.58	2.18
PDP-G	2.83	.61	2.58	.20	4.67	.26	4.33	.75
PDP-V	3.25	.52	3.00	.26	4.67	.41	4.42	.74
PDP-P	3.58	.97	2.75	.49	4.33	.68	4.12	.68
PDP-F	2.92	.38	2.67	.45	4.67	.26	4.42	.59
PDP-DM	3.50	.71	3.25	.52	4.83	.26	4.58	.59
PDP Total	16.08	3.02	14.25	1.64	23.17	1.29	21.92	2.96

Overall, the female participants scored higher than male participants did on both paired tests, not only in the composite measure, but also in analytic criteria. There was a tendency for female participants to get a lower score when paired with different-proficiency partners, even though the differences were minimal. In contrast, male participants tended to score higher in the same situation.

Table 4.4 displays the score differences between the PSP and PDP conditions when gender groups were analyzed separately. As with the scores presented above for the high- and low-proficiency groups, this table also shows the differences in terms of percentages.

Table 4.4 Percentile differences between the analytic and composite scores of the PSP and PDP conditions for test-takers by gender groups

	Low-proficiency				High-proficiency			
	Female		Male		Female		Male	
	Difference	%	Difference	%	Difference	%	Difference	%
G	0.25	8%	-0.16	-7%	-0.09	-2%	-0.16	-4%
V	---	0%	-0.17	-6%	0.08	2%	-0.09	-2%
P	-0.16	-5%	-0.33	-14%	0.17	4%	0.21	5%
F	0.16	5%	-0.17	-7%	0.08	2%	---	0%
DM	---	0%	---	0%	0.09	2%	-0.25	-6%
Total	0.25	2%	-0.83	-6%	0.33	1%	-0.34	-2%

The results indicate that the differences in total mean scores between the paired tests were minimal, showing decreases or increases ranging from 1% to 6%. Regarding both female high-proficiency (FH) and low-proficiency (FL) groups; they had a reduction in the composite scores of 1% and 2%, respectively, when working with different proficiency partners. In contrast, the male high- (MH) and low- (ML) proficiency groups had an increase in the composite scores of 2% and 6% when being paired with different-proficiency partners.

As displayed in Table 4.4, scores were also analyzed separately for each of the analytic criteria (for figures refer to Appendix J). Like previous results, there were no significant differences of scores between the two pairing conditions. Nonetheless, the distinct patterns in behavior between the female and male groups shed some light into participants' behavior in the paired speaking tests.

The percentages of differences show that the female groups (i.e, FL and FH) appeared to either score less or remain constant in almost all of the criteria when being paired with different-proficiency partners. For the FL

group, only pronunciation showed a 5% increase in scores in the PDP condition. The scores on vocabulary and discourse management did not change, while the grammar scores decreased by 8%. Similarly, the FH group marginally improved in only one criterion, grammar, with an increase of 2% when assessed with the lower proficiency partners. The rest of the criteria indicated a reduction ranging 2% to 4%. Therefore, this group appears to have performed slightly better when they were paired with high-proficiency partners than with low-proficiency ones.

The male group, on the other hand, benefited slightly more than the female group did when its members were paired with different-proficiency partners. In the case of the ML group, there was an increase in score for all of the criteria with the exception of discourse management which remained unchanged. The pronunciation scores increased the most (14%) followed by grammar, fluency, and vocabulary (7%, 7%, and 6%, respectively). The MH group also indicated an increase in scores for almost all of the criteria except for fluency, which did not change, and pronunciation, which decreased 5%. For this group, discourse management was the criterion for which the scores increased the most (6%) followed by grammar (4%) and vocabulary (2%).

4.2 Reliability Measures

Prior to analyzing the interlocutor proficiency effect, it was deemed necessary to provide evidence of concurrent and construct validities of the speaking scores as well as raters reliability for all test scores. Thus, this section reports the correlations among test scores, test tasks, and a third criterion, TEPS. It

then proceeds to examine the inter- and intra-rater reliability of test scores.

4.2.1. Correlations among Test Scores and Other Criterion Measures

Spearman rank-order correlation coefficients were calculated in order to examine the relationships among the non-interactive and paired tests and other criterion measures. The assumption is that high coefficients between item scores and test scores and between different test scores would indicate a close relationship among items and tests. Therefore, this also serves as evidence of item discrimination and the concurrent validity of the tests designed for the present study. The results of a two-tailed Spearman rank-order correlation among the NI Test tasks and total scores, the paired tests (PSP and PDP conditions), and TEPS are reported in Table 4.5.

Table 4.5 Spearman rank-order correlation coefficients between tests

Test Scores	NI- Test				Paired Test				TEPS
	T1	T2	T3	Total	Café	Film	PSP	PDP	
NI-T1	1								
NI-T2	.89 (.90)	1							
NI-T3	.81 (.86)	.80 (.89)	1						
NI-Total	.96 (.96)	.95 (.97)	.88 (.95)	1					
Café	.82 (.83)	.78 (.81)	.78 (.83)	.81 (.85)	1				
Film	.82 (.83)	.79 (.83)	.80 (.86)	.82 (.87)	.90 (.94)	1			
PSP	.86 (.85)	.79 (.83)	.78 (.84)	.83 (.88)	.97 (.97)	.94 (.97)	1		
PDP	.80 (.83)	.77 (.81)	.81 (.85)	.81 (.86)	.94 (.97)	.94 (.97)	.90 (.94)	1	
TEPS	.78 (.78)	.69 (.74)	.69 (.71)	.77 (.76)	.75 (.76)	.79 (.81)	.77 (.78)	.74 (.77)	1

Notes: All significant at .01 level (2-tailed)
NI-T1/T2/T3= Non-Interactive Test Tasks 1, 2, 3; NI-Total = aggregated scores
( ) = Pearson Product-Moment correlation coefficient

As shown in Table 4.5, high positive correlations were obtained between TEPS scores and NI-Test tasks and aggregated scores, and paired tests (Café and Film Tests; PSP and PDP conditions) with coefficients of .77, .75, .79, .77, and .71, respectively. As presented in the table, the scores of the three separate items in the NI Test were highly correlated among themselves, with a range of .80 to .89. In terms of the relationship between the NI Test and paired tests, the coefficients also showed a significantly high correlation ranging from .81 to .87.

4.2.2 Rater Reliability

The inter-rater reliability was assessed in several ways. First, the Spearman rank-order correlation coefficient was computed to examine the relationship between rater R and rater J in the NI test and paired tests. In terms of the NI test, since the scores were based on a holistic rubric, correlations were based on the scores of each rater for each task. For the paired tests, the correlations were calculated for the composite scores and the analytic scores (i.e., grammar, vocabulary, pronunciation, fluency, and discourse management). Second, the agreement indexes between raters were calculated to examine the percentage of perfect, adjacent, and non-adjacent scores between both raters. The final measure of inter-rater reliability was kappa coefficients. All these measures demonstrated the proportion of agreement between raters with regards to the scores given to all participants on each task in the case of the NI test and each analytic criterion in the case of paired tests.

Table 4.6 Correlation coefficients between raters in the non-interactive test

	Spearman Rho coefficients	Pearson Coefficients
Task 1	.70	.70
Task 2	.88	.86
Task 3	.69	.69
Total Score	.87	.86

Notes: All significant at .01 level (2-tailed)

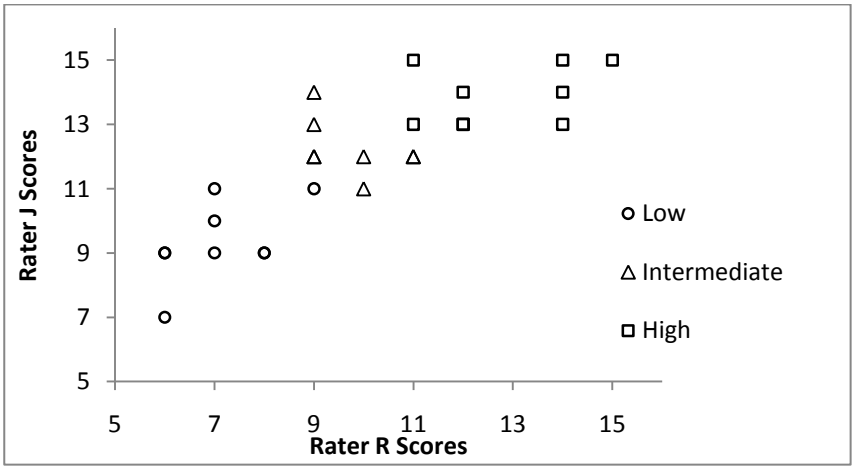


Figure 4.1 Scatterplot of the aggregated scores by both raters for the non-interactive test

As observed through the results of the Spearman rank-order correlations in Table 4.6, there was a positive correlation between raters who scored the NI Test tasks and aggregated scores, r_s ranging from .69 to .88, $n = 32$, $p = .000$. In other words, there was significant agreement between both raters in terms of the decisions made when they classified all participants into low-, intermediate-, and high-proficiency groups (see Figure 4.1). In addition, the close relationship between the scores of both raters also suggests a high degree

of intra-rater reliability.[1]

Table 4.7 lists the results of the Spearman rank-order correlation coefficients between rater R and rater J in the paired tests according to each rating criterion. The values only include the analysis of the low- and high-proficiency groups, since the intermediate groups were paired with same-proficiency partners both times.

Table 4.7 Spearman rank-order correlation coefficients between raters in the paired tests according to each criterion

	PSP Condition		PDP Condition	
	Spearman Rho	Pearson Correlation	Spearman Rho	Pearson Correlation
Grammar	.83	.81	.80	.76
Vocabulary	.66	.66	.83	.79
Pronunciation	.88	.90	.76	.78
Fluency	.83	.83	.73	.70
Discourse Management	.75	.74	.68	.67
Total Score	.86	.90	.92	.90

Notes: All significant at .01 level (2-tailed);

Overall, there was a significantly high correlation between both raters in the composite scores from the PSP condition (r_s = .86, n = 24, p = .000) as well as the PDP condition (r_s = .92, n = 24, p =.000). Furthermore, regarding each criterion, the correlation values ranged from .66 to .88 in the PSP

[1] The assumption is similar to that of the test-retest reliability approach, where test-takers take the same test twice. Then a correlation between two sets of scores is computed to examine stable scores over time. A positive correlation indicates the stability of the scores (Bachman, 1990: 181-82). Likewise, the high correlation found in the two sets of scoring from rater R and rater J indicate that there is stability of scores across the raters.

condition and .68 to .83 in the PDP condition. Interestingly, when comparing the PSP and PDP conditions, the scores from the latter condition show lower correlation coefficients on almost all the criteria except for vocabulary. Thus, there was a greater degree of disagreement in scores for the participants paired with different-proficiency partners. This could be an indication that performance in this test condition was more difficult to score, specifically the discourse management criterion.

Table 4.8 presents agreement indices as well as kappa coefficients which serve as further evidence of inter-rater reliability in the scoring of the NI test tasks and paired tests.

Table 4.8 Score agreement rates and Kappa coefficients between raters in the non-interactive test

NI Test	Perfect Agreement		Adjacent Agreement		Perfect + Adjacent		Non-adjacent Agreement		Kappa
	N	Rate	N	Rate	N	Rate	N	Rate	
T 1	16 (14)[a]	.50 (.58)	14 (9)	.44 (.38)	30 (23)	.94 (.96)	2 (1)	.63 (.04)	.23 (.42)
T 2	8 (6)	.25 (.25)	23 (17)	.72 (.71)	31 (23)	(.97) (.96)	1 (1)	.31 (.04)	.01 (.09)
T 3	13 (10)	.41 (.42)	16 (13)	.50 (.54)	29 (23)	.91 (.96)	3 (1)	.09 (.04)	.18 (.25)

Notes: T1, T2, T3 = Tasks 1, 2, 3

[a]( ) = results of calculations excluding the intermediate proficiency group.
Adjacent Agreement = the agreement rate of scores differing by +/- 1 point.

