문학석사 학위논문

Assessing Interactional Competence of Advanced-Level Korean EFL Students Through L2 Paired Discussion Tasks

2인 영어 토론 과제를 활용한 영어 능력 수준 상위권 한국 학생들의 상호작용 능력 평가 연구

2017년 8월

서울대학교 대학원

영어영문학과 영어학 전공

박 희 민

Assessing Interactional Competence of Advanced-Level Korean EFL Students Through L2 Paired Discussion Tasks

지도 교수 이 용 원

이 논문을 문학석사 학위논문으로 제출함

2017년 4월

서울대학교 대학원

영어영문학과 영어학 전공

박 희 민

박희민의 문학석사 학위논문을 인준함

2017년 6월

위 원 장 이 재 영 (인)

부위원장 소 영 순 (인)

위 원 이 용 원 (인)

Abstract

Assessing Interactional Competence of Advanced-Level Korean EFL Students Through L2 Paired Discussion Tasks

Park, Heemin

Department of English Language and Literature

Seoul National University

In recent years, interaction-based speaking tests, including paired and group oral tests, have been considered as viable formats of second language (L2) speaking assessment. A promising aspect of such tests is that they can provide authentic and real conversation-like tasks for test takers to participate in, enable us to investigate L2 learners' interactional competence as one crucial dimension of L2 speaking proficiency, and thereby increase the validity of speaking scores obtained from such tests. This promising characteristics of paired speaking tests has led to a number of research studies exploring the nature of interactional competence both theoretically and empirically as administering actual interaction-based tests in varying formats and contexts. However, little research studies have been conducted

on the characteristics of interactive tasks and raters including rating scales they use which highly affect the quality of test takers' speech performance and their ratings as the core variables in the context of speaking assessments. In response to this need for research, this study aims to investigate the feasibility of using paired discussion tasks to assess interactional competence and an existing scoring rubric for the construct in the Korean EFL context through both quantitative and qualitative methodologies.

For this study, two paired discussion tasks were prepared and administered to 40 Korean EFL students at an advanced level of English proficiency. A total of 20 pairs of interlocutors were created and each pair completed two of the same discussion tasks. The elicited samples were rated first by two native English speaking raters on the dimension of the interlocutors' interactional competence and then by two EFL trained raters on their general speaking proficiency. The elicited speech samples were also analyzed in terms of the types of interactional features operationalized during the test tasks and how raters perceived the features for assessing interlocutors' interaction. Along with the patterns of salient interaction-related features in the test takers' spoken responses, the reliability and validity of the obtained speaking scores were also examined. In addition, to investigate the raters' perception of the tasks and scoring rubrics used for this study, their comments were collected and analyzed.

The results of the study suggest that the paired discussion tasks not only achieved the acceptable levels of inter rater reliability but also elicited

a variety of interactional features from the interlocutors. First, a total of 18 interactional features were identified through the discussion tasks, each of which appeared with different frequencies in this study. This result has provided some validity evidence supporting the use of paired discussion tasks to elicit interaction-based speech performance. Second, the qualitative analysis of the interlocutors' response based on the raters' comments has demonstrated that in-depth investigation of interactional features in conjunction with task characteristics may be useful and effective. For example, considering the characteristics of discussion tasks, the raters focused on how the topics were appropriately developed during the interaction for the given tasks, enabling the raters to make sophisticated identification of interactional competence. In this regard, clear limitations of the existing rating scale for interactional competence were exposed since the descriptors of the used rating scale did not fully reflect the actual speech performance. In addition, raters' feedback on the rating scale revealed that making consistent and valid judgments was difficult for them due to the limited levels of interaction patterns and task completion status.

Some implications and future research directions seem to arise from the major findings of this study. First, the current study empirically examined not only the reliability and validity of scores from paired discussion tasks but also the feasibility of using such paired oral tasks to elicit abundant interactional features as evidence of the interlocutors' interactional competence. This has also suggested the need for further

research on exploring interactional resources in the context of varying forms and structures of tasks. Second, the study provided helpful suggestions for constructing rating scales for interactional competence. For teachers and assessors, it is recommended to develop a rating scale with several levels of criteria and detailed descriptors of each level based on empirical analysis. Lastly, further studies on raters' perspectives on the conceptualization of interactional competence and elicited speech performance are needed to provide them with guidance for making consistent and valid judgments, thereby ensuring the fairness of administering paired discussion tasks in the context of L2 speaking assessment.

Keywords: Interactional competence, paired speaking tests, second language assessments, task characteristics, rating scales, raters' perspectives, Korean EFL learners

Student Number: 2015-20050

Table of Contents

CHAPTER 1 INTRODUCTION 1

1.1 Background and Motivation 1

1.2 Research Questions 6

1.3 Organization of the Thesis 7

CHAPTER 2 LITERATURE REVIEW 9

2.1 Speaking Ability and Interactional Competence 9

2.1.1 Defining Speaking Ability as a Construct 9

2.1.2 Conceptualization of Interactional Competence 13

2.1.3 Interactional Competence as a (Sub) Construct 17

2.1.3.1 Patterns of Interaction at a Macro-level 18

2.1.3.2 Interactional Features at a Micro-level 21

2.2 Theoretical Framework for L2 Speaking Tests 29

2.3 Speaking Test Formats for Interactional Competence: Oral Proficiency Interviews and Paired Speaking Tests 32

2.4 Task-based Approach to L2 Speaking Tests 35

2.4.1 Task Types in the Paired Speaking Test 36

2.4.2 The Role of Interlocutors in the Paired Speaking Test 39

2.4.3 The Role of Topic in the Paired Speaking Test 41

2.5 Rating Scales for Interactional Competence 43

CHAPTER 3 METHODS 48

3.1 Participants 48

3.1.1 Test takers 48

3.1.2 Raters 49

3.1.3 Examiner 51

3.2 Instruments 52

3.2.1 Paired Discussion Tasks 52

3.2.2 Rating Scales 56

3.3 Procedures 58

3.3.1 Data Collection from Interlocutors 58

3.3.2 Data Transcription and Collection of Data from Raters 59

3.4 Methods of Analysis 64

CHAPTER 4 RESULTS 68

4.1 Descriptive Statistics 68

4.2 Reliability Measures 73

4.2.1 Inter rater Reliability 75

4.3 Correlations among Tasks and Other Criterion Measures 77

4.4 Frequency Analysis on Interactional Features 81

4.5 Qualitative Analysis of Interlocutors' Response 90

4.6 Raters' Feedback on the Rating Scale for Interactional Competence 96

CHAPTER 5 DISCUSSION 99

5.1 The Reliability of Interaction Scores and Their Relationships with Other Criterion Measures 99

5.1.1 The Reliability of Interaction Scores 99

5.1.2 The Relationships Between the Scores from Discussion Tasks and Other Criterion Measures 102

5.2 Salient Interactional Features in Paired Discussion Tasks 106

5.2.1 Types of Interactional Features in Paired Discussion Tasks . 106

5.2.2 The Effect of Topic on Interlocutors' Response..... 107

5.2.3 Differences in Raters' Salience on Interactional Features 111

5.3 Raters' Perception of Interlocutors' Response 112

5.4 Raters' Feedback on the Rating Scale for Interactional Competence 114

CHAPTER 6 CONCLUSION 117

REFERENCES 122

APPENDICES..... 134

Appendix A Paired Discussion Tasks 134

Appendix B Rating Scale for General Speaking Proficiency 136

Appendix C Rating Scale for Interactional Competence 141

Appendix D Raters' Comments on Interactional Features..... 143

Appendix E Raters' Feedback about the Rating Scale for Interactional Competence 147

Appendix F The Number of Turns and Utterances 150

Appendix G Research Consent Form for the Participants 152

국문 초록 155

List of Tables

Table 2.1 Interaction features in hypothesized categories (Ducasse and Brown, 2009).....23

Table 3.1 Background information about the participants 49

Table 3.2 List of interactional features perceived by raters in the paired discussion tasks (n=18) 67

Table 4.1 Means and standard deviations of TEPS, general speaking proficiency, and interaction scores of interlocutors 69

Table 4.2 Descriptive statistics for the scores of general speaking proficiency in the paired discussion tasks 72

Table 4.3 Descriptive statistics for interaction scores in the paired discussion tasks 73

Table 4.4 Score agreement rates, Spearman rank-order and Pearson correlation, weighted Kappa, and reliability coefficients for general speaking proficiency scores.....75

Table 4.5 Score agreement rates, Spearman rank-order and Pearson correlation, weighted Kappa, and reliability coefficients for interaction scores.....76

Table 4.6 Coefficients of Pearson correlations among scores for TEPS, general speaking proficiency and interaction.....79

Table 4.7 Descriptive statistics for the perceived non-verbal interpersonal features in the paired discussion tasks.....84

Table 4.8 Descriptive statistics for the perceived interactional features of interactive listening in the paired discussion tasks.....85

Table 4.9 Descriptive statistics for the perceived interactional features of interactional management in the paired discussion tasks.....88

Table 4.10 A summary of the most frequently perceived interactional features.....90

Table 4.11 Interactional features related to topic development.....92

Table 4.12 Interactional features related to interaction at a macro level....94

List of Figures

Figure 2.1 Three patterns of interaction in dimensions of mutuality and equality (Galaczi, 2004, p. 107) 19

Figure 2.2 Fulcher's (2003, p. 115) expanded model of speaking test performance 31

Figure 3.1 Paired Discussion Task 1.....54

Figure 3.2 Paired Discussion Task 2.....55

Figure 4.1 Histogram of interlocutors' TEPS scores 70

Figure 4.2 Histogram of interlocutors' general speaking proficiency scores 71

Figure 4.3 Histogram of interlocutors' interaction scores 71

Figure 4.4 Scatter plot of total scores of general speaking proficiency and TEPS scores 80

Figure 4.5 Scatter plot of total scores of interaction and TEPS scores 80

Figure 4.6 Scatter plot of total scores of interaction and general speaking proficiency 80

Figure 4.7 Salient interactional features perceived by raters for each task 82

Figure 4.8 Salient interactional features perceived by each rater during Task 1.....83

Figure 4.9 Salient interactional features perceived by each rater during Task 2.....83

Chapter 1 Introduction

1.1 Background and Motivation

For decades, various formats of interaction-based speaking tests including Oral Proficiency Interview (OPI), Semi-direct Oral Proficiency Interview (SOPI), and paired and group oral tests have been proposed and used to measure test takers' speech performance. As one of the popular formats, the OPI has been widely utilized and investigated so far. Numerous researchers, however, have criticized the OPI for its clear limitations in terms of authenticity and validity. As Van Lier (1989) pointed out, the purpose of administering the OPI is to successfully elicit ratable samples of test takers' language use in speaking, not to simply enable them to have a "successful conversation" with another interlocutor. To be more specific, the format of the OPI hardly lends itself well in creating an interactional situation in which the basic features of authentic conversation can be manifested, which can include "face-to-face interaction, unplannedness, potentially equal distribution of rights and duties in talk, and manifestation of features of reactive and mutual contingency" (Van Lier, 1989, p. 495). This is because the interaction during the OPI is dominantly controlled by the interviewer while the interviewee merely answers the predetermined

questions posed by the interviewer. Therefore, the OPI has severe drawbacks in terms of eliciting collaborative patterns of interaction with various interactional features that can be observed in an authentic communicative context.

In recent years, to overcome these critical problems of the OPI in the assessment of second language (L2) speaking ability, paired speaking assessments along with group oral tests are being considered as one potential test format that can provide authentic and real conversation-like tasks for test takers to participate in. These promising characteristics of paired speaking assessments enable investigation of L2 learners' interactional competence as one crucial dimension of speaking proficiency in L2, hence increasing the validity of speaking scores obtained from such tests. However, as for the assessors whose responsibility is to assign scores to each of the interlocutors, the paired- or group oral test is challenging in that speech performance elicited from such tests is jointly constructed by multiple interlocutors, which makes it difficult to evaluate the interactional competence in isolation of a single interlocutor. Accordingly, many conceptual and empirical research studies thus have been conducted to explore the nature of interactional competence, including types of interactional features elicited from such tests, factors that may affect test takers' speech performance such as personality or proficiency of

interlocutors to talk with, and characteristics of test tasks (Brooks, 2009; Davis, 2009; Ducasse & Brown, 2009; Galaczi, 2008, 2013; Jungheim, 2001; Leaper & Riazi, 2014; Ockey, 2009; Sato, 2014).

Another important trend is that not much research has been carried out on the potential effects that the various factors of test tasks have on interlocutors' speech performance. Noticing the possible impact that different topics of test tasks have on interlocutors' speech performance, Nakatsuhara (2006) and Leaper and Riazi (2014) examined the role of different task types, resulting in contradictory findings from each other. Nakatsuhara (2006) investigated how four different topics of a problem solving discussion task have an influence on the relationship between interlocutors' levels of proficiency and their interaction patterns. The findings showed that there was no significant effect that the different topics had on the quality of interlocutors' speech performance. On the other hand, the findings of Leaper and Riazi's study (2014) revealed that different topics and questions that were combined differently for each topic affected syntactic complexity, fluency and turns taken during interaction but there was no significant difference in test takers' scores.