The results indicate that the rate of perfect + adjacent agreement ranges from .91 to .96 (i.e., 90.6% to 95.8%) for all three tasks. The adjacent agreement indexes represent the proportion of cases which differ by only 1 band. Therefore, the scores considered truly discrepant were those that differed by 2 bands or more. These score discrepancy rates ranged from .03

to .09, which demonstrates that the two raters had a relatively high agreement rate overall.

Given that the agreement rates are subject to chance (Powers, 2000), the Kappa coefficients were also computed to provide a better assessment of the inter-rater reliability. The results indicate that, in the case of Tasks 1 and 3, there was a fair to slight level of agreement (Landis & Koch, 1977, cited in Sim & Wright, 2005). In contrast, Task 2 showed poor agreement between the raters' scores. As these scores were used to assign participants into different proficiency groups for the paired tests, it was crucial for the ratings from the two raters to be consistent. Therefore, as mentioned above, the truly discrepant ratings were adjudicated by a third rater.

A point worth noting is that when the data from the intermediate-proficiency group was excluded, the Kappa coefficients increased. This means that the raters mostly disagreed when scoring intermediate-proficiency participants. Additionally, the scores from the NI test were only used to determine each student's speaking proficiency level, and the results show that the raters certainly agreed on this classification. One noteworthy pattern was that high correlations were obtained between the raters' scores but only a slight level of agreement was achieved. This indicates that while the two raters rank-ordered the participants' speaking performances in a similar way, they exercised different levels of severity in scoring them.

Similar to the NI test, the agreement indices in the scoring of analytic criteria in the paired tests were also high. Table 4.10 reports the agreement rates together with the kappa coefficients.

Table 4.9 Score agreement rates and kappa coefficient between raters in the paired test

Paired Test	SC	Perfect Agreement		Adjacent Agreement		Perfect + Adjacent		Non-adjacent Agreement		Kappa
		N	Rate	N	Rate	N	Rate	N	Rate	
PSP	G	11	.46	13	.54	24	1.00	-	-	.31
	V	11	.46	12	.50	23	.96	1	.04	.23 (.29)
	P	12	.50	12	.50	24	1.00	-	-	.36
	F	18	.75	5	.21	23	.96	1	.04	.66 (.71)
	DM	13	.54	10	.42	23	.96	1	.04	.39 (.42)
PDP	G	8	.33	15	.63	23	.96	1	.04	.19 (.23)
	V	14	.58	10	.42	24	1.00	-	-	.42
	P	10	.42	14	.58	24	1.00	-	-	.24 (.33)
	F	9	.38	14	.58	23	.96	1	.04	.17 (.22)
	DM	12	.50	10	.42	22	.92	2	.08	.34

Notes: ( ) = Kappa coefficient when considering the third rater's adjudication
Adjacent Agreement = the agreement rate of scores differing by +/- 1 point.
SC = Scoring criteria

As observed in the table above, the rate of perfect agreement + adjacent agreement in the PSP condition ranged from .95 to 1.00 (i.e., 95.8% to 100%). While grammar and pronunciation had no cases of non-adjacent agreement, vocabulary, fluency, and discourse management had only one case. Moreover, the values of kappa coefficients were considerably high ranging from .23 to .66, which indicates a fair to substantial strength in agreement between raters. When the scores of the adjudicator were taken into account, higher coefficients were obtained, which ranged from .29 to .77. Fluency was the criterion with the highest kappa value, having the highest rate of perfect agreement (.75) in the PSP condition. In contrast, the same criterion showed the lowest rate of agreement (.17) in the PDP condition.

The rate of perfect + adjacent agreement in the PDP condition was

slightly lower than that found in the PSP condition, ranging from .92 to 1.00 (i.e., 91.7% to 100%). In this test condition, vocabulary and pronunciation had no true discrepancies in scores whereas grammar, fluency, and discourse management had up to 2 cases of non-adjacent agreement. In terms of kappa coefficients, the values ranged from .17 to .44. Unlike the PSP condition, raters' scores of vocabulary proved to have the highest kappa coefficient and fluency the lowest. With the sole exception of vocabulary, all other criteria had lower kappa coefficients than the PSP condition did. This echoes the results of the Spearman correlation analysis suggesting that raters tended to agree slightly less under the PDP condition.

4.3 Analysis of Variance (ANOVA)

The previous section demonstrated high reliability of the speaking scores used in the present study. These results constitute the pillars for the following inferential statistics, which attempt to answer the research questions. The split-plot ANOVAs were conducted to observe the interaction between proficiency groups and the pairing factor.

4.3.1 Interlocutor Proficiency Effect on the Composite Scores

Table 4.10 and Figure 4.2 show the results of the split-plot ANOVA of the composite scores for high- and low-proficiency groups across two different pairing conditions. It reports the score differences between the two proficiency groups (i.e., high and low) as well as the differences within the groups under different pairing conditions (i.e., when paired with the same and different-

proficiency partners).

Table 4.10 Split-plot ANOVA for the composite scores of the paired tests

Source Variation	*SS*	*df*	*MS*	*F*	*p*
Between Groups					
Proficiency	678.755	1	678.755	70.746	.000
Error	211.073	22	9.594		
Within Groups					
Pairing	.255	1	.255	.180	.675
Pairing*Proficiency	.255	1	.255	.180	.675
Error	31.115	22	1.414		

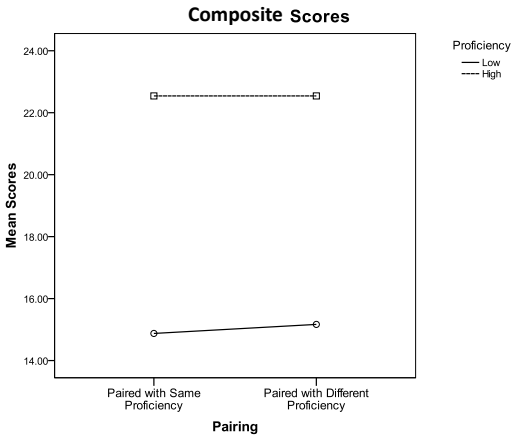


Figure 4.2 Composite scores for the PSP and PDP conditions

The results show that there was a statistically significant between-groups difference (high- and low-proficiency groups) in terms of their speaking performance in both paired tests, $F(1,22) = 70.746$, $p < .05$. These results were expected, since the participants were recruited according to their different proficiency levels. The participants with high speaking proficiency consistently scored higher than the participants with low speaking proficiency on all three speaking tests.

When the results were analyzed separately for each proficiency group,

there was no statistically significant difference on the composite scores of the participants when they were paired with same- and different-proficiency partners, $F(1,22) = .180$, $p > .05$. In other words, the interlocutor's proficiency had no effect on the speaking performance of participants. The test-takers' performances were nearly identical in both the PSP and PDP conditions. For the high-proficiency group, the composite scores from the PSP and PDP conditions were identical. As previously reported in Table 4.1, the composite score for high-proficiency participants was 22.54 for both paired tests. Similarly, the low-proficiency group composite scores in both paired tests were nearly identical, with a difference of only .29 on a scale of 5 to 25. The ANOVA results showed that this difference was not statistically significant.

Regarding the interaction between the different pairings and proficiency, there was no statistically significant result, $F(1,22) = .180$, $p > .05$. This means that the variances between the high- and low-proficiency groups stayed almost the same in both paired tests. The nearly parallel lines observed in Figure 4.2 indicate that the interlocutor's proficiency had virtually no effect in both high and low groups.

4.3.2 Interlocutor Proficiency Effect on the Analytic Scores

Although the results of the split-plot ANOVA in the composite scores showed no statistically significant results, there was a need to further explore the analytic scores separately. This section also reports the results of ANOVA for each of the analytic criteria: grammar, vocabulary, pronunciation, fluency, and discourse management.

Table 4.11 Split-plot ANOVA for the grammar mean scores of the paired tests

Source Variation	*SS*	*df*	*MS*	*F*	*p*
Between Groups					
Proficiency	35.021	1	35.021	63.434	.000
Error	12.146	22	.552		
Within Groups					
Pairing	.021	1	.021	.328	.572
Pairing*Proficiency	.083	1	.083	1.313	.264
Error	1.396	22	.063		

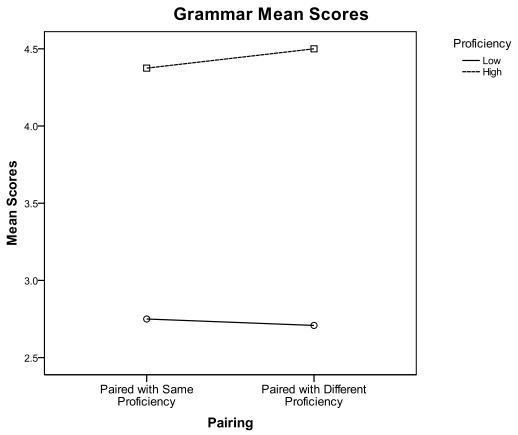


Figure 4.3 Grammar mean scores of the paired tests for high- and low-proficiency groups

When analyzing the grammar scores separately, the split-plot ANOVA indicated that the proficiency groups were significantly different in their grammar proficiency $F(1,22) = 63.434$, $p < .05$ (See Table 4.11 and Figure 4.3). The high-proficiency test-takers scored considerably higher than the low-proficiency test-takers did in this criterion. As for the interlocutor's proficiency, the results demonstrated that it had no statistically significant effect on participants' grammar, $F(1,22) = .328$, $p > .05$. The interaction between pairing × proficiency also showed no significant results, $F(1,22) = 1.313$, $p > .05$. Even though there was a slight increase and decrease

in the grammar scores for the high- and low-proficiency groups respectively when being paired with different-proficiency partners, these differences were not statistically significant.

Table 4.12 Split-plot ANOVA for the vocabulary mean scores of the paired tests

Source Variation	*SS*	*df*	*MS*	*F*	*p*
Between Groups					
Proficiency	25.521	1	25.521	65.732	.000
Error	8.542	22	.388		
Within Groups					
Pairing	.021	1	.021	.169	.685
Pairing*Proficiency	.021	1	.021	.169	.685
Error	2.708	22	.123		

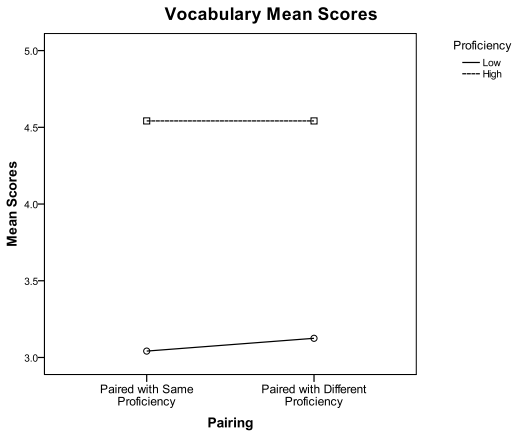


Figure 4.4 Vocabulary mean scores of the paired tests for high- and low-proficiency groups

Similar results were obtained for the vocabulary scores. As shown in Table 4.12 and Figure 4.4, the vocabulary scores for the high- and low-proficiency groups were substantially different, $F(1,22) = 65.732$, $p < .05$. However, when examining the within groups effects, as with grammar,

vocabulary was not affected by the interlocutor's proficiency, $F(1,22) = .169$, $p > .05$. Furthermore, the interaction between the pairing and proficiency was not significant $F(1,22) = .169$, $p > .05$. The high-proficiency group remained consistent in both paired conditions while the low-proficiency group did marginally better when paired with higher-proficiency students.