The previous studies above are meaningful in that they noted the primary role of task characteristics in the L2 assessment and investigated how variables of a task had impact on test takers' speech performance and

scores to be obtained. However, both studies are limited in the sense that they were quantitatively-oriented research without using qualitative methods that enable to observe what interactional features are operationalized during a particular task and how the features affect the raters' judgments on assigning scores. Furthermore, among various types of test tasks, a limited range of tasks was used in the previous studies, such activities as describing pictures and story-telling. As Nunan (2004) proposed, diverse task types of various topics can be designed, developed, and used for different testing and pedagogical purposes. Therefore, more studies are needed to investigate what interactional features are elicited from certain tasks or task types, what types of interaction patterns are observed, and how the elicited features and patterns affect raters' perception about test takers and their performance.

Also surprisingly, little research has been done on raters' perception of elicited interactional features and rating scales for assessing interactional competence. For decades, theoretical models of language proficiency have been used together with some pilot-test results to design and develop language tests and scoring rubrics. However, not enough attention has been paid to the investigation of the validity of conceptualized constructs based on performance data, which has resulted in the lack of validation research to examine the validity of language tests and rating scales for them (Bachman & Savignon, 1986; Canale, 1983; Fulcher, 1995). When it comes to raters'

perception about elicited interactional features, May (2009) found that raters positively perceived interactional features from a collaborative pattern while they were having difficulty in assigning scores for interactional competence when interlocutors showed an asymmetric pattern of discourse during interaction. In addition, May (2011) investigated raters' perception of salient interactional features on the basis of qualitative data such as raters' summary statements, verbal reports, and discussion. Then, the collected data were organized into three categories which include: the ability to understand the interlocutors' message, the ability to respond appropriately to interlocutor, and the use of appropriate communicative strategies. In each category, two subcategories that illustrated raters' positive and negative perception about elicited interactional features were included. The findings of May (2011) identified a range of interactional features that were salient to raters and introduced why raters perceived a certain feature as positive or negative evidence of interlocutors' interactional competence. However, no findings were reported in terms of how well the adopted rating scale functioned and how raters' ideas on observed interactional features were reflected in it, along with the observed quality of interaction, in the process of rating the interlocutor performance.

With these as a backdrop, the main goals of the present study are to: (a) investigate the validity and feasibility of using paired discussion tasks

for assessing interactional competence; (b) examine the feasibility of the existing rating scale designed by Wang (2015) for assessing interactional competence in the context of free discussion for Korean students at an advanced level of English proficiency; and (c) identify salient interactional features to raters and their perceptions on the used rating scale, with a view to gleaning some insights for developing reliable and adequate rating scales.

1.2 Research Questions

In examining the validity and feasibility of paired discussion tasks for assessing interactional competence, the current study aims to address the following research questions:

1) Do raters' scores for test takers' interactional competence in paired discussion tasks achieve acceptable levels of inter rater and score reliabilities?

2) What are the relationships among paired discussion tasks and other criterion measures? How do the scores from paired discussion tasks and other criterion measures correlate with each other?

3) What are the salient interactional features captured from the interlocutors' speech performance on paired discussion tasks? Are

paired discussion tasks effective in eliciting a wide range of interactional features that can be used to assess test takers' interactional competence?

4) What are the raters' general perceptions on the functioning and adequacy of the existing rating scale for interactional competence?

1.3 Organization of the Thesis

The current study is organized as follows. Chapter 2 provides an overview of previous studies on conceptualizations of speaking ability and interactional competence as assessment constructs, a major theoretical framework for L2 speaking tests, speaking tests formats for assessing interactional competence, task-based approaches to L2 speaking tests, and rating scales for interactional competence. Chapter 3 presents the methods and designs of data collection for the present study. Chapter 4 reports the descriptive statistics including raters' raw scores, inter rater and score reliability values, coefficients of correlations among tasks and measures, and other results of data analyses. Chapter 5 discusses the major findings of the current study in terms of their implications not only for assessment of interactional competence but also second language speaking assessment.

Lastly, Chapter 6 concludes the thesis with the summary of the findings, limitations and suggestions for further research.

Chapter 2 Literature Review

2.1 Speaking Ability and Interactional Competence

2.1.1 Defining Speaking Ability as a Construct

Speaking ability in its broad sense is defined to be the ability of an individual to produce language that is syntactically comprehensible, lexically task-appropriate, grammatically accurate, and displaying pronunciation in a manner that approximates how a native speaker speaks (Payne & Whitney, 2002). Fulcher (2003) defined the speaking ability to be the capacity to use oral language to communicate with others. In an academic setting, Jamieson, Eignor, Grabe, and Kunnan (2008) defined the speaking proficiency as the speakers' competence for "the use of oral language to interact directly and immediately with others for the purpose of engaging in, acquiring, transmitting, and demonstrating knowledge" (p.74). Based on these broad definitions of speaking ability, Ockey and Li (2015) specified four major components of oral communication ability: (a) interactional competence; (b) appropriate use of phonology; (c) appropriate and accurate use of vocabulary and grammar; and (d) appropriate fluency.

The four components of oral communication ability identified by Ockey and Li (2015) can be divided into two categories: linguistically

related constructs and socially grounded ones. First, the linguistic constructs refer not only to the appropriate use of phonology but also the appropriate and accurate use of vocabulary and grammar. On one hand, appropriate use of phonology is related to whether a language user has the ability to articulate words in an appropriate manner at both segmental and suprasegmental levels (Fulcher, 2003). However, in the context of the L2 speaking assessment, the focus is put on the overall comprehensibility rather than the accuracy of segmented sounds in that the purpose of communicating with others by using language is to create meaning in a particular context. Therefore, raters, in the process of scoring speech, usually make judgments about the test takers' speaking ability based on their task performance, and the scores are very likely to reflect a rater's perception on the degree of comprehensibility of a test taker's speech.

On the other hand, appropriate and accurate use of vocabulary relates to the breadth, size, and depth of vocabulary and in what manners the words are used, while appropriate and accurate use of grammar refers to not only the morphological and syntactical diversity but also the accuracy and effective use of grammar. However, although vocabulary and grammar have been treated as two distinctive dimensions of speaking ability, some research studies have shown a strong relationship between the scores on the two constructs when raters assign scores for them (Hunston, Francis, &

Manning, 1997). Skehan (2009) also pointed out that lexis should be integrated with fluency and grammatical complexity and accuracy. However, it has not been much investigated whether vocabulary and grammar in oral communication can be treated as separable components of linguistic knowledge in the context of using various types of test tasks.

Here, the concept of fluency involves the processing of all three components of linguistic knowledge discussed above, which can include lexis, grammar, and phonology. Appropriate fluency is defined as "rapid, smooth, accurate, lucid, and efficient translation of thought or communicative intention into language" (Lennon, 2000, p. 26, cited in Sato, 2014). A number of studies have indicated that the temporal and acoustic properties such as the number or length of syllables, pauses, corrections or repetitions are important components used as measures of fluency (Pinget, Bosker, Quené, & de Jong, 2014). However, there have been studies that re-conceptualize fluency as the speaker's automatic procedural skill and the listener's perceptual phenomenon. In other words, oral fluency is reconsidered to entail not only fast and automatic processing of utterance, but also the ability to retrieve necessary grammar and vocabulary items while using them in an appropriate manner (Derwing, Munro & Thomson, 2008).

However, fluency as a sub-construct (or sub-dimension) of speaking proficiency needs to be carefully defined according to task types. Sato (2014) investigated whether fluency is perceived differently in monologic and interactive tasks, respectively. The findings revealed that some features of fluency, such as pauses and turn-taking, were differently perceived by raters according to the condition of performance, lending support to previous studies that speech performance during peer interaction has to do with perceived fluency. In the monologic task, the raters seemed to consider test takers' pauses as inability to phrase words or formulate sentences. On the contrary, in the interactive task, pauses were differently perceived as in conjunction with the ability to manage and facilitate communication. In terms of turn-taking features in the interactive task, the raters related test takers' management of turns with the levels of their engagement in communication. In addition, scaffolding was regarded as one important concept that dominantly affected the raters' perception in the interactive task. With these fundamental differences, Sato (2014) named fluency in interactions as interactional oral fluency, suggesting that fluency and features of different task conditions are interwoven with each other.

Second, Ockey and Li (2015), viewing speaking ability from the socially grounded perspective, also introduced interactional competence as one of the components that constitutes speaking ability. For decades, many

researchers from the socially grounded perspective have strongly argued that spoken language itself is intrinsically social in nature (McNamara, 1997). That is, language is used as a powerful mediating mechanism in a communicative context where more than two interlocutors engage in interaction (Markee, 2000; Vygotsky, 1980; He & Young, 1998). In this regard, the notion of interactional competence emerged, and Young and He (1998) proposed the idea of the interactional competence that reflects interactional skills of turn and topic management. In the next section, the definition and components of interactional competence will be elucidated in depth.

2.1.2 Conceptualization of Interactional Competence

Interactional competence is the ability to engage in the dynamic process of co-constructing discourse (He & Young, 1998; Young, 2008). Its central concept is “co-construction” which was originally highlighted by Kramsch (1986) who insisted that the participants engaging in a communication have a distributed responsibility for co-constructing meaning during interaction. Kramsch also indicated that interlocutors share an internal context while interacting with others as putting oneself in the shoes of another interlocutor. This philosophical concept is called

“intersubjectivity”, and Kramsch recognized that interactional competence presupposes it. On the basis of this core concept of co-construction during interaction, He and Young (1998) proposed the term ‘interactional competence’ that has to do with skills of turn and topic management as an additional and distinguishable component of communicative competence proposed by Canale and Swain (1980).

The conceptualization of communicative competence for second language was firstly established by Canale and Swain (1980) on the basis of Hymes’ (1972) framework for communicative competence for first languages. However, in Canale and Swain’s (1980) framework, the distinction between linguistic knowledge and language use was absent. This was because the authors strongly believed that language use cannot be modeled since the actual language use not only underlies both memory and perceptual strategies which are non-specific factors of communication but also is limited due to general psychological constraints on communication. Consequently, communicative competence theorized by these scholars was recognized as a single trait that resided in individuals as becoming highly influential in the field of second language teaching and testing.

Reflecting upon these theories of communicative competence, interactional competence emerged as the ability to use oral language to interact with others. However, the preceded conceptualization on

communicative competence was too limited to involve interactional competence as one of its components since interactional competence is not regarded as a single trait that resides in individuals. In other words, interactional competence is not what an individual knows about language but what one does together with others in social contexts (Young, 2008). Focusing on this fundamental difference between the notions of communicative competence and interactional competence respectively, Young (2008, p.71) introduced seven resources that individuals bring to interaction.

1. Identity resources
    - Participation framework: the identities of all participants in an interaction, present or not, official or unofficial, ratified or unratified, and their footing or identities in the interaction
2. Linguistic resources
    - Register: the features of pronunciation, vocabulary, and grammar that typify a practice
    - Modes of meaning: the ways in which participants construct interpersonal, experiential, and textual meanings in a practice

3. Interactional resources
    - Speech acts: the selection of acts in a practice and their sequential organization
    - Turn-taking: how participants select the next speaker and how participants know when to end one turn and when to begin the next
4. Repair: the ways in which participants respond to interactional trouble in a given practice
5. Boundaries: the opening and closing acts of a practice that serve to distinguish a given practice from adjacent talk

As in Young's (2008) interactional resources, interactional competence involves the employment of both interlocutors' linguistic knowledge and resources related to speech acts, turn-taking, repairs, and the opening and closing acts during interaction. This conceptualization of interactional competence shows that the competence is distributed across all participating interlocutors in an interaction and is differently operationalized depending on varying interactional practices, which is fundamentally different from the notion of communicative competence (Young, 2011).

2.1.3 Interactional Competence as a (Sub) Construct

As discussed in Section 2.1.2, interactional competence emerged as a core and distinguishable dimension of communicative ability that is operationalized during interaction with others. Influenced by the shift in conceptualizing communicative competence, numerous researchers in L2 speaking assessment have viewed interlocutors' interactional competence as one essential aspect of their speaking proficiency. May (2009) argued that if speaking proficiency is, as defined by Butler, Eignor, Jones, McNamara, and Suomi (2000, p. 2), the capacity for "the use of oral language to interact directly and immediately with others," the co-construction of discourse is the core of successful interaction.

When it comes to assessing interactional competence, however, it is challenging for assessors to evaluate individual interlocutors' interactional competence since ratings assigned to each individual interlocutors based on the co-constructed single performance is jointly produced by more than two interlocutors in paired or group oral tests. This challenging aspect of assessing interactional competence has led to a great deal of research on identifying its components, resources operationalized during real performance tests, and other relevant factors that might affect ratings. In the

next Subsections 2.1.3.1 and 2.1.3.2, the findings of identification of interactional features at a macro- and micro-level respectively are reviewed.

2.1.3.1 Patterns of Interaction at a Macro-level

At a macro-level of interactional competence, interaction patterns and the quality of interaction have been investigated on the basis of two main characteristics of interaction patterns: mutuality and equality (Damon & Phelps, 1989; Storch, 2001, 2002). The mutuality indicates the degree of interlocutors' engagement of constructing shared ideas, and the equality refers to the dominance level of each of the interlocutors in their interaction. Based on the levels of mutuality and equality, respectively, Galaczi (2004) proposed four types of interaction patterns in a peer-peer speaking test: (a) collaborative talk; (b) asymmetric talk－passive speaker; (c) asymmetric talk－dominant speaker; and (d) parallel talk.