Table 4.13 Split-plot ANOVA for the pronunciation mean scores of the paired tests

Source Variation	*SS*	*Df*	*MS*	*F*	*p*
Between Groups					
Proficiency	20.021	1	20.021	24.413	.000
Error	18.042	22	.820		
Within Groups					
Pairing	.021	1	.021	.124	.729
Pairing*Proficiency	.521	1	.521	3.090	.093
Error	3.708	22	.169		

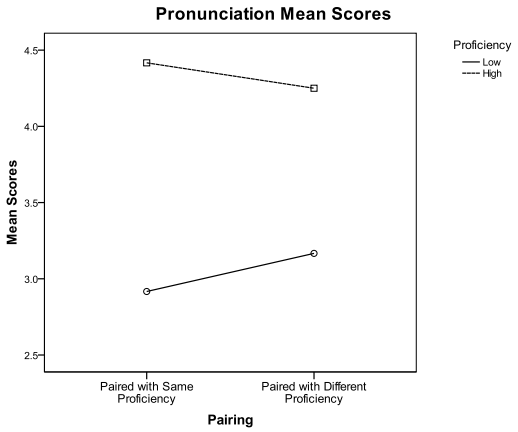


Figure 4.5 Pronunciation mean scores of the paired tests for high- and low-proficiency groups

In the same manner, both proficiency groups differed significantly in terms of pronunciation, $F(1,22) = 24.413$, $p < .05$ (see Table 4.13 and Figure 4.5). However, despite the minimal differences between the PSP and PDP scores, the interlocutor proficiency had no statistically significant effect on the

pronunciation scores for both proficiency groups, $F(1,22) = .124$, $p > .05$. The pairing × proficiency interaction effect was not statistically significant: $F(1,22) = 3.090$, $p > .05$. The variation between the pronunciation scores for both proficiency groups decreased slightly during the paired test with different-proficiency interlocutors. Even though the results of the split-plot ANOVA indicated that this change in variation was not significant, this criterion showed the highest interaction effect compared to the other analytic criteria.

Table 4.14 Split-plot ANOVA for the fluency mean scores of the paired tests

Source Variation	*SS*	*df*	*MS*	*F*	*P*
Between Groups					
Proficiency	37.630	1	37.630	118.443	.000
Error	6.990	22	.318		
Within Groups					
Pairing	.005	1	.005	.037	.850
Pairing*Proficiency	.005	1	.005	.037	.850
Error	3.115	22	.142		

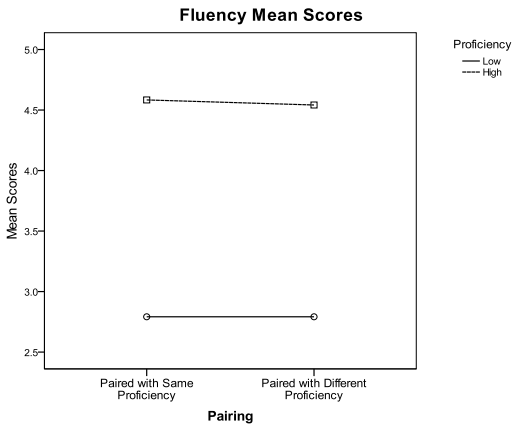


Figure 4.5 Fluency scores of the paired tests for high- and low-proficiency groups

As with all of the other criteria, fluency showed similar results. The between groups effect was statistically significant, $F(1,22) = 118.443$, $p < .05$ (See Table 4.14 and Figure 4.6), indicating a considerable difference in fluency between high- and low-proficiency groups. On the other hand, the interlocutor's proficiency, as in the other criteria, had no significant effects on fluency, $F(1,22) = .037$, $p > .05$. In fact, the fluency scores of the high-proficiency group had an increase of only .04 when paired with low-proficiency partners, and the scores of the low-proficiency group showed no changes.

Table 4.15 Split-plot ANOVA for the discourse management mean scores of the paired tests

Source Variation	*SS*	*df*	*MS*	*F*	*p*
Between Groups					
Proficiency	20.021	1	20.021	45.369	.000
Error	9.708	22	.441		
Within Groups					
Pairing	.021	1	.021	.208	.653
Pairing*Proficiency	.021	1	.021	.208	.653
Error	2.208	22	.100		

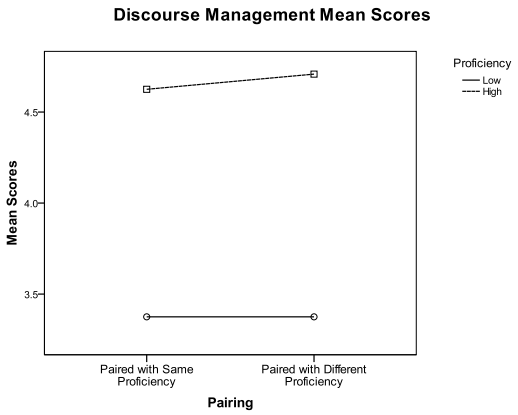


Figure 4.7 Discourse management mean scores of the paired tests for high- and low-proficiency groups

The results for the final criterion, discourse management, concurred with the previous findings. While there was a statistically significant difference between the high- and low-proficiency groups on this evaluation criterion, $F(1,22) = 45.369$, $p < .05$ (see Table 4.15 and Figure 4.7), the within groups effect was not significant, $F(1,22) = .208$, $p > .05$. In other words, pairing conditions did not affect participants' discourse management. Moreover, the interaction between the different pairings and proficiencies was also non-significant, $F(1,22) = .208$, $p > .05$.

Overall, the results show that interlocutor proficiency had no significant effects on the speaking performance of participants in any of the analytic scores. Even though there was a substantial difference in the scores of high- and low-proficiency groups, the test-takers were not affected by the interlocutor's proficiency. Both high- and low-proficiency test-takers performed similarly across the same and different dyad conditions.

4.3.3 Interlocutor Proficiency Effect by Gender Groups

In order to examine whether gender moderates the interlocutor proficiency effect, the scores in both paired speaking tests were also analyzed separately for each gender group. This section uses the same method of analysis as the previous section did by analyzing the total mean scores and then each of the different criteria.

Even when each gender group was examined separately, the interlocutor's proficiency turned out to have no statistically significant effects on the composite scores of female as well as male participants. These results are presented in Table 4.16 and Figure 4.8.

Table 4.16 Gender comparison of the split-plot ANOVA for the composite scores

	Female	Male
Proficiency	$F(1,10) = 33.649$, $p = .000$	$F(1,10) = 52.471$, $p = .000$
Pairing	$F(1,10) = .450$, $p = .518$	$F(1,10) = 1.178$, $p = .303$
Pairing*Proficiency	$F(1,10) = .009$, $p = .926$	$F(1,10) = .216$, $p = .652$

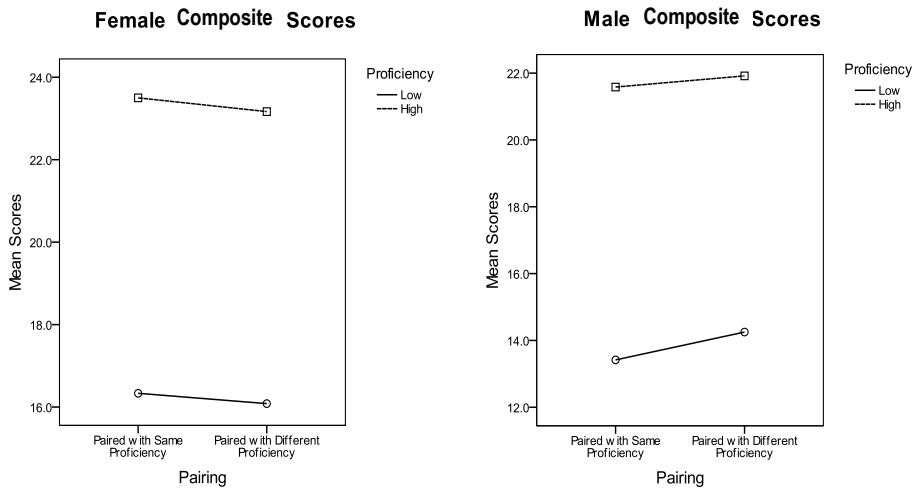


Figure 4.8 Composite scores for the PSP and PDP conditions according to gender

In both gender groups, there was a statistically significant difference between the performances of high- and low-proficiency students with $F(1,10) = 33.649$, $p < .05$ for the female group and $F(1,10) = 52.471$, $p < .05$ for male test-takers. However, there were no statistically significant results in terms of pairing and the pairing × proficiency interaction. This means that the interlocutors' proficiency had no significant effect in the speaking performances of both female and male groups and that the variation in scores between high- and low-proficiency participants within these groups remained almost identical under the different pairing conditions.

Split-plot ANOVAs for analytic scores were also computed for each

gender group. These results were consistent with those for the composite scores. Despite clear differences in language proficiency between high and low proficiency groups, there was no interlocutor proficiency effect for any of the criteria when each gender group was examined separately (see Table 4.17).

Table 4.17 Gender comparison of the split-plot ANOVA for analytic scores by gender groups

		Female	Male
G	Proficiency	$F(1,10) = 30.075$ $p = .000$	$F(1,10) = 40.833$ $p = .000$
	Pairing	$F(1,10) = .769$ $p = .401$	$F(1,10) = 2.857$ $p = .122$
	Pairing×Proficiency	$F(1,10) = 3.077$ $p = .110$	$F(1,10) = .000$ $p = 1.000$
V	Proficiency	$F(1,10) = 38.043$ $p = .000$	$F(1,10) = 33.108$ $p = .000$
	Pairing	$F(1,10) = .172$ $p = .687$	$F(1,10) = .464$ $p = .511$
	Pairing×Proficiency	$F(1,10) = .172$ $p = .687$	$F(1,10) = .052$ $p = .825$
P	Proficiency	$F(1,10) = 5.216$ $p = .045$	$F(1,10) = 16.667$, $p = .000$
	Pairing	$F(1,10) = .000$ $p = 1.000$	$F(1,10) = 1.053$, $p = .329$
	Pairing×Proficiency	$F(1,10) = .200$ $p = .664$	$F(1,10) = 1.800$ $p = .209$
F	Proficiency	$F(1,10) = 129.309$ $p = .000$	$F(1,10) = 49.388$ $p = .000$
	Pairing	$F(1,10) = .529$ $p = .484$	$F(1,10) = .357$ $p = .563$
	Pairing×Proficiency	$F(1,10) = .059$ $p = .813$	$F(1,10) = .357$ $p = .563$
DM	Proficiency	$F(1,10) = 22.975$ $p = .001$	$F(1,10) = 26.118$ $p = .000$
	Pairing	$F(1,10) = .172$ $p = .687$	$F(1,10) = .652$ $p = .438$
	Pairing×Proficiency	$F(1,10) = .172$ $p = .687$	$F(1,10) = .652$ $p = .438$

Consequently, the results of this study indicate that the interlocutor's proficiency had little effect upon the speaking performances of both male and

female test-takers. The results showed no significant difference in any of the five analytic criteria between the scores from the PSP and PDP conditions.

4.4 Raters' Post-Rating Feedback

In order to supplement the quantitative findings, the raters provided post-rating feedback by responding to a short questionnaire about paired tests, scoring rubric, and test-takers' performance (refer to Appendix L). In terms of the overall opinion about the paired test, the raters responded positively to this format of speaking assessment. One of the raters commented that she found the elicited speech sample to be more authentic than those of other tests she had graded before. Furthermore, both raters agreed on the fact that this type of test elicited more speech among test-takers. They stated: "There seemed to be more speaking among test-takers generally than occurs in computer tests" and "The prompt was good – clear, and generated a lot of language."