Figure 2.1 shows the four major types of interaction patterns that can occur in a dialogic communication situation, which were re-conceptualized on the basis of three representative interaction patterns, including collaborative, asymmetric, and parallel talk.

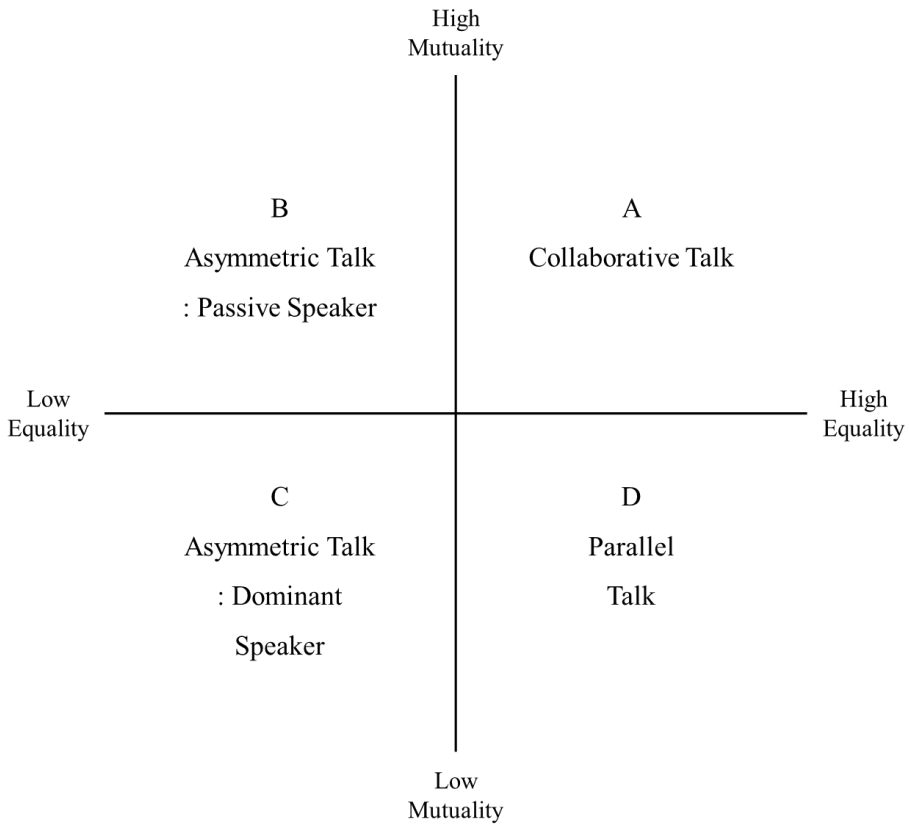

Figure 2.1 Three patterns of interaction in dimensions of mutuality and equality (Galaczi, 2004, p. 107)

First, Quadrant A indicates a collaborative talk with high mutuality and high equality where interlocutors work cooperatively and contribute to their interaction equally. When interlocutors' equality is low, it can be viewed as an asymmetric talk and further divided into two types of interaction patterns depending on the degree of mutuality. Quadrant B refers

to asymmetric talk with a passive interlocutor while Quadrant C represents an asymmetric interaction pattern with a dominant speaker. Lastly, Quadrant D illustrates a parallel talk in which both interlocutors do not actively respond to each other. That is, both interlocutors equally dominate their own discourse without responding appropriately to each other.

Based on the four representative patterns of interaction, Galaczi (2004) investigated the relationship between interaction patterns and speaking scores assigned by raters. Her findings revealed that interlocutors whose talk was collaborative obtained the highest mean scores while those who showed a parallel talk received the lowest average scores. The interlocutors who showed blended interaction patterns were assigned the moderate mean scores.

In order to examine raters' perceptions of salient interaction features, May (2009, 2011) collected and analyzed a range of qualitative evidence including rater notes, summary, verbal reports, and paired rater discussions for speech performance in the paired speaking test. From the collected data, she found that interactional features that belonged to a collaborative talk were positively perceived by raters.

All these taken together, the findings of the previous research show that the interactional features of a collaborative discourse influence raters' perception positively, and the raters assign high scores to collaborative

interlocutors. However, there have been few studies on how interactional competence is operationalized in testing contexts, thereby leaving many problems waiting to be resolved in the future studies. For example, Galaczi (2004) reported that raters felt difficulties in identifying the characteristics of the asymmetric interaction pattern. That is, since the discourse is co-constructed by interlocutors, it is difficult for raters to make clear boundaries on the impact of one interlocutor upon the other, which causes problems in assigning individual scores. With regard to the raters' perception, May (2011) also pointed out that further in-depth studies are needed to investigate raters' perception about operationalized interactional features as the first step in designing descriptors for the co-constructed discourse which can guide raters' fair judgments on interlocutors' interactional competence.

2.1.3.2 Interactional Features at a Micro-level

In recent years, there have been studies on interactional features observed in a real test situation, which were conducted to explore the nature of interactional competence (Ducasse & Brown, 2009; Gan & Davison, 2011; Jungheim, 2001). To identify the interactional features, Ducasse and Brown (2009) investigated teachers' and raters' views on the discourses elicited from the paired test. Based on the collected data, these researchers

identified interactional features and categorized them into three interaction patterns including non-verbal interpersonal communication, interactive listening, and interactional management. Table 2.1 lists the interactional features in the hypothesized categories.

First of all, non-verbal interpersonal features have been addressed as a natural aspect of communication (Ducasse & Brown, 2009; Jungheim, 2001). Jungheim (2001) investigated non-verbal interpersonal features such as head nods, gaze directions, gestures, and the relationship between these features and scores assigned by raters. In Jungheim's (2001) study, it is revealed that non-verbal scores and given ratings were correlated at a moderate level, but no meaningful association was captured between the two variables. Regarding the raters' perception, Ducasse and Brown (2009) found that raters positively perceived interlocutors' use of non-verbal features such as gaze and body language, but the relationship between operationalized interactional features that were salient to raters and speaking scores were not investigated in their study.

Table 2.1 Interaction features in hypothesized categories (Ducasse and Brown, 2009)

Parameters	Subcategories	Interactional features
Non-verbal interpersonal communication		▪ Gaze ▪ Body language
Interactive listening	Signaling comprehension	▪ Filing a silence ▪ Making comments ▪ Agreeing/ Disagreeing ▪ Correcting a mistake ▪ Offering or requesting clarification (prompt)
	Signaling support	▪ Back-channeling
Interactive management	Topic management	▪ Topic initiation ▪ Topic development ▪ Topic connection
	Turn management	▪ The number of turns ▪ Turn interruption ▪ Turn overlapping ▪ Length of turn
	Using Questions	▪ Agreement questions ▪ Confirmation questions ▪ Opinion questions ▪ Information questions ▪ Floor-offer questions

In addition to the non-verbal interpersonal features, Ducasse and Brown (2009) introduced the verbal interaction features that were divided into interactive listening and interactional management categories. First,

interactive listening refers to the interlocutors' manner of displaying their engagement of attention and demonstrating their comprehension while responding to other interlocutors. It comprises signaling comprehension and signaling support as two subcategories.

The first subcategory of interactive listening, signaling comprehension, consists of interactional features at a micro-level such as filling a silence, demonstrating comprehension, offering or requesting clarification (prompt), (dis)agreement and correcting a mistake. Filling a silence can be exemplified by a case in which the listener provides the word that the other partner is searching for, which indicates the listener's full comprehension. Demonstrating comprehension refers to the listener's relevant comments on the partner's contribution. The offering or requesting clarification feature indicates an interlocutor asking the partner to elaborate what she or he has already mentioned. (Dis)agreement represents when interlocutors disagree or agree with each other's idea. It should be noted that disagreement does not affect the perceived quality of interaction. That is, disagreement is one of the natural aspects of communication which does not negatively affect raters' perception. Correcting a mistake indicates the case that the listener helps out the speaker by correcting awkward or incorrect use of language.

The second subcategory of interactive listening is signaling support which includes back-channeling. Back-channeling refers to the listener's audible body language that encourages the other speaker to continue.

Second, interactional management, which is the second category of verbal interaction, is composed of three subcategories: topic management, turn-taking management, and using questions. The first subcategory, topic management, consists of topic development, topic connection, and closing a topic. Topic development indicates the interlocutors' expansion of a topic to further develop their discussion. Topic connection refers to one interlocutor's response based on the other interlocutor's idea, which helps to continue their interaction. Lastly, topic closure is the move to close a certain topic.

As the second subcategory of interactional management, turn-taking management consists of three features including turn length, turn speed, and turn domination. Turn length indicates the mean length of utterances spoken during communication. Turn speed refers to how quickly interlocutors respond with each other. Turn-domination has to do with how long one interlocutor takes floor during a conversation.

Using questions is the third subcategory of interactional management. Asking questions, confirmation questions, opinion questions,

information questions, and floor-offer questions are included in this subcategory.

Inspired by Ducasse and Brown's (2009) classification system for interactional features at a micro-level, a number of research studies have been conducted recently to examine a range of factors that might influence speech performance, such as interlocutors' proficiency, raters' perception, and task characteristics. Gan and Davison (2011) investigated how language learners in Hong Kong can engage in group interactions. The participants were divided into either a higher or lower proficiency group and took a group oral test which was implemented at that time in Hong Kong. The findings showed that the higher-scoring group used a wide range of speech functions including suggestions, (dis)agreement, explanation, challenges, and turn overlapping, which led participants to building mutually co-constructed discourse. On the other hand, from the lower-scoring group, interactional features of minimal participation such as minimal responses were displayed. The participants in the latter group showed tendencies of using more questions to deal with linguistic problems and most of the questions were confirmation checks and clarification questions.

Regarding raters' perception of operationalized interactional features, Galaczi (2013) identified several types of interactional features salient to the raters that showed differences among interlocutors at different levels of

proficiency: topic development and organization including the degree of topic development and topic extensions of 'own' vs. 'other' topics; listener support moves including back-channeling and confirmation of comprehension; turn-taking management in terms of no-gap-no-overlap manner; an overlap/latch; and a gap/pause. May (2011) also investigated raters' perception of salient interactional features in the paired speaking test. To identify what interactional features are salient to raters, May (2011) collected rater notes, summary statements, and verbal reports from four trained raters, and made them have a paired rater discussion on elicited speech performance. Then, collected data from raters were characterized into three categories including understanding interlocutors' message, ability to respond appropriately to interlocutor, and use of communicative strategies. Each category consists of raters' comments on operationalized interactional features which were perceived as either positive or negative evidence of interlocutors' interactional competence. For instance, 188 comments on the ability to respond appropriately to the partner were collected, and it was revealed that the raters positively perceived when one interlocutor precisely and coherently expressed one's ideas. On the other hand, when the interlocutors had difficulty expressing one's ideas clearly, the raters considered it to serve as a negative evidence of the interlocutors' interactional competence.

As one of the main factors that have impact on interaction, task-related factors have also been studied in terms of their influence on elicited speech performance. However, there have been few studies on the impact of varying forms and structures of interactive tasks in paired and group oral tests. Leaper and Riazi (2014), for example, examined the effects of different prompts on the features of turn-taking, syntactical complexity, accuracy, and fluency in the context of a paired discussion task. The authors found that the prompts requiring test takers to talk about their personal experiences yielded longer and more complex turns while the prompts with factual contents elicited shorter and less complex turns. The findings suggest the need of further investigation on task characteristics to reach better understanding of elicited interaction during L2 oral communication and to develop task-specific rating scales if needed.

To conclude, the interactional features at a micro-level have been studied to explore and conceptualize interactional competence as a construct to be assessed. In the review of previous studies, the features of interactional competence have been identified, and Ducasse and Brown (2009) identified the three categories of interactional features including non-verbal interpersonal communication, interactive listening, and interactional management. However, the deficiency of available studies in regards to this idea calls for the necessity of further research to provide a comprehensive

set of interactional features that can be displayed in varying forms and structures of interactive tasks and how elicited interactional features influence the quality and pattern of interaction and speaking (or interaction) scores respectively.

2.2 Theoretical Framework for L2 Speaking Tests

In the field of L2 speaking assessment, several speaking competence (and performance) models (Fulcher, 2003; McNamara, 1996; Skehan, 1998) have been proposed and expanded to reflect the increasing understanding of the constructs of speaking ability. Such attempts to develop new theoretical models or revise the existing models for speaking seem to be related to the expectation that the change in identifying and defining constructs of interest would make the score-based inferences more valid, which are also grounded in L2 learners' oral performances on performance tests. For decades, conceptualizing the constructs of speaking ability has been limited in that the focus has been heavily placed on structural aspects of language (or linguistic knowledge), such as pronunciation, grammar, and vocabulary. Researchers from socially-grounded perspectives, however, have strongly advocated the idea that speaking ability needs to be reconsidered with regard to the inherently social nature of spoken communication that can be

operationalized in language use while interacting with others (He & Young, 1998; Markee, 2000; McNamara, 1997). In other words, language should be seen as a powerful tool that enables language users to negotiate intended meanings when they communicate with others. With this shift in conceptualizing constructs of speaking ability, Fulcher (2003) expanded Skehan's (1998) model, which includes a number of factors, such as test takers' ability to use language, task characteristics, and rating scales.

Figure 2.2 presents Fulcher's (2003) model of speaking test performance that illustrates a number of factors interacting with each other on the basis of three main variables that affect test scores: (a) test taker's capacity on constructs to be assessed, (b) task implementation, and (c) rating. In Fulcher's model (Figure 2.2), these three main variables interact with each other, and each main variable consists of a number of factors. Many of these variables and factors have been researched in the field of performance assessment to explore the magnitude of influence of each variable on speaking performance and rating and to investigate the relationship among these variables in the context of speaking assessment.

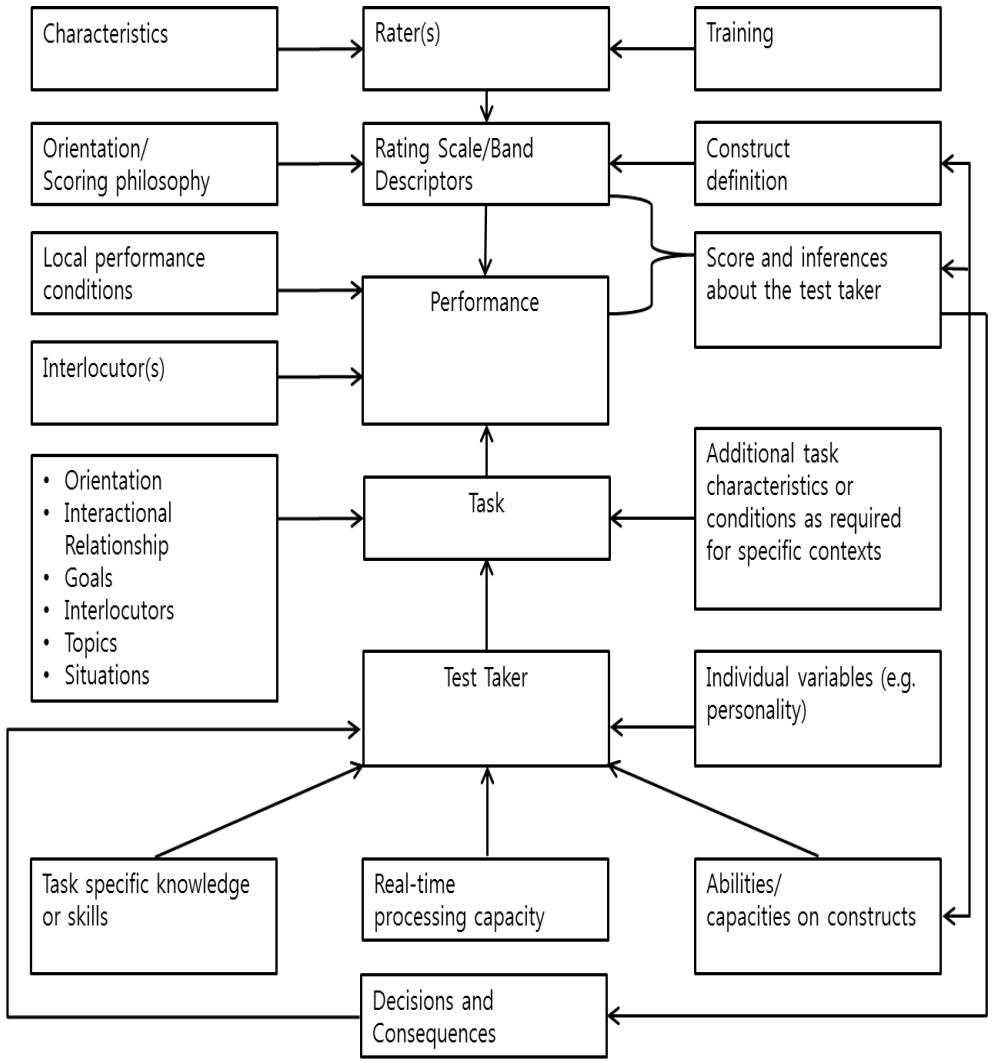

Figure 2.2 Fulcher's (2003, p. 115) expanded model of speaking test performance

2.3 Speaking Test Formats for Interactional Competence: Oral Proficiency Interviews and Paired Speaking Tests

To provide test takers with a real-conversation like testing situation, various types of interactive tasks have been used, which can include Oral Proficiency Interview (OPI), Semi-direct Oral Proficiency Interviews (SOPI), paired and group oral tests, and simulated tests.

As one of the most popular types of interactive tests, the OPI has been used for assessing oral communication ability. This type of a test involves two interlocutors, an interviewer and an interviewee. During the test, an interviewer poses topics and asks pre-determined questions to successfully elicit language to be rated. The use of the OPI for assessing oral communication ability, however, has been highly criticized due to its limitations in terms of its own characteristics different from that of real-life conversation. Specifically, it was revealed in Van Lier's (1989) study and other more recent research that there were significant differences between the OPI and realistic conversation, indicating that an examiner takes a dominant role over test takers in the OPI whereas interlocutors in real-life conversation have equal communicative rights and duties when engaging in the conversation (He & Young, 1998; Johnson, 2001). Moreover, Brown (2003) found that the choice of interviewer in the International English

Language Testing System (IELTS) test affected raters' perception of a test taker's speaking proficiency and thereby influencing scores assigned, creating potential unfairness for test takers.

The semi-direct oral proficiency (SOPI) interview is a procedure of elicitation that records test taker's performance for immediate or later retrieval to be evaluated by raters or computers. It became an alternative format of the OPI in that it is economical and efficient compared to the administration of the OPI. However, using the SOPI instead of the OPI has been controversial, with many concerns raised as to whether the OPI and the SOPI measure the same constructs or not (Shohamy, 1994, 1995; Luoma, 1997). In addition to the construct-related issues, the SOPI has been criticized due to its 'tape-like' discourse with a limited number of language functions (Shohamy, 1994).

Paired and group oral tests require at least two test takers to elicit interaction among them. As a way to address the limitations of the OPI and SOPI and find alternatives, numerous research studies have examined the potentials of paired and group oral tests and confirmed the vitality and usefulness of the paired and group oral tasks in eliciting discourse with a mutual contingency pattern (Brooks, 2009; Csépes, 2009; Kang & Wang, 2014). However, the paired and group oral tests do not always ensure interaction among interlocutors. According to He and Dai's (2006) research,

candidates were aware of testing situations where they had to prove their English ability. This resulted in a majority of candidates' performance with less negotiation of meaning and limited use of language functions. This finding indicated that test developers should provide test takers with properly designed tasks which encourage them to use a variety of language functions in the process of negotiating meaning. In addition, since the paired test format along with group oral tests involve more than one interlocutor, concerns have been raised regarding construct-irrelevant factors, such as interlocutor proficiency, personality, gender or familiarity that might decrease the validity of inferences made from test takers' performance (Bonk & Van Moere, 2004; Davis, 2009; Gan, 2010; O' Loughlin, 1995, 2002). Furthermore, some factors related to the characteristics of task such as task difficulty, task types, and task familiarity might have impact on the validity issue (Brooks, 2009; Ockey, 2014; Kang & Wang, 2014; O'Loughlin, 1995; Van Moere, 2013). On this issue, however, a limited amount of research has been conducted so far and few studies have been done on the effects of task-related variables on L2 interaction. Nakatsuhara (2010), for example, insisted that conditions of task implementation resulted in different interactional patterns by forcing candidates to take different interactional roles. Leaper and Riazi (2014) studied the effect of four different prompts on the group oral test and found that test takers' discourse

was different for each of the prompts although the difficulty of the four prompts was roughly the same.

In simulated tasks, test takers are required to assume a particular role in a particular context. For example, role-play tasks such as a making a dentist appointment or meeting with a buyer are examples of the simulated tasks. A teaching task to assess the teaching ability of international students as candidates of teaching assistants is another example. In addition, the simulated tasks are commonly used for occupational purposes, such as an oral communication test for air traffic controllers (Ockey & Li, 2015).

2.4 Task-based Approach to L2 Speaking Tests

Tasks have played a primary role in classroom language learning as teaching and learning tools. For pedagogical purposes, tasks are investigated and developed to be used in classroom as helping students to improve their proficiency and achieve final target goals of a course. Additionally, researchers in the field of Second Language Acquisition (SLA) have paid close attention to the role of tasks which allow them to explore developmental stages of L2 learners based on elicited performance from tasks. In the same vein, test tasks have been researched as one main variable consisting of the performance testing process. Bachman and Palmer (1996),

for example, acknowledged the impact of test tasks on test takers' oral performance and proposed a framework that involves components of task characteristics: setting, test rubrics, input, expected response, and relationship between input and response. Focusing on the cognitive demands that task characteristics impose on test takers, Skehan (1998) also emphasized that tasks play a crucial role in L2 oral assessment. He viewed communication as an information processing system which requires language users to meet cognitive and linguistic demands. In other words, in the context of speaking proficiency tests, the characteristics of test tasks determine both cognitive and linguistic demands which are required to complete a task and influence the quality of speech performance and test takers' speaking scores. Moreover, the fact that test developers can have full control over task variables suggests that much more efforts should be made on the studies of task characteristics in the field of L2 speaking assessment.

2.4.1 Task Types in the Paired Speaking Test

For a pedagogical or testing purpose, various types of tasks have been developed and used. The first time that different types of tasks for language teaching were used was in the Bangalore project using three principal task types: information-gap, reasoning-gap, and opinion-gap

(Nunan, 2004). The information-gap task involves a transfer of given information from one interlocutor to another, which encourages interlocutors to engage in interaction while decoding or encoding given information from or into language. The reasoning-gap task is described as the activity that requires interlocutors to use practical reasoning, inference or deduction while transferring and comprehending information. The opinion-gap task is an activity in which interlocutors identify and convey one's preference, ideas or attitude in a given situation. To argue or justify one's ideas, the last type of a task might involve the use of factual information or supporting resources. Therefore, no objective procedure for developing task performance is predetermined.

When it comes to task types in the context of the paired or group oral proficiency tests, little attention has been paid to the studies on the influence of task types that affects the quality of elicited speech performance. Foster and Tavakoli (2009), for example, examined how narrative task design affected EFL learners' speech performance in terms of their syntactic complexity, accuracy, fluency, and diversity. Through the comparison between two narrative tasks with different structures, loose or tight, the findings revealed that interlocutors who took a tight narrative structure task achieved greater accuracy in their speech performance than that achieved during a loose narrative task. In addition, syntactic complexity

was supported by narrative storyline complexity, and grammatical accuracy was supported by an inherently fixed narrative structure.

In line with research on task characteristics, not much research has been done on the impact of task complexity which is determined by task types and task difficulty. In Nuevo's (2006) research, simple versus complex versions of interactive tasks were analyzed in terms of interlocutors' learning opportunities. The results showed that the simple task led more interaction and clarification requests while the complex version of interactive task did not have great influence on accuracy of interlocutors' speech. In Michel, Kuiken, and Vedder's (2007) study, the complex and simple versions of tasks were investigated in two different conditions for L2 learners of Dutch: monologic condition and dialogic condition. The results of the dialogic setting showed that the complex version led to more accurate but less fluent production while linguistic complexity showed a marginal effect. The interaction between task complexity and task conditions showed a significant effect especially on measures of accuracy. The results of the research above can be viewed as clear evidence that shows the influence of task complexity depending on task types and difficulty on the change of the quality of discourse.

2.4.2 The Role of Interlocutors in the Paired Speaking Test

One of the challenging aspects of interactional competence assessment is assessing speech performance that is co-constructed by more than two interlocutors in a communicative context. Due to this challenge, researchers have raised many concerns about possible effects that interlocutor-related factors might have on the pattern or quality of co-constructed speech performance and thereby affecting rating outcomes. So far, the effects of these factors including gender, acquaintanceship, background knowledge, personality, the number of interlocutors, and different levels of proficiency have been investigated (Bonk & Van Moere, 2004; Davis, 2009; Gan, 2010; Iwashita, 1996; McNamara & Lumley, 1997; Ockey, 2009; O'Loughlin, 2002; Lumley & O'Sullivan, 2005; Nakatsuhara, 2011; Van Moere & Kobayashi, 2003).

In terms of the interlocutors' personalities, Bonk and Van Moere (2004) found that outgoing students regardless of their proficiency levels took a significant but small advantage over their shy peers. In Van Moere and Kobayashi (2003), more talkative interlocutors were assigned higher scores regardless of their proficiency levels. In Ockey's (2009) study, the levels of assertiveness of group members were examined, and the results showed that more assertive group members were awarded with higher

scores than expected. Regarding the degree of acquaintanceship between interlocutors, O' Sullivan (2002) found that Japanese EFL learners received more scores when they performed three different tasks with a friend. With regard to gender effect, Lumley and O'Sullivan (2005) showed that female interlocutors tended to obtain slightly higher scores but a significant difference was not captured. In addition, it is revealed that the bias for or against test takers of either gender was identified as significant in four test tasks out of 27.