Although both raters had no prior experience rating this type of oral test, they found the rating of the paired oral test to be generally easy to conduct. Nevertheless, when asked to mention the most difficult part in the process of rating the speech samples, they pointed out some factors that might have caused some difficulty. On the one hand, one of the raters indicated that the discourse management criterion was the hardest since it was "the most subjective." On the other hand, the other rater explained that rating test-takers from similar language proficiency levels was hard given that their proficiency was too similar and that at times it was hard to differentiate the test-takers' voices. Moreover, she suggested that the scores of 2 or 3 were the hardest to discern.

Question 3 asked raters to rank the analytic criteria according to their level of difficulty in scoring. Both raters disagreed on which criterion was the hardest. While one rater considered discourse management to be the most difficult, the other rater regarded pronunciation to be the hardest. As for discourse management, the rater mentioned that it was hard because of its level of subjectivity. She stated that it was "hard to separate proficiency from someone's conversational style." In respect to pronunciation, the other rater explained that it was not easy to award low scores since she had grown accustomed to Korean EFL learners' pronunciation. Therefore, it was fairly easy to understand participants. The difficulty was due to the fact that she had to think as other North Americans would do.

In regards to vocabulary, both raters agreed that it was the second hardest criterion to rate. Their comments included: "Vocabulary errors are easy to spot, but it is a little harder to rate the range"; "Vocabulary was a little tricky, because there are no clear guidelines as to what "good" or "bad" vocabulary is." Additionally, the raters also agreed on grammar being one of the easiest criteria to rate. They commented that grammatical errors were easily identified.

Finally, when asked to comment about the test-takers' speaking performances, the raters indicated that there were generally no noticeable differences between the different pairings (i.e., high-high, low-low, or high-low). One of the raters, however, mentioned that in some dyads of high-low proficiency test-takers, the high-proficiency test-takers asked more questions to their low-proficiency partners in order to elicit more speech.

CHAPTER V

DISCUSSION

This chapter discusses the major findings with respect to the interlocutor proficiency effect presented in the previous chapter in terms of: 1) composite scores; 2) each analytic score; and 3) composite and analytic scores when the effect was analyzed separately for gender groups. It also examines some of the raters' answers to the short questionnaire in relation to the rater reliability coefficients.

5.1 Interlocutor Proficiency Effect in the Composite Scores

The statistical results indicated that the interlocutor proficiency effect had no significant influence on the composite scores of participants. When the interlocutor proficiency effect was analyzed through raw scores, there was a slight difference in patterns between high- and low-proficiency groups. Whilst the high proficiency group showed no difference in mean scores when paired with same- or different-proficiency partners, the low-proficiency group showed a slight increase in mean scores of 6% when being paired with high-proficiency partners. However, the results of the split-plot ANOVA demonstrated that this variation was not statistically significant. This echoes the findings by Davis (2009), who indicated no statistically significant

differences between scores.

In regards to the descriptive statistics, the findings seem to agree partially with the results found in Iwashita (1998), which indicated that both high- and low-proficiency test-takers benefited slightly from working with different-proficiency peers, showing an increase of 15% and 50% respectively. Like Iwashita (1998), the present study also indicated that the low proficiency group showed a slight increase when being paired with high proficiency test-takers. However, the high proficiency group presented no changes under both conditions. Furthermore, the trend of the variations differs from that reported in Davis (2009). In the latter study, high-proficiency test-takers presented a minimal increase in mean scores of 3% and low-proficiency test-takers a minimal decrease in mean scores of 5%, when paired with different proficiency partners.

Similar to the findings reported by both Iwashita and Davis, the present study also revealed the presence of individual variations. When examining the differences in the composite scores of individual test-takers (Appendix K), all but one of the participants, from high- and low-proficiency groups indicated individual variations ranging from -35% to + 11%. This range is relatively smaller than the one reported in Davis' data (i.e., -50% to +21%). In addition, only 7 out of 24 test-takers showed a difference greater than 10%. Based on this range and number of test-takers, it appears that there was a considerably small amount of variation within this data. Moreover, these variations did not appear to show any patterns. Some participants within the high proficiency group scored slightly higher when paired with lower proficiency partners,

while others scored lower. The same was true for the lower proficiency group. As suggested by Davis (2009), these differences could be explained as a representation of "the overall variability present in the testing process as a whole" (p. 386).

5.2 Interlocutor Proficiency Effect in the Analytic Criteria

Another major finding in the current study was that interlocutor proficiency had no statistically significant effect on any of the analytic scores assigned for the rating criteria, namely grammar, vocabulary, pronunciation, fluency and discourse management. This is parallel to the findings in Davis (2009). Test-taker performance in all five criteria remained mostly consistent when being assessed with the same- or different-proficiency partners. The variation observed in the high proficiency group was minimal, ranging from -3% to 4%. Similarly, the low-proficiency group indicated a variation from -9% to 1%.

Interestingly, the criterion with the highest percentage of variation was pronunciation, with a reduction of 4% in the high-proficiency group and an improvement of 9% in the low-proficiency group. Although the differences proved to be non-significant, the patterns of these results can be explained by the Speech Accommodation Theory. This theory describes how speakers sometimes tend to adjust their speech depending on who their interlocutor is (Giles, 1977; Thakerar, Giles, & Cheshire, 1982; Beebe, 1988). A study by Beebe & Zuengler (1983) provided evidence in support of this theory by examining Chinese-Thai speakers who accommodated their speech in terms of

pronunciation according to who the interlocutor was. The results in the current study might therefore be an indication of the accommodation of high-proficiency test-takers to low-proficiency, and vice versa. This accommodation was minimally present in other rating criteria such as grammar and vocabulary. Thus, relative to pronunciation, it appears that test-takers found it harder to accommodate their speech in terms of grammar and vocabulary.

Since the data from the present study were not analyzed qualitatively, the study cannot directly contradict the findings by Norton (2005) who mentioned the possibility of lower language ability test-takers benefiting from being paired with higher level test-takers. Nevertheless, the results showed that within the context of this study, the scores failed to represent such discourse features as "appropriation" of syntactic and lexical structures. Even though there were some differences in the scores of grammar and vocabulary in the low-proficiency group, these variations were minimal, with a reduction of .04 and an increase of .09 on a scale of 5, respectively. They had no positive effect on scores for the participants at the low-proficiency level. An alternative reason could be that a wide range of grammatical and lexical structures was not a deciding factor in awarding scores. In other words, even though the descriptors in the grammar and vocabulary criteria include the usage of a wide range of structures, other factors such as appropriacy and accuracy could have had a greater weight on the scores. In fact, as observed in the raters' post-rating feedback, the raters indicated that measuring the range of vocabulary was very difficult.

In regards to the discourse management criterion, differences in the interlocutor's proficiency had no effect on the scores for this particular criterion. The scores of the participants from the low-proficiency group remained unchanged while the participants from the high proficiency group increased their scores by .08 accounting for only 2% of the variation. This is tantamount to Nakatsuhara (2004)'s findings which explained that the interlocutor's proficiency had little effect on the candidates' features of interaction. However, the results of the present study should be interpreted cautiously since the correlation coefficients between the scores of raters in discourse management showed some disagreement, especially in the PDP condition.

5.3 Interlocutor Proficiency Effect by Gender Groups

The analyses of the scores according to gender groups concurred with the results on the first two research questions. The interlocutor's proficiency had nearly no effect on both female and male candidates' overall speaking performance. The inferential statistics showed no statistically significant difference between the overall scores when paired with a similar- or with a different-proficiency partner in either gender groups.

In addition, the patterns of the variations of scores, while minimal, revealed interesting characteristics of female and male participants. First, the variation of the composite scores was different between female and male test-takers. It appears that as opposed to the female group, the male group benefited slightly from working with a different-proficiency partner. The low-

proficiency male group benefited a little more, with an increase of 6%, than did the high-proficiency male group, whose scores increased by only 2%. Accordingly, the male group seems to have followed the trend found in Iwashita (1998).

It is possible that in previous studies, the mixed gender pairs could have had an effect on the results. The gender variable should be explored even further due to its complexity (Brown & McNamara, 2004) and should be explored in association with other variables (Nakatsuhara, 2011).

The aforementioned trend in the scores of pronunciation also appeared in the female and male groups. Although there was no statistically significant difference overall for the raw scores, pronunciation was the only criterion that showed a similar pattern between gender groups. While the pronunciation scores of low-proficiency female and male test-takers increased when being paired with high-proficiency partners, the scores of high-proficiency female and male test-takers decreased when working with low-proficiency peers. The results were not statistically significant, but this trend seems to be in accordance with the speech accommodation theory explained in the previous section. Moreover, the interaction effect between pairing and proficiency in this criterion could reach a statistically significant level if the sample size is increased.

In terms of the other criteria, the results showed no statistically significant variation due to the interlocutor's proficiency level. However, analysis of the patterns of minimal variations, displayed a tendency for the female group to be negatively influenced when paired with different

proficiency partners whereas the male group seemed to gain an advantage in the same situation. This could be explained by degrees of *affective schemata*[1] (Bachman & Palmer, 1996). It is possible that female test-takers felt more self-conscious when working with different-proficiency partners than male test-takers did. For a more complete study, future research should also include questionnaires that could shed light on this particular matter.

5.4 Raters' Post-Rating Feedback

The raters' answers to the post-rating questionnaire provide valuable feedback on the overall paired test and scoring criteria, highlighting some of the most important aspects that should be improved for future studies. As indicated in the previous chapter, both raters found the paired test format to be advantageous for its ability to elicit a larger speech sample than the computer-based oral tests.

Although, they found the scoring rubric to be "straightforward," they indicated some difficulties when scoring certain analytic criteria. Interestingly, the criteria with the lowest correlation coefficient between raters concurred with those they considered most difficult to rate. According to the correlation coefficients, the raters' scores tended to disagree more in the vocabulary and discourse management criteria. In spite of rater training on the rubric, these results reveal that there is need for more specificity of the descriptors in the rubric. In terms of vocabulary, the range of vocabulary should be clearly

[1] Affective schemata is one of the components of Bachman & Palmer (1996)'s model of language use and performance on language tests. It attempts to explain the test-taker's emotional responses to the test and how this might positively or negatively affect her performance in the test.

specified. As for discourse management, more training could be necessary to reduce the amount of "subjectivity," which was the major concern of one of the raters. In addition, raters' perception of the test-takers' pronunciation is another issue that should be dealt with cautiously. Previous research revealed that the raters' characteristics, such as exposure to the test-takers' non-native English accents and training status, might influence the pronunciation ratings (Kang, 2009; Carey, Mannell, & Dunn, 2010).

Furthermore, both raters asserted that they did not notice any obvious differences between different types of pairings (i.e., high-high, low-low, and high-low). This is tantamount to the quantitative results, which showed that the interlocutor's proficiency had little effect on the test-taker speaking performance. However, since the raters completed the post-rating questionnaire several weeks after scoring the speech samples, they might have forgotten some of the minor differences. As mentioned in the previous chapter, one rater vaguely recalled that, in some dyads, the high-proficiency test-takers tended to ask more questions to their low-proficiency peers. For future studies, raters' feedback should be collected through verbal recalls and discussions, as in May (2011), in order to have more complete data on raters' perception.

CHAPTER VI

CONCLUSION

6.1 Conclusions and Implications

The results of this study indicated that the interlocutor's proficiency had no statistically significant effect on the test-takers' performance in the paired speaking tests. The test scores for the composite as well as all five analytic criteria showed little indication of differences in performance across the two different pairing conditions: (i.e., paired with same- and paired with different-proficiency partners). Contrary to studies that have stated the possibility of positive or negative impact on the scoring in paired speaking tests, this study found that the scores were not affected by the pairing condition at the statistically significant level.

Based on the findings of this study, one major implication is the use of this format of speaking assessment. Since the language proficiency level of interlocutors had no influence on test-taker scores, it appears that the paired format of speaking assessment could be a viable assessment tool for measuring learners' speaking ability in a valid, reliable, and fair way.