Regarding the effect of interlocutors' proficiency, Iwashita (1996) found that the proficiency level of the paired interlocutors could affect the discourse by making it more symmetrical, but candidates were not assigned higher scores. However, when the interlocutors talked with candidates whose proficiency level was higher rather than lower, they felt more comfortable. On the contrary, in Bonk and Van Moere (2004) and Davis (2009), little influence on raw scores was captured when interlocutors were paired with either higher-proficiency or lower-proficiency candidates. However, in terms of the quality of the discourse, candidates at a lower proficiency level talked 35% more when they were paired with candidates at a higher proficiency level than when they were with the candidates at the same proficiency level. When more proficient candidates were grouped, their speech performance showed either primarily collaborative or

asymmetric interaction while a wide range of interactional features were used by candidates at a lower proficiency level.

To sum up, though the results of previous studies on the effect of interlocutor characteristics are not all consistent, it may be reasonable to think that the proficiency differences between the two interlocutors participating in paired oral tests may not significantly have influence on their speaking scores.

2.4.3 The Role of Topic in the Paired Speaking Test

In recent years, there has been a significant increase in research on the potential of using a paired or group speaking test format as L2 speaking proficiency test. In contrast, few studies have been done on various topics presented in test tasks despite ongoing claims that a lack of particular knowledge of a certain topic makes test takers difficult to perform. Lumley and O'Sullivan (2005), for example, examined the interaction between task topics and gender of audience. In the study, six topics were identified according to topic orientation including female, male, and neutral. Then, each task was designed to have different types of audience: male, female or group audience. The results suggest that particular tasks were favored by candidates of one gender with statistically significant bias, but tasks affected

individual candidates differently without clear conclusion on the role of the topic itself on interlocutors' elicited speech performance. Nakatsuhara (2006) also investigated the possible effects of task topics on the relationship between interlocutors' proficiency levels and their interaction patterns. Within a problem-solving discussion task, four different topics were used in the study: seven jobs in a hotel, six things that single men might want, two events that affect people's lives and three things necessary to cross the Arctic on foot. The results indicated that there was no statistically significant effect of four different topics on interlocutors' speech performance. On the other hand, Leaper and Riazi (2014) proposed contradictory findings. The two investigated the effect of four different prompts with different topics on group oral tests. To be specific, the authors attempted to examine the role of different types of tasks and combinations of questions on test takers' discourse. The topics of four different types of tasks included mobile, outdoor activities, comparison of a single and marriage life, and family. Each task was composed of a series of questions related to a given topic. The results of the study were analyzed in terms of syntactic complexity, accuracy, and fluency of test takers' speech performance. The findings revealed that the prompts of the mobile and outdoor activities elicited significantly shorter and less complex turns whereas interlocutors who took the other two prompts produced longer

conversation with more complex turns. However, this difference in the quality of speech performance did not affect the interlocutors' speaking scores.

2.5 Rating Scales for Interactional Competence

As illustrated in the Fulcher's (2003) expanded model of the speaking performance test, various factors are involved in the assessment of L2 speaking performance such as raters, rating scales, test tasks, and characteristics of test takers. Among these factors, constructing rating scales is of great importance in that rating scales provide an operational definition of a construct in the context of L2 assessment (Fulcher, 2003). In this regard, numerous studies have been administered to develop rating scales and examine their validity and reliability.

For decades, a number of rating scales have been derived from general theory of language abilities and criticized since the descriptions of criteria do not lead assessors to make valid judgments on test takers' language proficiency. That is, the theory-based rating scales include too arbitrary and inconsistent contents as the descriptors of constructs and do not reflect what actually happens in a real test situation (Fulcher, 1987; Knoch, 2007). To resolve this critical problem that the theory-based rating

scales have, many researchers have investigated test takers' performance that underlies constructs of interest and developed empirically supported rating scales.

Following this shift in constructing rating scales, those for interactional competence have been developed on the basis of interlocutors' actual performance elicited from interactive tasks (Fulcher, 1996; Fulcher, Davidson, & Kemp, 2011; Hirai & Koizumi, 2013; Nakatsuhara, 2007). Nakatsuhara (2007), for example, developed a rating scale for assessing English speaking skills of Japanese EFL learners. To design a rating scale, experienced school teachers were recruited and asked to review the existing rating scales for speaking tests. After reviewing, descriptors were selected and modified from the collected rating scales, and a new rating scale that had five criteria, including pronunciation and intonation, grammar, vocabulary, fluency, and interactive communication was created. Then, to examine its validity and reliability, the authors asked raters to assess speech performance samples from 42 students. The findings of the study showed that the new rating scale was highly reliable with valid descriptions that reflected students' speech performance.

Hirai and Koizumi (2013) proposed empirically derived rating scales for a story retelling test and compared three types of rating scales in terms of their valdity and usefulness. All the rating scales included "Communicative

Efficiency" as one of the criteria which indicates interactional competence in a communicative context. To construct score descriptors, students at different levels of proficiency were classified and then prominent features that distinguished the differing levels were identified. Based on the salient features to raters that differentiated each test takers' different proficiency levels, three types of empirically derived rating scales were developed: an empirically derived, binary-choice, boundary-definition (called EBB) rating scale 1, EBB2 as a modified version of EBB1, and a multiple-trait scale as a conventional analytic scale. The three rating scales were the same in that the descriptors of each rating scale were empirically supported, but these rating scales were different in terms of their wording and contents of descriptors and the number of criteria. The findings showed that the differences in the characteristics of the three rating scales resulted in remarkable difference in terms of the degree of their validity and relibility. Specifically, the modified rating scale, EBB2, with carefully defined descriptors on features that distinguished speaking proficiency at different levels was the most valid and reliable rating scale in the context of story retelling. Additionally, Hirai and Koizumi (2013) found that different topics of four tasks used in the study imposed certain influence on interlocutors' speech performance. This suggests that factors within the same task type such as information and materials that are provided for test takers should be carefully selected, and

rating scales should be modified on the basis of the task-specific characteristics.

Wang (2015) also examined the validity of three different types of rating scales for assessing interactional competence with using four different task types. In Wang's (2015) research, the interactional competence quality scale was newly proposed with two criteria: interaction patterns and task completion status. Then, descriptors of the same rating scale were slightly modified according to different task characteristics of the four tasks. In the study, it was revealed that the developed rating scale worked as a credible measure of assessing interactional competence, showing a close relationship with ratings. In addition, in line with Hirai and Koizumi's (2013) research findings, Wang's (2015) study found that difference in task type had impact on the distribution of patterns of interactional features.

In general, the previous studies on rating scales for interactional competence suggest that to construct rating scales, in-depth observations of performance data should be made, and then carefully illustrated descriptors that distinguish different levels of L2 speaking proficiency should be developed with using both qualitative and quantitative methods. Moreover, the influence of task characteristics on speech performance should be taken into consideration to reflect differences in terms of a range of interactional

features or the quality of interaction that can be affected by different types of tasks.

Chapter 3 Methods

3.1 Participants

3.1.1 Test takers

A total of 40 Korean students (20 pairs) were recruited and asked to participate in the current study (Male=10, Female=30). Through online postings and on-site promotions, the researcher recruited the participants whose TEPS (Test of English Proficiency developed by Seoul National University) scores were above 801 out of 990 or whose converted score from TOEFL® iBT or TOEIC® score belonged to this range (2 participants submitted their TOEFL® iBT scores and 4 participants TOEIC® scores). According to the descriptors of TEPS grades, participants whose TEPS scores range from 801 to 900 are considered to be learners at a near-native level of communicative competence, and participants whose TEPS scores range from 901 to 990 are at a native level of communicative competence. The participants' mean score of TEPS was 896.7, ranging from 811 to 967. The participants consisted of university students, graduate students, and graduates of Seoul National University. They were from various academic departments such as the department of Humanities, Social science,

Education, Law, Business administration, Technology, Engineering, Life science, and Food and nutrition. Half of the participants belonged to majors in the department of Humanities. Participants' age ranged from 19 to 35 years of age and the average age was 23.5 years old.

Table 3.1 Background information about the participants

Interlocutors	Male	Female	Total
Number	12	28	40
Age	19~29	19~35	23.5
A major area of departments	Humanities, engineering, science	Humanities, education, business administration, social science	
No. of students living in foreign countries	4	18	
Average of TEPS scores	886.2	900.1	896.7

3.1.2 Raters

Two native English speakers participated in the rating of the speech samples on interactional competence and the researcher served as a third rater in this rating session for adjudicating scores when the two native raters' scores were more than two scales apart. To assess interlocutors' general

speaking proficiency, one Korean ESL rater and the researcher participated in the rating process where a holistic rating scale was used. The obtained scores from the Korean rater and the researcher were used without any rearrangement. In terms of native English raters, the first rater (Rater 1) lived in the Philippines for 11 years and one year in the United States. She majored in Teaching English as Second Language (TESL) at her university and has 5 years of experience in teaching English to Korean students. The second rater (Rater 2) lived in Canada for 10 years. She also majored in TESL in her undergraduate study and is currently in a graduate program specializing in the same area. Rater 2 has taught English to Korean students for 6.5 years. The raters' role in this study was to assign scores to each test takers on the performance of each task with using the given rating scale for interactional competence and make comments on perceived interactional features. Before they participated in the rating of all performance samples of the 20 pairs, they received training from the researcher in order to understand the descriptors of the given rating scale in depth and practice the way to appropriately score each test taker for each task. For the rater training, two performance samples from a pair were selected and used for rating practice. These samples were excluded in the actual rating process. During the training, they discussed how to interpret the descriptors of the

given rating scale and exchanged opinions about them with each other. The average age of the two raters was 26, and they were both female.

The Korean rater (Rater 3) majored in English education at her university and specialized in English linguistics in her graduate program. Using the given rating scale for rating interlocutors' general speaking proficiency, Rater 3 was also trained with the two samples that the native raters used. During the training, the Korean rater and the researcher had a discussion session to talk about the way to score interlocutors on the basis of in-depth understanding on the descriptions of the rating scale and exchanged opinions with each other. The age of Rater 3 was 26.

3.1.3 Examiner

The researcher participated in the current study as an examiner by overseeing and managing the whole process of administrating the paired discussion test. Before two test takers of a pair started their discussion, the examiner gave them instruction about the amount of preparation and discussion time on a given prompt. Test takers' speaking performance on each task was video-recorded and no interruption occurred during a discussion by the examiner.

3.2 Instruments

This section explains the four types of instruments used in this study: two paired discussion tasks, rating scales for interactional competence and general speaking proficiency, and interactional features at micro and macro-levels.

3.2.1 Paired Discussion Tasks

According to Kim (2006), East Asian international graduate students faced the greatest difficulty leading class discussion and participating in whole-class discussion, which was viewed as one of the most important skills to be successful in academic settings. Reflecting the need of practicing discussion for students, paired discussion tasks were selected in the current study not only for exploring the nature of interactional competence in a paired discussion task, but also for providing opportunities to diagnose students' weaknesses in their interactional competence so that they can practice communicative skills required to lead discussion. Within the type of a paired discussion task, the components of tasks including goals, inputs, procedures, and authenticity of tasks related to the target language use (TLU) domain were considered (Bachman, 1990; Ellis, 2000; Nunan, 2004; Wang, 2015). As opinion exchange tasks, the paired discussion tasks used in

this study required interlocutors to engage in discussion through arguing their opinions and making negotiations. Unlike problem-solving or decision-making tasks, interlocutors in paired discussion tasks do not need to come to an agreement. It means that interlocutors have no clear goals or outcomes to be achieved through a discussion. However, interlocutors were required to fulfill what the given prompt instructed them to do during a discussion: 1) to exchange opinions with each other and 2) to provide specific reasons and examples to support one's own opinion. In terms of the authenticity of the tasks, two discussion topics in university setting were selected: choosing a roommate to live with in a dormitory and the necessity of lectures given in English in university. Considering that the majority of the participants in this study were students, both topics were appropriate to be discussed in university setting, which encouraged them to use authentic resources during the discussion, such as the current policies implemented in their university and their personal experiences. Therefore, interlocutors' performance will be evaluated in terms of whether they communicate with each other based on the negotiation of logical and reasonable ideas and providing supporting resources. During the test, interlocutors were given 2 minutes for preparation of each task and 7 minutes for each discussion. During the discussion, they were allowed to use a pen to jot down notes. The examiner read aloud the instruction in each prompt before letting the interlocutors

proceed in preparing for their discussion for 2 minutes. Figures 3.1 and 3.2 show the prompts of paired discussion tasks used in this study.

The first prompt briefly described two possible situations and instructs the interlocutors to choose one side. Given that the most of the interlocutors were students in university, it was assumed that the interlocutors felt comfortable and were familiar with the situation, where students have the possibility of living in a dormitory and may choose their roommates.

Task 1

Directions **It has been recently announced that dormitory rooms at university must be shared by two students. Would you rather have the university assign a student to share a room with you, or would you rather choose your own roommate?** Exchange opinions with your partner, and provide specific reasons and examples supporting your position on the announcement.

Figure 3.1 Paired Discussion Task 1

Task 2

Directions During the English discussion class, you and your partner are going to talk about the topic provided below.

"Lectures in university should be given in English to help students improve their global competence."

Exchange opinions with your partner, and provide specific reasons and examples supporting your position on the announcement.

Figure 3.2 Paired Discussion Task 2

The second task is to discuss the necessity of university lectures to be given in English to improve students' global competence. Lectures given in English have been discussed as a controversial issue in educational contexts in South Korea. The term 'global competence' has been widely used not only in educational settings, but also in everyday social contexts. Regarding the authenticity of the task, the two themes of the second topic, lectures given in English and global competence, seemed to be appropriate to be discussed in a university setting since English has been widely chosen and used as a medium of instruction in many universities in South Korea to

the effect that it can help university students develop their global competence.

3.2.2 Rating Scales

Each interlocutor's speech performance in paired discussion tasks was rated by two native English raters for interactional competence and the Korean rater for general speaking proficiency. As mentioned earlier, the researcher also took part in the rating processes. For assessing interlocutors' interactional competence, the interactional competence quality scale (the CAP scale) developed by Wang (2015) was used, and a holistic rubric was used for assessing general speaking proficiency.

Wang's (2015) CAP scale is a holistic scoring rubric for interactional competence and consisted of two main dimensions: interactional features at a macro-level and task completion status. Interactional features have been scrutinized as factors affecting the overall interaction quality. Galaczi (2008) distinguishes discourses of test taker dyads into "Collaborative," "Parallel," and "Asymmetric" patterns of interaction on the basis of "Mutuality," "Equality," and "Conversational dominance". With these distinctions on the global patterns of interaction, multiple research on interactional features has been conducted to explore the nature of interactional competence and

investigated the relationship between manifested interactional features and test scores for interactional competence (Brooks, 2009; Galaczi, 2008, 2014; Gan & Davison, 2011). In terms of the task completion status as the other crucial aspect in the assessment of task performance, completing a task is valued as a satisfactory indicator for successful performance, while the partial complete status of task is less valued (Bygate, Skehan, & Swain, 2001; Skehan, 1996). On the basis of two main dimensions of assessing paired speaking tasks, Wang (2015) developed 4 task-specific rating scales describing six levels of interaction quality and task completion status. Given that the paired discussion tasks used in the current study belong to the type of an opinion exchange task with no explicitly required task outcome, the CAP scale with the two major criteria of open task outcome and divergent negotiation results was used by two native raters (see Appendix C).

In terms of assessing general speaking proficiency, a holistic scoring rubric was designed for the current study. The scoring rubric is composed of five criteria: Pronunciation, Grammar, Vocabulary, Fluency, and Organization. The descriptors of five criteria at different levels is adapted from other descriptors of scoring scales used for the TOEIC® Speaking, TOEFL® Speaking, Council of Europe (2001), Luoma (2004) and Sato (2014). The raters considered the descriptions of five criteria while making

judgments on assessing interlocutors' general speaking proficiency. The total score was decided on the range of 1 to 7 (refer to Appendix B).

3.3 Procedure

This section explains the procedures of the study, including the process of recruiting and collecting data from the interlocutors, transcription and data collection from raters, labeling raters' perceived interactional features, and classification of their written reports on the features and the used rating scale.

3.3.1 Data Collection from Interlocutors

Participants were recruited through on-line postings and on-site promotions. The researcher was contacted by the participants that were willing to take part in this study voluntarily and the consent form of this study was sent to them afterwards. Once the participants fully understood the process and purposes of this study, they finally decided to participate in this study. Then, they were randomly assigned to a pair according to their available time schedule.

The collection of interlocutors' speech performance proceeded from November 29th to December 9th in 2016. As a pair of two interlocutors

arrived in a test room, they sat side by side and signed the informed consent form after the examiner briefly explained once again the purpose and procedures of the study. Then, the examiner handed out an examination sheet on which the prompt of the first task was written. The examiner read aloud the first prompt and the interlocutors were able to ask any questions related to the test for a little while before preparing for their discussion. They were allowed to use 2 minutes for preparation and 7 minutes for their discussion after greeting one another. The second task was conducted in the same method. Interlocutors' speech performance for each task was video-recorded using a SONY NEX-3N digital camera. The whole process of the test for each pair took an average of 25 minutes, ranging from 23 minutes to 27 minutes.

3.3.2 Data Transcription and Collection of Data from Raters

After collecting all the video-recorded speech performance from the 20 pairs of interlocutors, 4 transcribers, including the researcher, transcribed the collected data following orthographic conventions. After the transcription work was finished, the researcher checked all the transcription once again to identify and correct errors if there were any. Then, the transcription data were printed and bounded in a book format to be given

and used for investigating raters' perception of elicited interactional features during the test.

In order to complete the ratings of the recorded spoken responses for the interlocutors, two separate rating sessions were conducted as part of this study, one for the holistic rating of speaking proficiency and another for the rating of interactional competence. The researcher and one Korean ESL rater participated in the rating of the interlocutors' general speaking proficiency, while two native English speaker raters were recruited and trained to rate their interactional competence.

First, for the holistic rating of the interlocutors' performance, the two raters (i.e., the researcher and one Korean ESL rater) had a joint rater training session in which they familiarized themselves with a holistic rating scale developed for assessing general speaking proficiency and used it to score two selected speech performance samples for the practice run. During the rater training session, each of the two raters assigned a separate score to each interlocutor on each task, and had a question-and-discussion session. After having a detailed and extended discussion on the elicited speech performance and the rating scale used for the scoring, the two raters rated speech performance samples from 20 pairs of interlocutors, but did not include any written comments on their performance and the rating scale.

Second, for the assessment of interactional competence of the interlocutors, the two native English raters were trained on the basis of the rating scale to be used in the actual rating. The two speech performance samples were collected from a pair of interlocutors to train raters, but these samples were not included in the actual rating process. Before rating the samples for training, the researcher explained the rating scale in detail and had a question-and-answer session with the raters for some time. During the session, the raters were asked to focus on interlocutors' interactional skills and the task completion status as described in the given rating scale. Then, the raters watched the first sample and gave a score for each interlocutor. In the second step, the raters watched the same sample once more to write their comments on the operationalized interactional features from each interlocutor on the transcribed book prepared in advance. To be specific, the raters were required to use a '+' sign to indicate an interactional feature that was positively perceived by raters and '－' to indicate an interactional feature which worked as negative evidence in terms of interlocutors' interactional competence. With either a '+' or '－' sign, the raters were asked to write their comments specifically about their reasons why each interactional feature was considered as either positive or negative evidence of an interlocutor's interactional competence. To do this, the raters stopped

the video-recorded samples at necessary intervals whenever they noticed significant interactional features and wrote down relevant comments. After the raters wrote comments on prominent interactional features, they were asked to write other comments specifically about the given rating scale.

The ratings during the first phase and the written comments during the second phase were produced by the raters individually. After the two phases of rating were completed, the raters reflected upon the entire process of rating, including their ratings, written comments, and score level descriptors of the rating rubric, in order to have a deeper understanding of the whole process and to make more sensible interpretations of the interlocutor performance. When the raters' scores for the same interlocutor were different, the researcher asked them several questions about particular features of interaction that led them to deduct or assign certain points and encouraged them to discuss the rationales and reasoning for their score assignment with each other more specifically. During the discussion, it was revealed that each rater focused on various interactional features from different interaction categories on the basis of the level descriptors from the rating rubric. To resolve the discrepancies between the raters' conceptualization of the interactional competence and judgments on the levels of task completion status for the interlocutors, they re-discussed the score level descriptors from the rating rubric used in this study and various

textual components indicating task completion status in the interlocutor response until they reached an agreement. The same procedures were conducted for the second speech sample.

From December 28th, 2016 to January 9th, 2017 the two raters individually rated all the speech performance collected from the 20 pairs. Then, the researcher identified and labeled the types of interactional features perceived by both raters on the basis of their comments. The identification of interactional features was conducted on the basis of Ducasse and Brown's (2009) three parameters in interaction: non-verbal interpersonal communication, interactive listening, and interactional management. Figure 5 presents interactional features in the hypothesized categories.

As shown in Table 3.2 in the next section, each parameter of interaction included subcategories consisting of multiple interactional features at a micro-level. Signaling comprehension which was a subcategory of interactive listening included filling a silence, making comments, agreeing/disagreeing, correcting mistake, and offering or requesting clarification (prompt). Signaling support which was the other subcategory of interactive listening consisted of back-channeling. Topic management, turn-taking management, and using questions were the subcategories of the interactional management parameter. For the topic management subcategory, topic initiation, development, and connection were included as interactional

features at a micro-level. Turn-taking management consisted of the number of turns, turn interruption, and turn overlapping. Lastly, the using questions subcategory included agreement, confirmation, opinion, information, and floor offer features. For each interactional feature, a label in a parenthesis is presented for convenience.

3.4 Methods of Analysis

The raters' scores for each interlocutor for each task and the frequency accounts of interactional features perceived by the two raters were submitted into a Microsoft Excel 2010, respectively. To obtain descriptive statistics for the collected data, IBM SPSS (IBM, 2011) Statistics Version 20.0 was used. The descriptive statistics includes the interlocutors' scores for each task, the frequency accounts of interactional features perceived by raters, reliability coefficients between raters' rating, and correlation coefficients between tests respectively. In terms of the descriptive statistics, the scores for interlocutors' general speaking proficiency and interactional competence from the two paired discussion tasks are presented.

To answer the first research question as to the values of inter rater and score reliabilities of the raters' scores for test takers' interactional

competence, both inter rater reliability and task-score-based reliability coefficients were computed. The inter rater reliability (alpha) coefficient was obtained by applying the Spearman-Brown Prophecy Formula (Henning, 1987) on the correlation coefficient computed between the first and second ratings for each task. Similarly, the task-score-based reliability (alpha) coefficients were also computed by applying the same formula on the obtained correlation between the averaged ratings for the two discussion tasks.

To answer the second research question in terms of the relationship among paired discussion tasks and other criterion measures, correlation coefficients were computed. This was done to examine not only the relationships between the two paired discussion tasks but also their relationships with other speaking and language proficiency measures.

To answer the third research question as to which interactional features are salient in paired discussion tasks, the operationalized interactional features during the tasks were identified and labeled on the basis of raters' comments by the researcher following Ducasse and Brown's (2009) hypothesized categories of interactional features. After, the frequency accounts of the perceived interactional features were calculated. In addition, the raters' written comments were collected and organized depending on the categories of interactional features. Some comments on a

certain category were further specified to investigate the raters' perception in depth. While identifying the types of operationalized interactional features perceived by raters, it was revealed that some features identified in Ducasse and Brown (2009) were not noticed by the raters of this test whereas some new features grabbed the raters' attention. Table 3.2 shows the types of interactional features with labels. Interactional features in parentheses indicate that those were not acknowledged by the raters in this study, but listed in the hypothesized categories while bolded features indicate features perceived by raters, but not listed in the categories.

Table 3.2 List of interactional features perceived by raters in the paired discussion tasks (n=18)

Types of interactional features in the paired discussion tasks (n=18)	
Gaze	NVG
Filling a silence (Helping out)	LCF
Making relevant comments	LCC
Agreeing/Disagreeing	LCA
(Correcting a mistake)	-
Offering or requesting clarification (prompt)	LCR
Back-channeling	LCB
Topic initiation	MTA
Topic development	MTD
Topic connection	MTC
Closing a topic	MTT
Irrelevant topic	MTO
(The number of turns)	-
Turn interruption	MTI
(Turn overlapping)	-
The length of a turn	MTL
Agreement questions	MQA
Confirmation questions	MQC
Opinion questions	MQO
Information questions	MQI
Floor-offer/-**taking questions**	MQF

Chapter 4 Results

This section reports the results of the collected data, including descriptive statistics, inter rater reliability coefficients, correlation coefficients among tasks and other criteria measures, types and frequency accounts of interactional features, and summaries of rater comments.

4.1 Descriptive Statistics

Table 4.1 shows the means and standard deviations of TEPS, general speaking proficiency, and interaction scores. As shown in Section 3.1, the mean of the interlocutors' TEPS scores was 896.7 which is an indicator of a near-native level of communicative competence according to the descriptors of TEPS score bands (TEPS, 2017). TEPS scores range from 10 to 990 with 10 major proficiency bands. In terms of the interlocutors' general speaking proficiency, the mean score of 5.6 corresponds to the interlocutors' proficiency level assessed by TEPS as indicating their speaking ability is very good in overall. Compared to the level of general speaking proficiency, the interlocutors' interactional competence is judged to be at a lower level with a score of 4. On the rating scale for interactional competence, the score

of 4 indicates the asymmetric pattern of interaction where one interlocutor takes a dominant role while the other passively engages in discussion.

Table 4.1 Means and standard deviations of TEPS, general speaking proficiency, and interaction scores of interlocutors

	TEPS	Speaking proficiency	Interaction
MEAN	896.7	5.6	4.1
SD	46.16	1.23	1.25

Notes: 0-990 scores for TEPS;
0-7 rating bands for general speaking proficiency;
1-6 rating bands for interactional competence

Figures 4.1, 4.2, and 4.3 present the histogram of the interlocutors' TEPS, general speaking proficiency, and interaction scores with a distribution curve included in each of them. In Figure 4.1, the distribution curve is slightly skewed to the left, indicating that most of the interlocutors in this study are at an advanced level of English proficiency. In addition, because achieving a score between 801 and 990 was a prerequisite requirement for participating in this study, the distribution curve is not only a little bit skewed to the left but also shows a "truncated" distribution. This suggests that the interlocutor participants of this study represent only a

limited range of TEPS scores, particularly towards the higher, more proficient end.