There are numerous benefits associated with this type of assessment not only in terms of practicality, but for authenticity as well. Given that test-takers interact with each other, this oral test assesses their interactive communication skills. Choi (2008) has emphasized the need to incorporate a speaking component in the Korean Scholastic Aptitude Test (KSAT) in order to encourage learners as well as instructors to focus on productive spoken skills.

The paired test could be considered feasible for assessing speaking abilities since they could bring positive washback in the Korean EFL education context. Learners would be motivated to practice speaking skills in order to score satisfactorily in these types of tests. Even though, more research is needed regarding matters of standardization, paired speaking assessment should be considered as a viable option for future English assessment in the Korean context.

6.2 Limitations and Future Studies

There were a number of limitations in terms of the methodology and analysis of data collected for the purposes of this study. Given that this study was based on a small sample size (n = 24), the findings should be interpreted with caution. Particularly for the use of inferential statistics a larger sample size would provide more accurate analysis of results. In addition, to achieve more generalizable results, mixed gender dyads should also be examined. Moreover, the paired speaking test consisted of one single task. This limits the analysis of data to the scores in the performance on this single task. The combination of several tasks could provide a more accurate measure of test-takers' speaking performances, as well as of score reliability. Furthermore, another limitation was that the perceptions of test-takers were not collected. Banerjee and Luoma (1997) stress the importance of qualitative validation techniques to investigate the test-takers perceptions of tasks and constructs.

Parallel to the previous studies, the analysis of results in this study also indicated the presence of individual variation. In other words, while some

participants scored slightly higher when being paired with different-proficiency partners, some scored lower in the same situation. This demonstrates how the research on interlocutor proficiency effect is "indirect and unpredictable, rather than simple and consistent" Davis (2009:388). Therefore, more empirical research needs to examine the interaction between different factors.

More specifically, qualitative analysis of the speech samples could provide a complete explanation of why some learners appear to perform differently depending on the different test conditions. As several qualitative studies of oral interviews (e.g. Brown, 2003; Ross, 2007) have shown, a misalignment between interlocutors could result in a difference in test-taker performance. The careful analysis of this difference could yield results that could help to improve the scoring rubrics by identifying the elements in speech that are currently not captured by the scoring descriptors. At the same time, finer-grained speaking scores could be provided to test-takers for a more diagnostically-oriented feedback.

In addition, this study found that rater score disagreement was more frequent in the PDP condition. Studies by May (2006, 2009, 2011) on raters' perceptions of the interaction in paired speaking tests pointed out that asymmetric dyads are more prone to disagreement in ratings. Moreover, McNamara and Lumley (1997) indicated that raters seemed to compensate if they perceive the candidates are having difficulties caused by variables other than their language ability, such as task and interlocutors. Hence, raters' perceptions as well as the rating criteria are crucial in the further investigation

of this topic. Even though the current study used a short questionnaire to examine the raters' opinions of the paired oral test, a more specific study of their perceptions in the process of rating the speech samples is needed. Such research will shed light on the most important factors that raters take into account when scoring the oral tests, providing valuable feedback for the scoring rubric and scales.

References

Audacity (version 2.0.0). (2012). Retrieved March, 14, 2012 from http://audacity.sourceforge.net/

Bachman, L.F. & Palmer A.S. (1982). The construct validation of some components of communicative proficiency. *TESOL Quaterly 16(4)*, 449-465.

Bachman, L.F. & Palmer A.S. (1996). *Language Testing in Practice*. Oxford: Oxford University Press.

Bachman, L.F., Davidson, F., Ryan, K. and Choi, I-C. (1995). *An investigation into the comparability of two tests of English as a foreign language*. The Cambridge-TOEFL comparability study. Cambridge: Cambridge University Press.

Bachman, L.F. (1990). *Fundamental Considerations in Language Testing*. Oxford: Oxford University Press.

Bachman, L.F. (2004). *Statistical Analyses for Language Assessment*. Cambridge: Cambridge University Press

Banerjee, J. & Luoma, S. (1997). Qualitative approaches to test validation. In Clapham, C. and Corson, D. (Eds.), *Encyclopedia of Language and Education, Vol. 7: Language Testing and Assessment*. Dordrecht: Kluwer Academic, 275-87.

Beebe, L. & Zuengler, J. (1983). Accommodation theory: An explanation for style shifting in second language dialects. In N. Wolfson and E. Judd (Eds.), *Sociolinguistics and Language Acquisition*. Newbury House Publishers, Rowley, MA, London, Tokyo.

Beebe, L. (1988). Five sociolinguistic approaches to second language acquisition. In L.M. Beebe (Ed.), *Issues in Second Language Acquisition: Multiple Perspectives*. New York: Newburry House, 43– 77.

Bejar, I. (1985). A preliminary study of raters for the test of spoken English. TOEFL Research Reports. Princeton, NJ: Educational Testing Service, 1 – 28.

Bennet, R. (2011). Is linguistic ability variation in paired oral language testing problematic? *ELT Journal,* 1-10

Berry, V. (1997). Ethical considerations when assessing oral proficiency in pairs. In Huhta, A., Kohonen, V., Kurki-Suonio, L. and Luoma, S. (Eds.), Current *Developments in Language Testing*. Jzväskzlä: Jzväskzlä University Press, 107-23.

Berry, V. (1998). *Personality and oral test score variability.* Paper presented at TESOL '98 Conference, Seattle, WA.

Berry, V. (1995). *A qualitative analysis of factors affecting learner performances in group oral tests.* Paper presented at the 17th Language Testing Research Colloquium, Long Beach, CA.

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.

Brown, A, & McNamara, T. (2004). 'The devil is in the detail': Researching gender issues in language assessment. *TESOL Quaterly, 38,* 524-538.

Cambridge ESOL. (n.d.). ALTE Can Do Statements: overall general ability. Retrieved October 16, 2011, from http://www.cambridgeesol.org/about/standards/can-do.html

Cambridge ESOL (2009). Cambridge ESOL Teacher Support. Retrieved October 15, 2011, from https://www.teachers.cambridgeesol.org/ts/teachingresources

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

Carroll, J. B. (1983). Psychometric theory and language testing. In J. W. Oller, Jr. (Ed.), *Issues in Language Testing Research.* Rowley, MA: Newbury House, 80 – 107.

Carey, Michael D., Mannell, Robert H., & Dunn, Peter K. (2010). Does a rater's familiarity with a candidate's pronunciation affect the rating in oral proficiency interviews? *Language Testing 28(2)*, 201–219.

Chapelle, C., Grabe, W. & Berns, M. (1997). *Communicative Language Proficiency: Definition and implications for TOEFL 2000.* TOEFL Monograph Series 10. Princeton, NJ: Educational Testing Service.

Choi, I-C. (2008). The impact of EFL testing on EFL education in Korea. *Language Testing, 25(1),* 39–62

Csépes, I. (2002). Is testing speaking in pairs disadvantageous for students? A

quantitative study of partner effects on oral test scores. *novELTy, 9(1)*, 22-45.

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.

Együd, G. & Glover, P. (2001). Oral testing in pairs – a secondary school perspective. *ELT Journal, 55*, 70-76.

Educational Testing Service. (1982). *Oral proficiency testing manual*. Princeton, NJ: Author.

ffrench, A. (1999). *Study of qualitative differences between CPE individual paired test formats* (Internal UCLES EFL report). Cambridge, UK: University of Cambridge Local Examinations Syndicate.

ffrench, A. (2003). The change process at the paper level. Paper 5, Speaking. In M. Milanovic & C, Weir (Eds.), Studies in Language Testing: Vol. 15. Continuity and innovation: Revising the Cambridge proficiency in English Examination 1913-2002. Cambridge: Cambridge University Press, 367 – 446.

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.

Fulcher, G. (1997). The testing of speaking in a second language. In C.

Clapham & D. Corson (Eds.), *Encyclopedia of Language Education. Vol. 7: Language Testing and Assessment*. Dordrecht Kluwer Academic Publishers, 75 – 85.

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

Giles, H. (1977). Social psychology and applied linguistics: towards an integrated approach. *ILT Review of Applied Linguisitcs*, 35, 27-42.

Iwashita, N. (1998). The validity of paired interview formal in oral performance assessment. *Melbourme Papers in Language Testing*, 5, 1-65.

Johnson, M. & Tyler, A. (1998). Re-analyzing the OPI: How much does it look like natural conversation? In R. Young and A.W. He (Eds.), *Talking and Testing: Discourse Approaches to the Assessment of Oral Proficiency*. Amsterdam and Philadelphia: John Benjamins, 27 – 51.

Kang, Okim. (2008). Ratings of L2 Oral Performance in English: Relative Impact of Rater. *Spaan Fellow Working Papers in Second or Foreign Language Assessment,* 6, 181–205.

Characteristics and Acoustic Measures of Accentedness

Lado, R. (1961). *Language Testing*. London: Longman

Landis, J. & Koch G. (1977). The measurement of observer agreement for categorical data. *Biometrics*. 33:159–174.

Lazaraton, A. (1996). Interlocutor support in oral proficiency interviews: The case of CASE, *Language Testing, 13*, 151-172.

Lazaraton, A. (2006). Process and outcome in paired oral assessment. *ELT Journal, 60*, 287-289.

Liskin-Gasparro, J. E. (2003). The ACTFL proficiency guidelines and the oral proficiency interview: a brief history and analysis of their survival. *Foreign Language Annals, 36(4),* 383-490.

Luoma, S. (2004). *Assessing Speaking*. Cambridge: Cambridge University 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. (2006). An examination of rater orientations on a paired candidate discussion task through stimulated verbal recall. *Melbourne Papers in Language Testing, 11(1),* 29–31.

May, L.A. (2009). Co-constructed interaction in a paired speaking test: The rater's perspective. *Language Testing, 26*, 397–422.

May, L.A. (2011). Interactional competence in a paired speaking test: features salient to raters. *Language Assessment Quarterly, 8,* 127–145

McNamara, T. (1996). *Second Language Performance Measuring*. London and New York: Longman.

McNamara, T. (1997). 'Interaction' in second language performance assessment: Whose pesformance? *Applied Linguistics, 16*, 159-179.

McNamara, T. & Lumley, T. (1997). The effect of interlocutor and assessment mode variables in overseas assessments of speaking skills in occupational settings. *Language Testing, 14,* 140-156.

Milanovic, M. & Saville, N. (1996). Introduction. In M. Milanovic & N.

Saville (Eds.), *Performance Testing, Cognition and Assessment.* Cambridge: Cambridge University Press, 1 – 17.

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. (2007). Developing a rating scale to assess English speaking skills of Japanese upper-secondary students. *Essex Graduate Student Papers in Language & Linguistics, 9,* 83 – 103.

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.

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

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.

Plough, I. & Gass, S. M. (1993). Interlocutor and task familiarity: Effects on interactional structure. In Crookes, G. and Gass, S. M. (Eds.), *Tasks and Language Learning. Integrating Theory and Practice*. Cleveland, Oh: Multilingual Matters, 35 – 56.

Powers, D. (2000). Computing reader agreement for the GRE writing assessment. ETS Research Memorandum.

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. (1992). The discourse of accommodation in oral proficiency interviews. *Studies in Second Language Acquisition, 14*, 159 – 176.

Ross, S. (2007). A comparative task-in interaction analysis of OPI backsliding. *Journal of Pragmatics, 39*, 2017 – 2044.

Ross, S. & Berwick, R. (1992). The discourse of accommodation in oral Proficiency interviews. *Studies in Second Language Acquisition, 14(2)*, 159 – 176.

Sacks, H., Schegloff, E. A., & Jefferson (1974). A simplest systematic for the organization of turn-taking for conversation. *Language*, *50(4)*, 696-735.

Sasaki, M. (1996). Second language proficiency, foreign language aptitude, and intelligence: *Quantitative and qualitative analysis*. New York: Peter Lang.

Saville, N. & Hargreaves, P. (1999). Assessing speaking in the revised FCE. *ELT Journal, 53(1)*, 42 – 51.

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.

Shin, S-K. (2005). Did they take the same test? Examinee language proficiency and the structure of language tests. *Language Testing, 22(1),* 31 – 57.

Sim, J. & Wright, C. (2005). The kappa statistic in reliability studies: Use,

interpretation, and sample size requirements. *The Journal of the American Physical Therapy Association. 85(3)*, 257 – 268.

SPSS Inc. (2008). SPSS Base 17.0. for Windows User's Guide. Chicago: SPSS Inc.

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.

Taylor, L. & Wigglesworth, G. (2009). Are two heads better than one? Pair work in L2 assessment contexts. *Language Testing 26(3)*, 325 – 339.

TEPS. (2009). Retrieved March, 2012 from http://www.teps.or.kr/

Thakerar, J. N., Giles, H., & Cheshire, J. (1982). Psychological and linguistic parameters of speech accommodation theory. In C. Fraser and K.R. Scherer (Eds.), *Advances in the Social Psychology of Language*. New York: Cambridge University Press, 205 – 255.

Van Lier, L. (1989). Reeling, writhing, drawling, stretching and fainting in coils: Oral proficiency interviews as conversations. *TESOL Quaterly, 23*, 480 – 508.

Weir, C.J. (1993). *Understanding and Developing Language Tests*. New York:

Prentice Hall.

Weir, C. J. (2005). *Language Testing and Validation*. New York: Palgrave Macmillan.

YBM (n.d.). *TOEIC Speaking Sample Test.* Retrieved March, 2012 from http://exam.ybmsisa.com/toeicswt/sampletest.asp

Young, R. & Milanovic, M. (1992). Discourse variation in oral proficiency interviews. *Studies in Second Language Acquisition, 14,* 403 – 424.

Appendix A

Personal Information Questionnaire

QUESTIONNAIRE SPEAKING ASSESSMENT

Please complete the following:

Name:

Gender (circle): Male / Female

Age:

Major:

Time spent abroad in an English speaking country (months or years):

What English test(s) have you taken? _____ Score(s):

_____

What was your score in the speaking section (if any): _____

How long ago did you take this/these test(s)? _____

Appendix B

Non-Interactive Test Examiner's Script

EXAMINER'S SCRIPT Non-Interactive Test

🕑 2:00 minutes in total

🕑 30 seconds to prepare in total

🕑 1:30 minutes to complete all three parts

This speaking test will be divided into two parts. In the first part you will describe two different pictures and in the second part you will answer a question related to these pictures.

Let's begin.

In this part of the test, you will describe two different pictures of *people doing different activities* (point at each picture).

You will have 20 seconds to prepare your response for both pictures. Then you will have 30 seconds to describe each picture. Please try to describe the pictures in detail.

Your 20 seconds of preparation start now.

(After 20 seconds)

🕑 Now you have 30 seconds to describe the first picture.

🕑 Now you have 30 seconds to describe the second picture.

In this part of the test, you will answer the question (point at the question):

Which activity would you enjoy doing more? Explain.

You will have 10 seconds to prepare your response. Then you will have 30 seconds to respond.

Your 10 seconds of preparation start now.

(After 10 seconds)

🕑 Now you have 30 seconds to answer the question.

Thank you.

Interactive Test Script and Prompts

EXAMINER'S SCRIPT

Pair test – Film Club

5:00 minutes in total

- Good afternoon / good evening, my name is _____
- What are your names?

  Let me write them down

In this test, you will have to discuss with each other a certain topic related to 7 pictures.

You will have about 5 minutes to discuss with your partner. Don't worry if you are interrupted in the end.

Remember you should discuss with each other, I will only be watching.

Here is the prompt card (show the prompt card).

There are two questions that you should answer.

(Read the prompt card out loud)

Any questions?

Imagine you are a member of a film club in the university. There will be a movie festival that would like to show two interesting movies on a big screen to all of the university students. Your job is to find the best two movie genres for the students. You should discuss with your partner and decide on the best two options of movie genres. Listen to your partner's opinion and also express your opinion.

First, <u>briefly</u> talk to each other about *what is interesting about <u>each movie genre</u> in the picture.*

Then, decide together *which <u>two movie genres</u> you would choose to show to the students in the university.*

IT IS IMPORTANT THAT YOU <u>REACH A DECISION</u> TOGETHER.

Here is a timer so that you know how long you've spoken.

You may begin discussing with your partner now.

Thank you.

EXAMINER'S SCRIPT

Pair test – Coffee Shop

5 minutes in total

- Good afternoon / good evening, my name is _____
- What are your names?

  Let me write them down.

In this test, you will have to discuss with each other a certain topic related to 7 pictures.

You will have about 5 minutes to discuss with your partner. Don't worry if you are interrupted in the end.

Remember you should discuss with each other, I will only be watching.

Here is the prompt card (show the prompt card).

There are two questions that you should answer.

(Read the prompt card out loud)

Any questions?

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, <u>briefly</u> talk to each other about *how successful <u>each suggestion</u> might be.*

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

IT IS IMPORTANT THAT YOU <u>REACH A DECISION</u> TOGETHER.

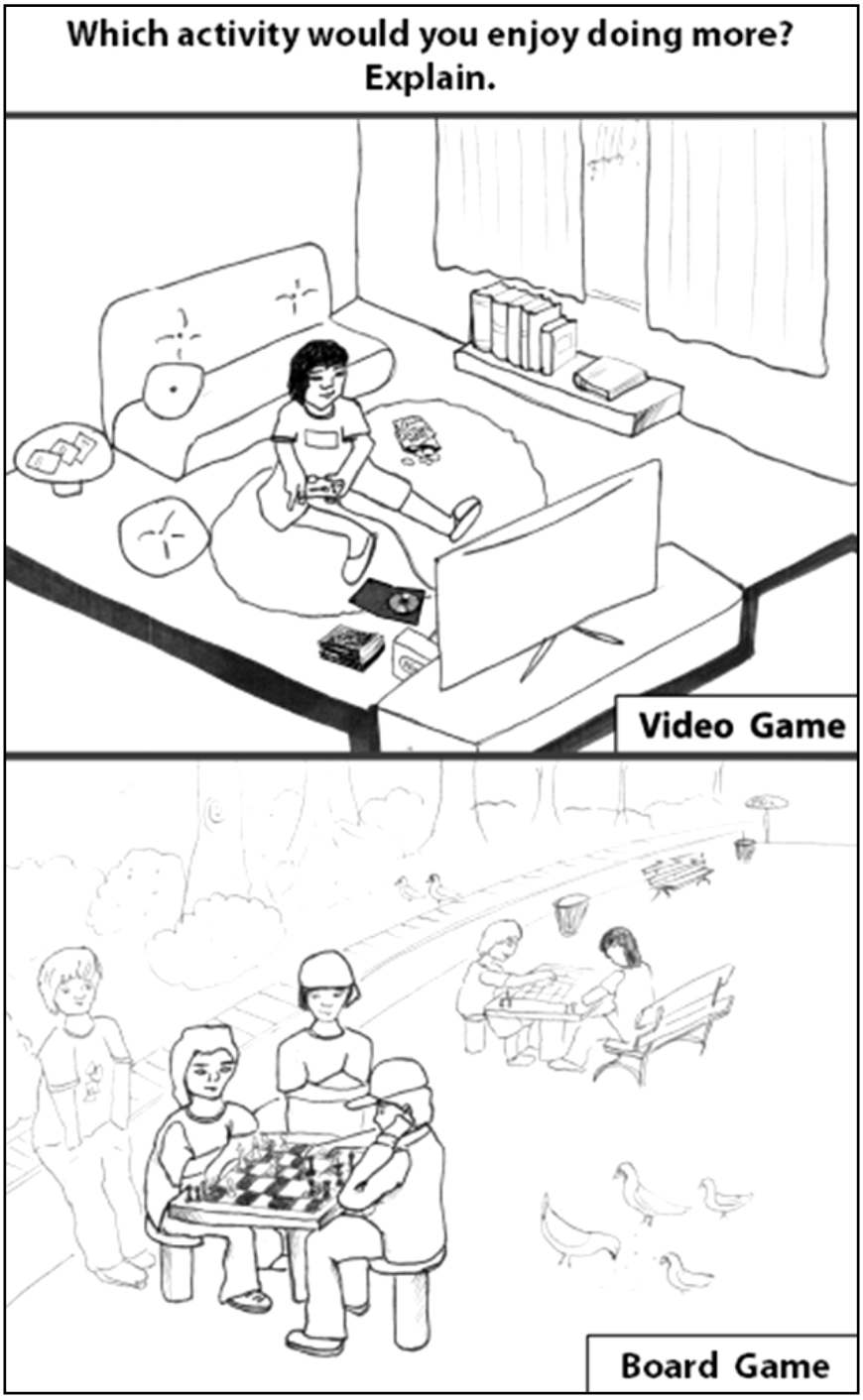
Here is a timer so that you know how long you've spoken.

You may begin discussing with your partner now.

Thank you.