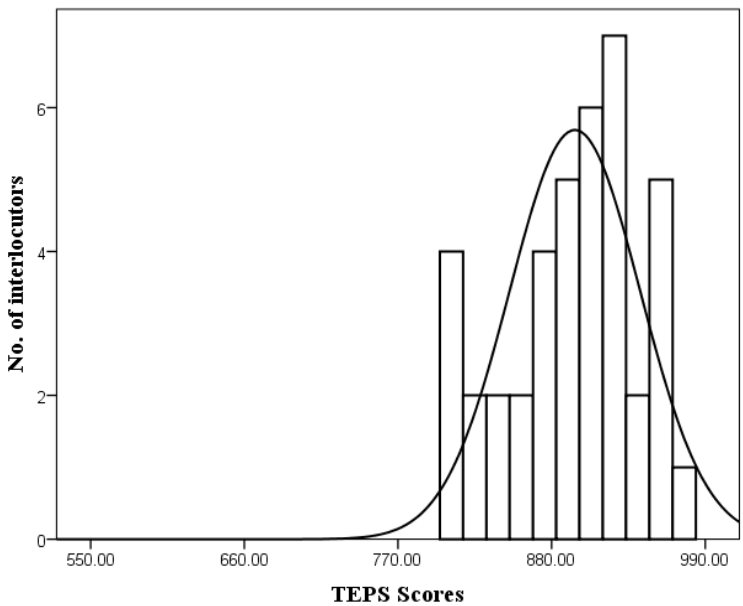

Figure 4.1 Histogram of interlocutors' TEPS scores

Figure 4.2 shows the histogram of the interlocutors' general speaking proficiency scores measured by the two paired discussion tasks. In Figure 4.2, the distribution curve is left-skewed. Considering that the scoring rubric for assessing general speaking proficiency consists of 8 subscales (0-7 scales), the mean score of 5.6 and the distribution curve in Figure 4.2 suggest that a majority of the interlocutors have a very good command of English.

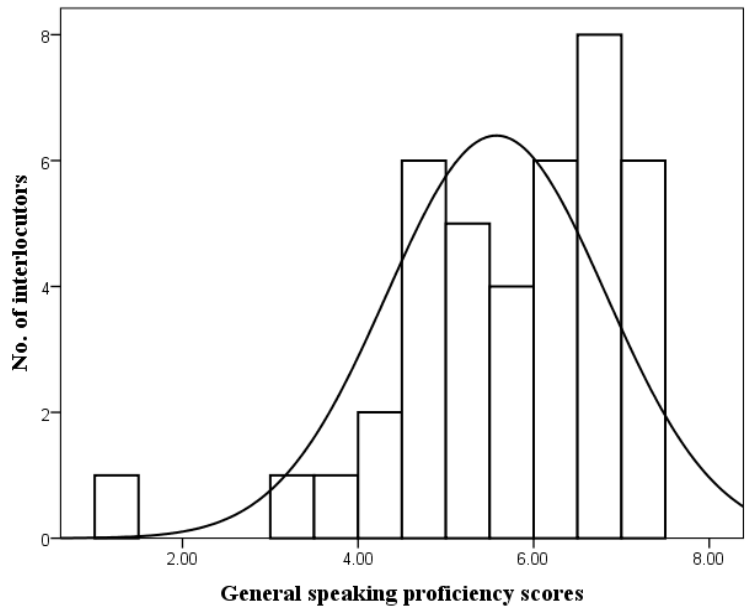

Figure 4.2 Histogram of interlocutors' general speaking proficiency scores

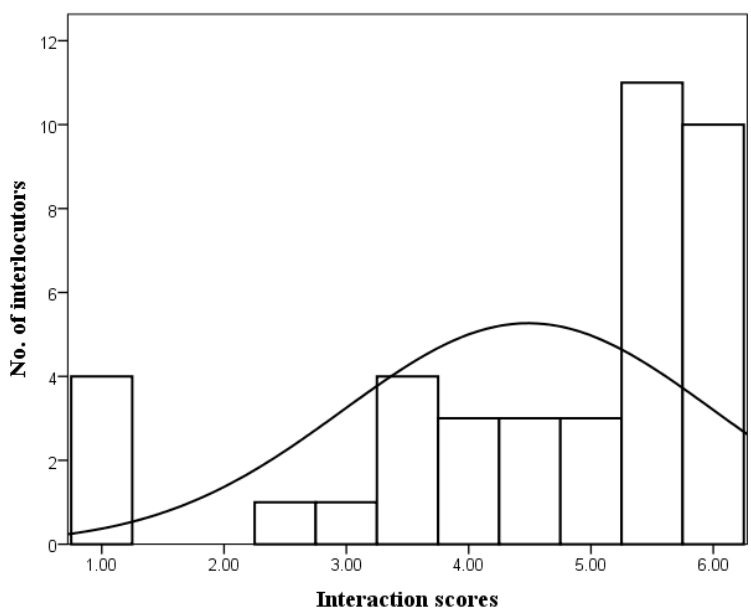

Figure 4.3 Histogram of interlocutors' interaction scores

Figure 4.3 presents the histogram of the interlocutors' interaction scores measured by the two paired tasks. As shown in Figure 4.3, the distribution curve is skewed to the left. In addition, more than half of the interlocutors are clustered on the extreme right and the other interlocutors are clustered all over the ranges of scores. This indicates that the levels of the interlocutors' interactional competence varied across the six levels.

Table 4.2 reports the descriptive statistics for the means and standard deviations of the interlocutors' general speaking proficiency scores. In Table 4.2, it is shown that Rater 3 was relatively severer in the rating compared to Rater 4. Moreover, both raters assigned more scores when they assessed the interlocutors' speech performances elicited from Task 2.

Table 4.2 Descriptive statistics for the scores of general speaking proficiency in the paired discussion tasks

	R3		**R4**		**Total**	
	M	**S**	**M**	**S**	**M**	**S**
T1	5.2	1.51	5.7	1.15	5.5	1.28
T2	5.4	1.45	6.0	1.18	5.7	1.26
Total	5.3	1.44	5.8	1.12	5.6	1.25

Table 4.3 Descriptive statistics for interaction scores in the paired discussion tasks

	R1		R2		Total	
	M	**S**	**M**	**S**	**M**	**S**
T1	3.7	1.82	5.1	1.44	4.4	1.48
T2	2.8	1.71	5	1.55	3.9	1.30
Total	3.2	1.43	5	1.47	4.1	1.27

Table 4.3 presents the descriptive statistics of the interlocutors' interaction scores measured by the two paired discussion tasks. As shown in Table 4.3, a considerable difference between Rater 1 and Rater 2 is captured as the total mean score is 3.2 for Rater 1 and 5 for Rater 2. This suggests that although raters used the same rating scale for interactional competence, their perspective on assessing the construct was highly different from each other, resulting in a substantial difference between their ratings. In terms of difference between the two tasks, interlocutors were assigned higher scores on their performances from Task 1 than Task 2.

4.2 Reliability Measures

This section reports various measures of inter rater and score reliability obtained for the general speaking proficiency and interaction

scores, which include correlations, exact/adjacent agreement rates, Kappa values, and reliability coefficients. First, the Spearman rank-order correlation and Pearson correlation tests are usually used to evaluate the null hypothesis that assumes there is no relationship between scores obtained from two raters. To be specific, the Spearman rank-order correlation test is a non-parametric procedure of Pearson correlation test, which is used to determine the strength and direction of the linear relationship between two ranked variables. The Pearson correlation test is a parametric procedure used to examine the strength and direction of the linear relationship that two continuous variables have (Rovai, Baker, & Ponton, 2013). Both of these coefficients are used here as measures of score consistency between raters. Second, Cohen's weighted Kappa coefficient (Cohen, 1968) is computed to provide a measure of agreement beyond expected agreement by chance and reflect the degree of raters' disagreement as putting greater weight on large differences between raters rather than small differences (Sim & Wright, 2005). This measure is considered to be a more valid index of score agreement between raters than an ordinary Kappa coefficient or simple score agreement rates. Lastly, the reliability coefficients are calculated to provide measures of score (rating) consistency across raters and tasks. This section will provide an overview of research results which will be further interpreted in the discussion section.

4.2.1 Inter rater Reliability

Table 4.4 Score agreement rates, Spearman rank-order and Pearson correlation, weighted Kappa, and reliability coefficients for general speaking proficiency scores

	Correlation		Agreement Rates				Wghtd. Kappa	A
	r_s	ρ	Per.	Adj.	Per. +Adj.	Non -adj.		
Task 1	.82**	.84**	.45	.43	.88	.13	.54	.91
Task 2	.80**[1]	.86**	.40	.53	.93	.08	.53	.92
Total	.85**	.89**	.35	.50	.85	.13	.57	.94

Notes: ** Significant at .01 level (two-tailed)

r_s=Spearman rank-order correlation coefficient

ρ=Pearson correlation coefficient

Table 4.4 reports agreement rates of scores, coefficients of Spearman rank-order and Pearson correlations, weighted Kappa, and reliability coefficients (alpha values) for interlocutors' general speaking proficiency scores. The coefficients of Spearman rank-order and Pearson correlation indicate that general speaking proficiency scores obtained from two raters

[1] Though it is not plausible to compare the magnitude of the coefficients obtained from the correlation tests, the discrepancy in the tendency between the measures of the Spearman rank-order and the Pearson correlations was captured. This is because the distributions of ratings assigned by each rater were different and there were some tied ranks due to the identical scores obtained from both raters (Putter, 1955; Rovai, Baker, & Ponton, 2013).

are highly correlated. Regarding Cohen's weighted Kappa, moderate strength of agreement is captured for all cases. In terms of the agreement rates, the values of perfect agreement rates indicate that raters' severity in rating was different (0.45 for Task 1, 0.40 for Task 2, and 0.35 for Total). Lastly, the reliability coefficients in Table 4.4 indicate that the use of the two paired discussion tasks for assessing general speaking proficiency increased the level of inter rater reliability.

Table 4.5 Score agreement rates, Spearman rank-order and Pearson correlation, weighted Kappa, and reliability coefficients for interaction scores

	Correlation		Agreement Rates				Wghtd. Kappa	α
	r_s	ρ	Per.	Adj.	Per. +Adj.	Non -adj.		
Task 1	.73** (.64**)	.89** (.61**)	.45 (.33)	.53 (.28)	.98 (.60)	.03 (.40)	.68 (.32)	.94 (.89)
Task 2	.65** (.30)	.85** (.27)	.38 (.15)	.58 (.35)	.95 (.50)	.05 (.50)	.60 (.11)	.92 (.85)
Total	.81** (.57**)	.94** (.53**)	.38 (.23)	.63 (.50)	1.0 (.73)	0.0 (.28)	.77 (.42)	.97 (.94)

Notes: ** Significant at .01 level (two-tailed)

r_s=Spearman rank-order correlation coefficient

ρ=Pearson correlation coefficient

( )=not considering the third rater's adjudication

Table 4.5 reports the measures of inter rater reliability of ratings for interactional competence. First, the Pearson correlation coefficients for Task 1 and Task 2 respectively indicate that there is a strong relationship between the scores obtained from different raters for each task. In terms of the Spearman rank-order correlation, the coefficients for Task 1 and Total indicate that raters' scores are highly correlated and the value for Task 2 is slightly lower as 0.65, indicating a moderate relationship between raters. Additionally, agreement rates show that perfect agreement rates for all cases are less than the adjacent agreement rates, suggesting that there was difference between raters' severity in rating. Regarding Cohen's weighted Kappa, substantial strength of agreement is captured for Task 1 and Total, and moderate strength of agreement is shown as 0.60 for Task 2. Lastly, the reliability coefficients in Table 4.5 indicate that using the two tasks for assessing interactional competence contributes to the increase of inter rater reliability.

4.3 Correlations among Tasks and Other Criterion Measures

Table 4.6 reports the coefficients of Pearson correlations among scores of interlocutors' TEPS, general speaking proficiency, and interaction. In Table 4.6, the correlation coefficient between interlocutors' TEPS scores

and scores for general speaking proficiency is 0.39 which indicates that these two variables are in a weak relationship. Between general speaking proficiency and interaction scores, the same strength of correlation is captured as 0.34. Considering correlations between scores for each task, however, only interaction scores for Task 1 show a significant relationship at a weak strength with general speaking proficiency scores from all cases. In addition, no correlation at a significant level was captured between interlocutors' interaction scores and TEPS scores for all cases. Lastly, coefficients of disattenuated correlation which assumes constant reliability are presented in italicized figures. The reliability information for TEPS scores was obtained by a TEPS staff (Lee, personal communication, March 7, 2017). After being corrected for attenuation, the values of coefficients increased, which suggest that when reliability among variables is controlled, the relationships among them become more highly correlated.