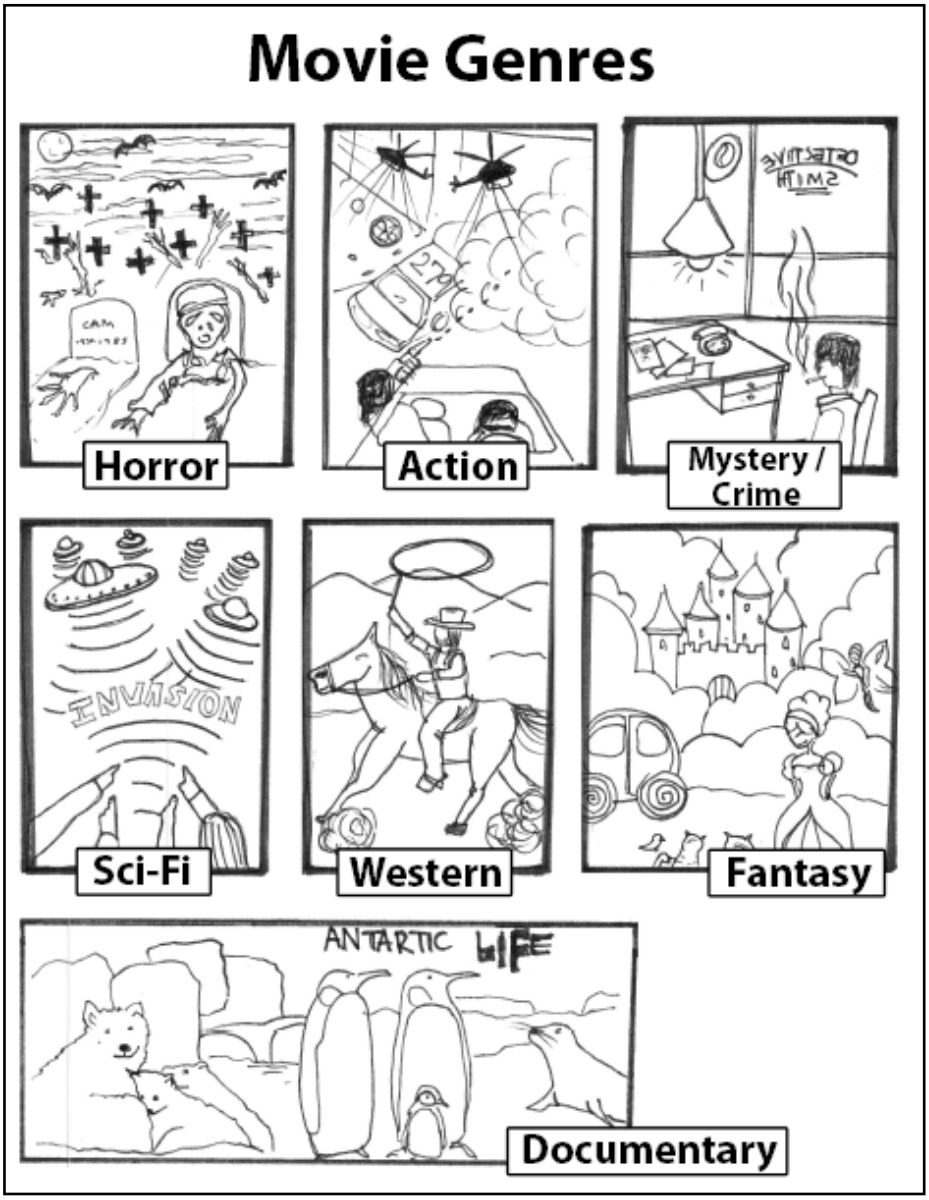
Appendix C

Non-Interactive Test

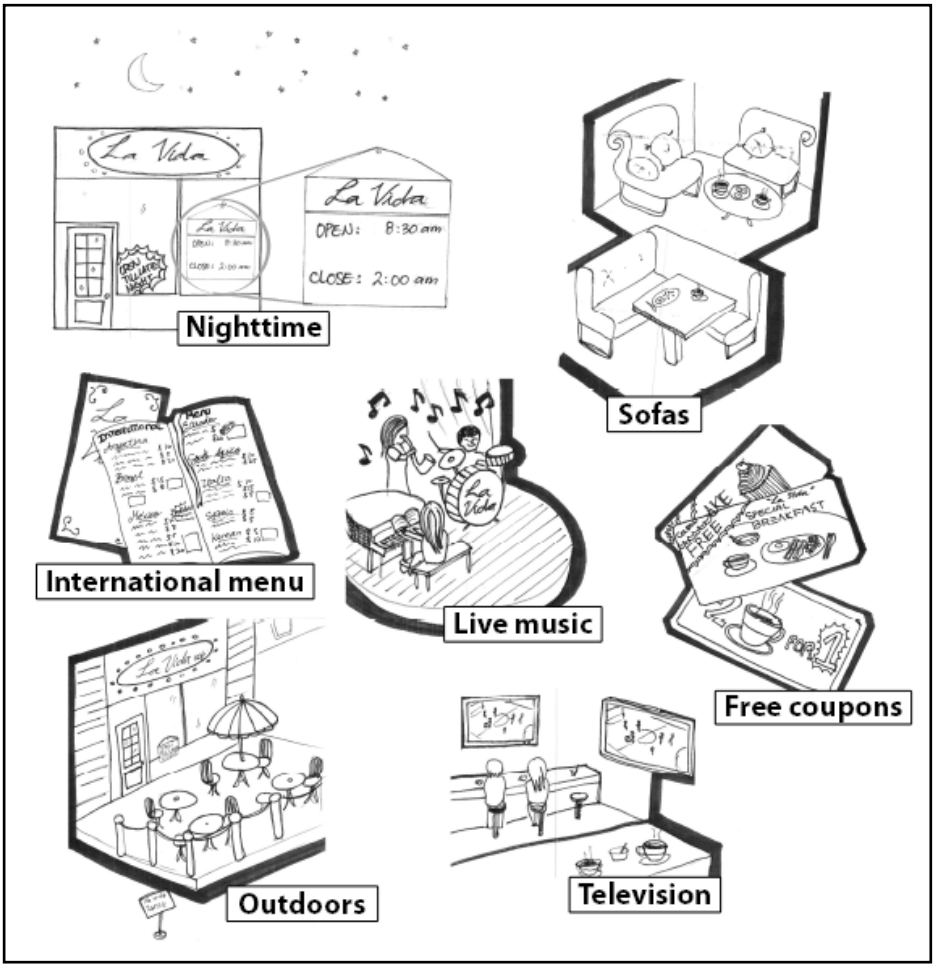


Appendix D

Interactive Test – Film Club



Interactive Test – Coffee Shop



Appendix E

Sample Interactive Test

Instructions for the next test

The second part of the experiment will assess your English conversational skills.
You will be asked to talk to another student about a certain topic related to seven pictures.
First, you will be asked to discuss all seven pictures with your partner.
Then, you will have to decide together which two options out of the seven are the best for a given situation.
The test will last 5 minutes.
The examiner will not participate in the conversation, she will only be watching.
After you finish the test, you will be asked to take the test again with different pictures with another student.

Here is a sample test:
Situation:
Imagine that a busy international hotel is looking to hire some people for the holiday season. Here are seven positions that are available:

First, <u>discuss with your partner</u> *how difficult it would be to do* ***each*** *of these jobs without training*.
Then, <u>decide together</u> *which* ***two*** *jobs you would find the most difficult to do*.
Express your opinion and also listen to your partner's opinion.
You will have to talk to each other for about 5 minutes. Please do not worry if you get interrupted in the end.

Appendix F

Experimental Consent

CONSENT FORM

English Paired Speaking Assessment

You are invited to participate in a research study about English speaking assessment. This study is being conducted by Ms. Young A Son. You are invited to participate in this study because you are a Korean university student who has studied English. Please read the following consent form carefully before you decide to participate in this study.

Purpose of the study:

This study aims to learn about Korean university students' English speaking performance under different testing situations. Your speech will be analyzed in order to see how your performance changes in different conditions.

Procedure:

The experiment will be divided into two parts and will take 2 days to complete it. You will be asked to take 3 brief speaking tests. The first test will assess your English speaking abilities individually, while the second and third tests will assess your speaking abilities with a partner.

Time required:

The time required for participation is about 15 minutes: 2-3 minutes for the first test and about 5-6 minutes for each of the other tests.

Access to existing records:

You were requested to provide us with your TEPS scores.

Risks and benefits:

You will be video-taped and audio recorded during the tests. Therefore, there is a small risk that someone else other than the researchers might see the video tapes or hear the recordings. However, this is very unlikely to happen, since the researchers are the only people allowed to see the videos and hear the recordings. The files will be kept under a secured folder and they will be destroyed once the research is finished. The videotapes and recordings will not be used for purposes other than this research study.

There are no direct benefits for participating in this experiment.

Compensation:

You will receive a payment of 10,000 won when you have finished taking all three tests.

Confidentiality:

Your identity will be kept confidential as your name or any other information that could possibly indicate your identity will be excluded from the final report of this research study.

Voluntary participation:

Your participation in this study is completely voluntary. If you choose not to participate in this study, this will have no effect on the services or benefits you are currently receiving. You may choose to stop participating in the study at any time. This will have no effect on your current or future relations with Seoul National University.

Contact information:
If you have any questions, you can contact the researcher at:
Ms. Young A Son 010 4842 3602 email young.ah.son@gmail.com

I have read and understood the information stated above and consent to participate in this study.

Participant:
Name :_____ Signature: _____

Researcher:
Name: _____ Signature: _____

Date: _____

Appendix G

Holistic Scale

5	Produces speech that is highly intelligible. The word-stress/rhythm/intonation is mostly accurate; thus effortless to understand. Uses a full range of complex structures and vocabulary appropriately and accurately. Hesitation and pauses are still present but they are few and only content-related situations rather than when searching for words or precise grammatical structures. Expresses thoughts and opinions without any difficulty connecting ideas coherently by appropriately using discourse markers.
4	Produces speech that is influenced by some L1 prosodic features. However, this has minimal effects on intelligibility; thus the speech is easy to understand. Uses a wide range of structures with few inaccuracies, especially when attempting to use more complex structures. Self-correction and rephrasing is sometimes present. Uses a wide range of vocabulary which is generally accurate. Hesitation and pauses are seldom present when searching for words. Expresses thoughts and opinions with minimal difficulty. There might be some inadequate use of discourse markers but speech is mostly coherent.
3	Produces speech that is marked by L1 prosodic features making it at times difficult to understand. Mostly uses basic structures that are occasionally inaccurate. There is some attempt to use complex forms but it is mostly imprecise. Self-corrections and rephrasing are frequently present but not always successful. Uses a limited range of vocabulary that is occasionally inaccurate. Hesitation and long pauses are frequent when searching for words but do not impede comprehensibility. Has occasional difficulty expressing thoughts and opinions generally providing basic ideas. Loses coherence at times and uses simple discourse markers repeating them throughout the speech sample.
2	Produces speech that is heavily influenced by L1 prosodic features which is nearly always difficult to understand. Only uses basic structures and vocabulary that are at times inaccurate. All attempts, if any, to use complex forms are inaccurate. Frequent hesitation and long pauses make speech slow and occasionally impedes comprehensibility. There are few abandoned sentences that impede the natural flow of speech. Only expresses simple ideas and has difficulty expressing thoughts and opinions. Rarely uses even simple discourse markers.
1	Produces speech that is unintelligible due to very frequent mispronunciations. Uses a very limited range of structures and vocabulary inaccurately. Speech is full of incomplete or abandoned sentences. Long pauses and constant hesitation makes speech hard to be rated. Unable to convey basic ideas.

Appendix H

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 I
Complete Data of Raw Scores

Tests	Low-Proficiency N	Low-Proficiency Mean	Low-Proficiency SD	Intermediate-Proficiency N	Intermediate-Proficiency Mean	Intermediate-Proficiency SD	High-Proficiency N	High-Proficiency Mean	High-Proficiency SD	All proficiencies combined N	All proficiencies combined Mean	All proficiencies combined SD
TEPS	12	728.08[a]	87.71	8	875.75	71.17	12	890.92	65.36	32	826.06	106.63
NIT		8.10[b]	.86		11.02	.49		13.15	.96		10.72	2.37
Task 1	12	2.75	.37	8	3.54	.28	12	4.33	.49	32	3.54	.80
Task 2		2.61	.39		3.63	.25		4.43	.42		3.55	.88
Task 3		2.74	.35		3.85	.31		4.39	.39		3.64	.82
PSP Total Score		14.88[c]	2.65		20.38	1.64		22.54	1.89		19.12	4.05
G		2.75	.66		4.00	.65		4.38	.53		3.67	.95
V	12	3.04	.45	8	4.19	.26	12	4.54	.45	32	3.89	.79
P		2.92	.85		4.13	.58		4.42	.29		3.78	.92
F		2.79	.54		3.69	.46		4.58	.56		3.69	.94
DM		3.38	.61		4.38	.23		4.63	.43		4.09	.73
PDP Total Score[d]		15.17	2.51		17.44	1.47		22.54	2.27		18.44	3.89
G		2.71	.45		3.19	.26		4.50	.56		3.50	.92
V	12	3.13	.53	8	3.38	.35	12	4.54	.58	32	3.72	.82
P		3.17	.86		3.63	.35		4.25	.66		3.69	.82
F		2.79	.33		3.38	.44		4.54	.45		3.59	.87
DM		3.38	.57		3.88	.44		4.71	.45		4.00	.76

Notes. NIT = non-interactive test; PSP = paired with same-proficiency; PDP = paired with different-proficiency

[a] The maximum score is 900

[b] The maximum score is 15 for the total score, and 5 for each task

[c] The maximum score is 25 for the total score, and 5 for each criterion (this applies for both paired tests)

[d] For the intermediate-proficiency group, the pairings were made with same-proficiency both times

Appendix J

Figures of Gender Score Variation for Each Criterion

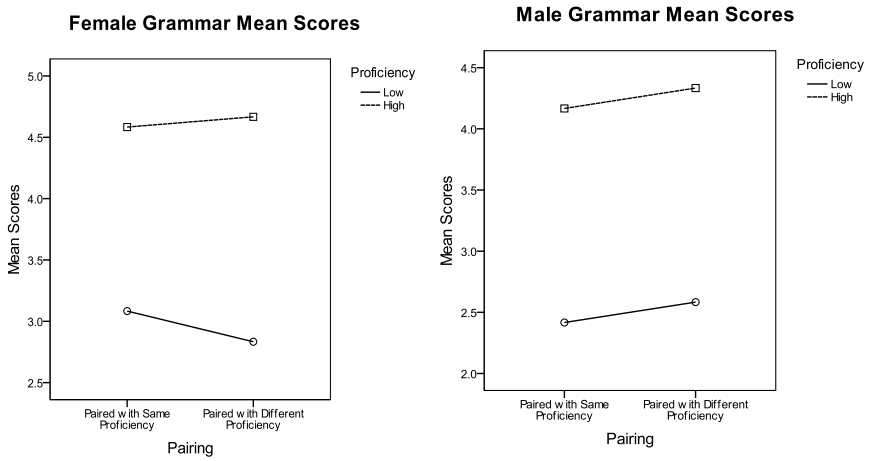


Figure J.1 Grammar mean scores in different pairing conditions according to gender

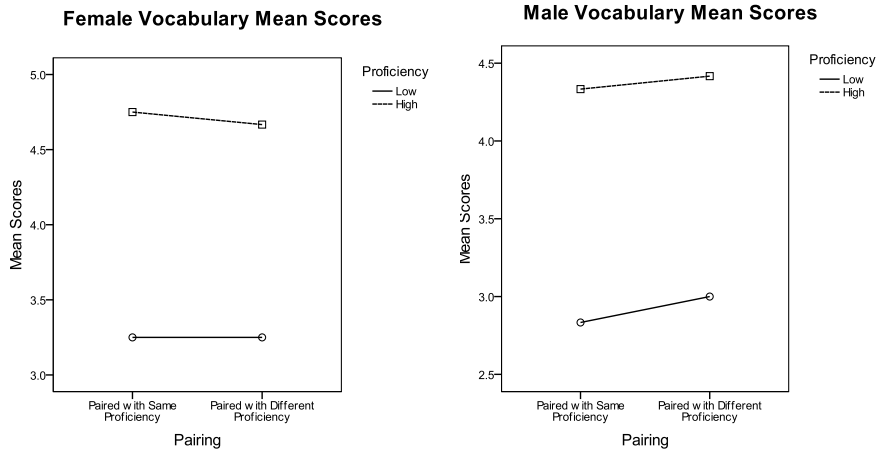


Figure J.2 Vocabulary mean scores in different pairing conditions according to gender

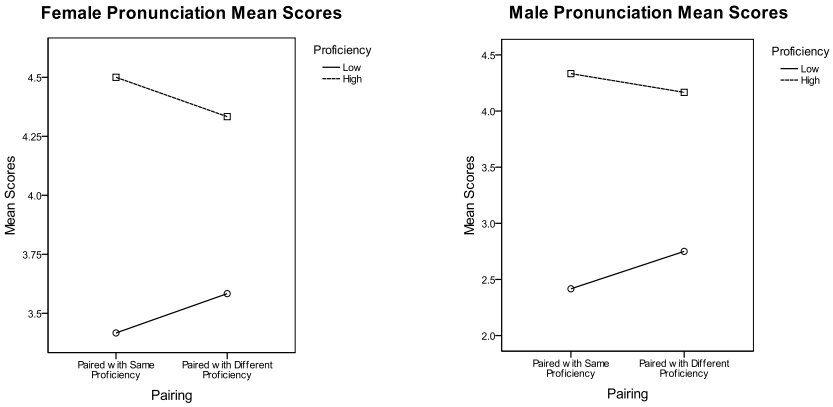


Figure J.3 Pronunciation mean scores in different pairing conditions according to gender

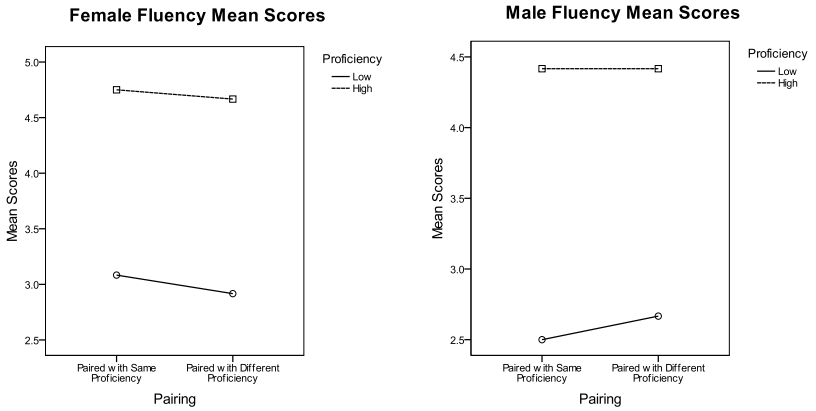


Figure J.4 Fluency mean scores in different pairing conditions according to gender

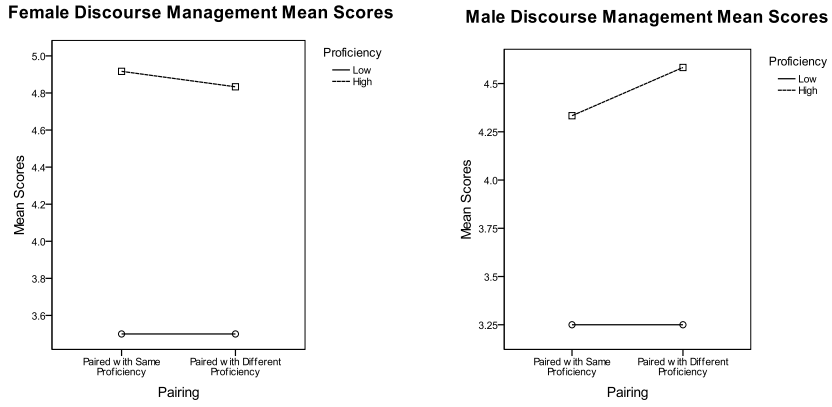


Figure J.5 Discourse management mean scores in different pairing conditions according to gender

Appendix K

Description of Raw Scores of Individual Participants

Group	Gender	Subject	NI Test	Paired Tests			
				PSP	PDP	Difference	%
1	Female	20	12.5	22.5	22.0	0.5	2%
		19	13.0	24.0	23.0	1.0	4%
		16	8.0	19.0	17.0	2.0	11%
		12	8.7	16.5	17.5	-1.0	-6%
2	Male	14	12.0	23.5	23.5	0.0	0%
		7	13.3	23.0	23.5	-0.5	-2%
		11	8.5	12.5	13.5	-1.0	-8%
		13	8.0	13.0	14.0	1.0	8%
3	Female	34	13.0	22.5	24.0	-1.5	-7%
		22	15.0	24.5	25.0	-0.5	-2%
		24	7.5	12.5	11.5	1.0	8%
		2	10.0	16.5	15.0	1.5	9%
4	Male	30	12.5	18.0	16.0	2.0	11%
		29	12.0	21.0	23.5	-2.5	-12%
		4	6.5	12.0	13.0	-1.0	-8%
		25	8.0	15.0	13.5	1.5	10%
6	Female	26	13.5	23.0	23.5	-0.5	-2%
		28	12.5	24.5	21.5	3.0	12%
		21	8.5	20.0	20.5	-0.5	-2.5%
		1	8.5	13.5	15.0	-1.5	-11%
7	Male	23	14.0	23.5	23.0	0.5	2%
		15	14.5	20.5	22.0	-1.5	-7%
		36	7.5	13.0	17.5	-4.5	-35%
		32	7.5	15.0	14.0	1.0	7%

Notes. Groups 5 and 8 were the intermediate-proficiency groups and were excluded from the analysis.
The first two scores in each group belong to high-proficiency participants

Appendix L

Raters' Post-Rating Questionnaire

POST-RATING FEEDBACK

1. What was your overall opinion about the speaking test? (e.g., were you familiar with it? How did you find it? You can compare it with other types of speaking tests you've rated before).

2. What was the most difficult thing when rating the speech samples?

3. How would you rank the analytic criteria according to their rating difficulty? (1 being the least difficult and 5 being the most difficult).

   _____ Grammar

   _____ Vocabulary

   _____ Pronunciation

   _____ Fluency

   _____ Discourse Management

   Why did you choose this particular order?

4. Was there any difference between the conversation of pairs of high-high, high-low, and low-low proficiency test-takers? Please explain your answer.

5. Do you have any other comments that are not addressed in the previous questions?

국문초록

짝 형식 말하기평가에서 대화상대자의 영어능력이 수험자의 말하기 수행에 미치는 영향

손영아

서울대학교 대학원

영어영문학과 영어어학 전공

짝 형식 말하기평가는 대화 능력을 측정하는 진정성 있는 평가 방법으로서, 영어 평가 영역에서 널리 활용되는 추세이다. 이러한 평가형식에서는 두 명 혹은 그 이상의 수험자가 일정한 말하기 과제를 함께 수행하면서 상호작용을 할 수 있는 장점이 있다. 하지만 짝 형식 말하기평가의 여러 장점에도 불구하고, 공정성, 신뢰도, 구성 타당도 측면에서 문제점이 제기되어 왔다. 특히 이러한 평가방식에서는 대화상대자의 영어 능력이 다른 수험자에게 미칠 수 있다는 점이 지적되어 왔다. 이 분야의 선행연구는 그동안 매우 적었고 일부 선행연구들에서도 상충되는 결과가 도출되었다.

본 논문은 짝 형식 말하기 평가에서 대화상대자의 말하기 능력이 수험자의 수행에 미치는 영향을 실험에 의거하여 조사하였다. 본 연구는 Iwashita (1998) 및 Davis (2009)와 유사한 실험디자인을 사용하였다. 24명의 한국인 대학생을 영어 능력에 따라 능력이 높은 그룹과 낮은 그룹으로 배정한 뒤, 조건을 달리하여 실행하는 두 번의 짝 형식 평가에 참여하게 하였다. 둘 중의 한 번은 자신과 비슷한 영어 수준의 대화상대자와 함께 평가에 참여하였고 다른 한 번은 자신과 다른 수준의 대화상대자와 함께 짝 형식 평가를 받았다.

데이터분석에서는 수험자의 영어능력과 짝 배정조건을 독립변인으로 삼아 혼합설계분산분석(ANOVA)을 시행하였다. 우선 두 번의 짝

형식 평가에서 조건을 다르게 한 것이 수험자의 말하기 수행에 어떤 영향을 주는지를 알아보기 위하여, 문법, 어휘, 발음, 유창성 및 담화 총 다섯 가지 분석적 영역 점수를 합한 합산점수를 종속변인으로 삼아서 혼합설계분산분석을 실시하였다. 다음으로 짝 배정 조건이 각 분석적 영역의 점수에 미칠 수 있는 효과를 분석하기 위하여, 각각의 분석적 영역 점수를 종속변인으로 삼아 총 5번의 혼합설계분산분석을 실시하였다. 마지막으로 각 수험자를 성별집단으로 나누어 분산분석을 실시하여, 성별이 대화상대자의 영어능력에 따른 수험자의 점수변화를 완화시키는 요인이었는지를 살펴보았다.

분석결과는 대화상대자의 영어능력이 수험자의 합산점수에 유의미한 영향을 미치지 않음을 보여주었다. 그리고 대화상대자의 영어능력은 각 분석적 영역의 점수에 통계적으로 유의미한 영향이 없었다. 또한 성별집단으로 나누어 분석한 결과, 대화상대자의 영어능력은 여성 수험자와 남성수험자 모두의 합산점수와 총 5개의 각 분석적 점수에 통계적으로 유의미한 영향을 미치지 않았다. 요컨대 대화상대자의 영어능력은 짝 형식 말하기평가에 참가하는 수험자의 수행에 통계적으로 유의미한 영향을 주지 않았다. 본 연구는 짝 형식 말하기 평가가 타당성, 신뢰성, 공정성을 갖는 말하기 능력 측정 방식임을 뒷받침하는 근거를 제시한다.

주요어: 대화상대의 언어능력, 상호작용, 짝 형식 말하기 평가
학번: 2010-22943