Table 4.6 Coefficients of Pearson correlations among scores for TEPS, general speaking proficiency and interaction

	TEPS	SP_T1	SP_T2	SP_Total	IC_T1	IC_T2	IC_Total
TEPS	1	*0.44*	*0.44*	*0.44*	*0.16*	*0.0*	*0.13*
SP_T1	.38*	1	*1*	*1*	*0.44*	*0.31*	*0.38*
SP_T2	.39*	.91**	1	*1*	*0.42*	*0.28*	*0.37*
SP_Total	.39*	.98**	.97**	1	*0.43*	*0.29*	*0.37*
IC_T1	.14	.38*	.37*	.38*	1	*0.71*	*0.96*
IC_T2	0.0	.26	.24	.25	.62**	1	*1*
IC_Total	.12	.34*	.33*	.34*	.88**	.89**	1

Notes: ** Significant at .01 level (two-tailed)
* Significant at .05 level (two-tailed)
SP: General speaking proficiency scores; IC: Interaction scores;
Italicized values: coefficients of disattenuated correlation

Figures 4.4, 4.5, and 4.6 represent how ratings from two different measures are associated by plotting data points. In Figure 4.4, the association between the interlocutors' total scores of general speaking proficiency and TEPS scores is shown. Between the two variables, a weak correlation with positive gradient was captured, but there was no linear relationship. The findings suggest that receptive areas of the interlocutors' English proficiency are positively correlated with productive areas of the language measured by the paired discussion tasks.

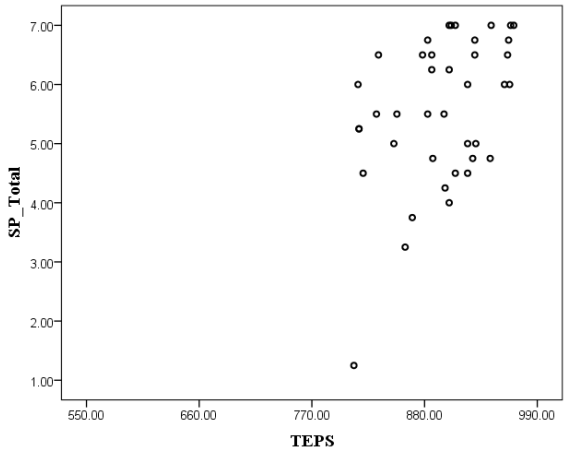

Figure 4.4 Scatter plot of total scores of general speaking proficiency and TEPS scores

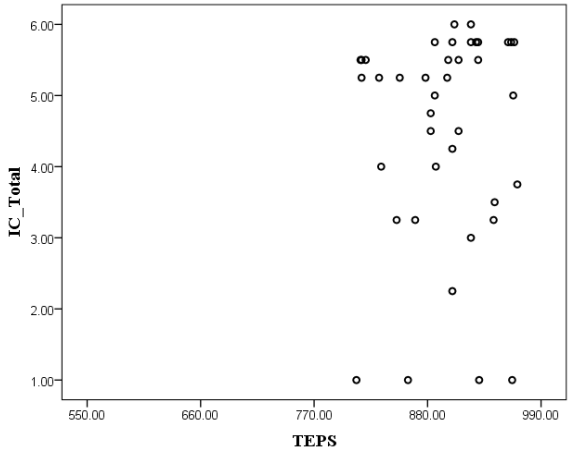

Figure 4.5 Scatter plot of total scores of interaction and TEPS scores

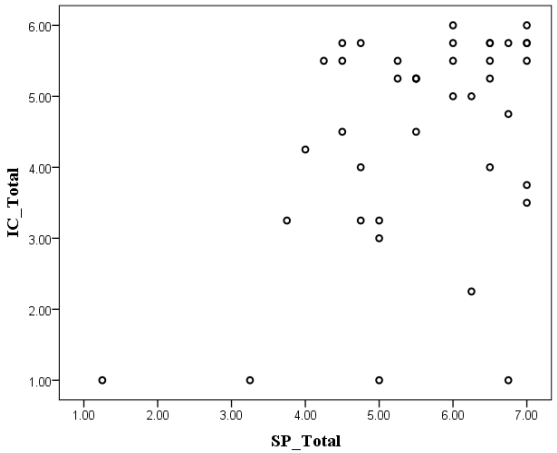

Figure 4.6 Scatter plot of total scores of interaction and general speaking proficiency

Figure 4.5 shows the association between TEPS scores and total scores of interaction during discussion tasks. In Figure 4.5, no association is captured between the two variables in either a positive or negative direction. This indicates that TEPS scores cannot be used as a predictable indicator of the interlocutors' interactional competence.

Figure 4.6 presents how the total scores of interaction and general speaking proficiency are associated with each other. As shown in Figure 4.6, there is no clear correlation or linear relationship between the two variables.

4.4 Frequency Analysis on Interactional Features

From Figures 4.7, 4.8, and 4.9 below, the line graphs of frequency accounts present the interactional features that were saliently perceived by raters. In each figure, the horizontal axis shows the labels of perceived interactional features with either a '+' sign or '－' sign and the vertical axis indicates the frequency accounts of individual interactional features perceived by raters in each task.

In Figure 4.7, the frequency accounts of perceived interactional features for each task show a similar pattern. Specifically, both raters prominently perceived topic-related interactional features, followed by

features of the interactive listening parameter in both tasks. In addition, features of using questions were least perceived by both raters for each task.

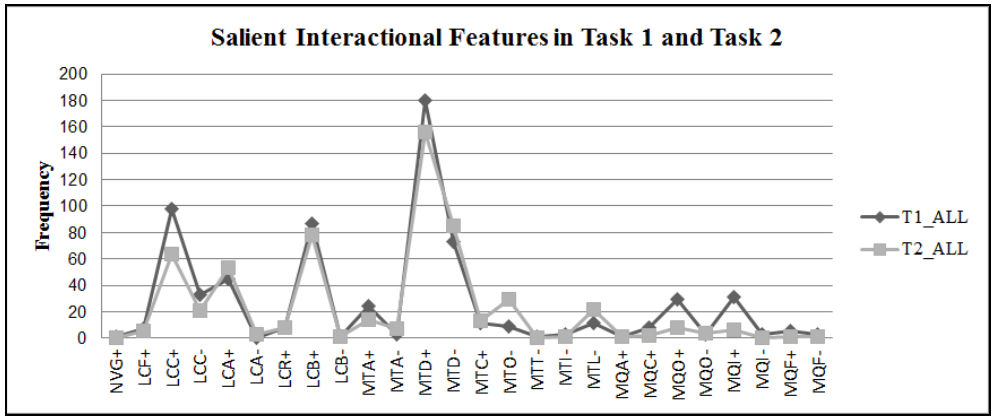

Figure 4.7 Salient interactional features perceived by raters for each task

In Figures 4.8 and 4.9, however, it is revealed that there are differences in the amounts of perceived interactional features between raters for each task. More detailed and specific descriptive statistics of the interactional features of each category of interaction patterns will be provided from Tables 4.7, 4.8, and 4.9.

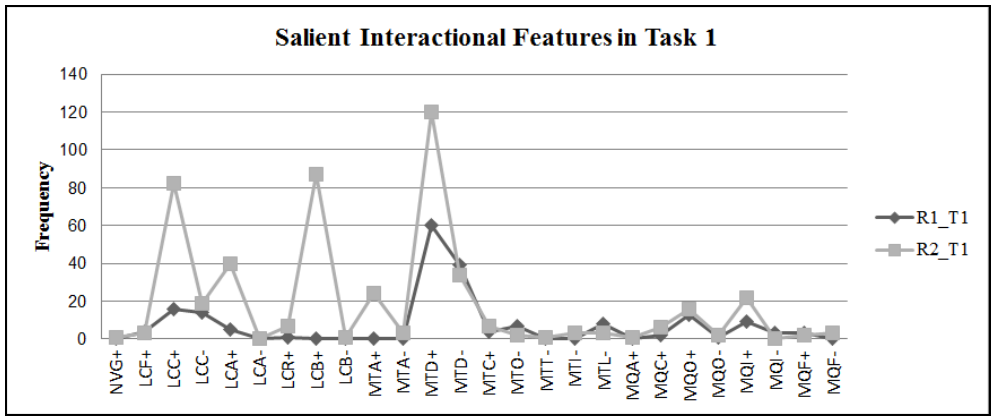

Figure 4.8 Salient interactional features perceived by each rater during Task 1

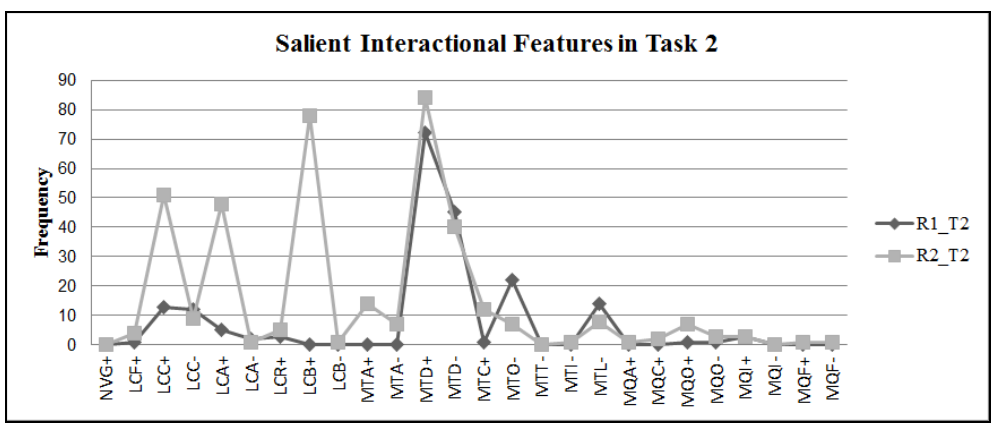

Figure 4.9 Salient interactional features perceived by each rater during Task 2

From Tables 4.7, 4.8, and 4.9, the types and frequency accounts of the interactional features saliently perceived by raters are reported. As mentioned in Subsection 3.5.2, an interactional feature perceived positively was described with a '+' sign while a feature perceived negatively with a '−' sign.

Table 4.7 Descriptive statistics for the perceived non-verbal interpersonal features in the paired discussion tasks

		T1				T2			
		R1		R2		R1		R2	
		Freq.	%	Freq.	%	Freq.	%	Freq.	%
Non-verbal interpersonal communication	NVG+	0	0.0	1	0.2	0	0.0	0	0.0
Total		0	0.0	1	0.2	0	0.0	0	0.0

Table 4.7 shows the descriptive statistics for the perceived non-verbal interpersonal features in the two discussion tasks. Though the rating scale used in the current study did not include the description of the non-verbal interpersonal features, Rater 2 made comments on one of non-verbal interpersonal features, gaze, while rating speech performances from Task 1. None of the non-verbal interpersonal features were captured in Task 2 by both raters.

Table 4.8 Descriptive statistics for the perceived interactional features of interactive listening in the paired discussion tasks

		T1				T2			
		R1		R2		R1		R2	
		Freq.	%	Freq.	%	Freq.	%	Freq.	%
Signaling comprehension	LCF+	4	2.1	3	0.6	1	0.5	4	1.0
	LCC+	16	8.5	82	16.8	13	6.7	51	13.1
	LCC-	14	7.4	19	3.9	12	6.2	9	2.3
	LCA+	5	2.6	40	8.2	5	2.6	48	12.4
	LCA-	0	0.0	0	0.0	2	1.0	1	0.3
	LCR+	1	0.5	7	1.4	3	1.5	5	1.3
	Total	40	21.2	151	30.9	36	18.5	118	30.4
Signaling Support	LCB+	0	0	87	17.8	0	0.0	78	20.1
	LCB-	0	0	1	0.2	0	0.0	1	0.3
	Total	0	0	88	18	0	0	79	20.4
Total		40	21.2	239	48.9	36	18.5	197	50.8

As mentioned above, Ducasse and Brown's (2009) three parameters of interaction patterns consisted of subcategories. In terms of the interactive listening parameter, features are divided into either a signaling comprehension or signaling support subcategory. With this distinction, Table 4.8 provides the descriptive statistics for the perceived interactional features of interactive listening in the two discussion tasks. In Table 4.8, it is shown that Rater 2 perceived more than about six times amount of interactional features of interactive listening (Frequency=239) than Rater 1

(Frequency=40) in Task 1. The percentage of the features suggests that Rater 2 put much more weight on the features of interactive listening as 48.9% while the percentage of the features perceived by Rater 1 is 21.2%. Regarding the features of signaling support in Task 1, no feature was perceived by Rater 1 whereas a feature of back-channeling was actively perceived by Rater 2 (18%).

Compared to the results from Task 1, Rater 1 perceived less amount of interactional features of interactive listening in the second task as 18.5% (21.2% for Task 1) while Rater 2 put nearly the same amount of focus on the features (50.8%) as she did in Task 1 (48.9%). With regard to the range of perceived interactional features, Rater 2 actively perceived interactional features of both signaling comprehension and signaling support subcategories in the two tasks. On the other hand, Rater 1 did not make any comment on the feature of a signaling support in both tasks.

Table 4.9 provides the descriptive statistics for the perceived interactional features of the third parameter of interaction, interactional management. Here, topic, turn-taking, and using questions are included as its three subcategories. During the rating process of Task 1, the features of interactional management were the most actively perceived among other features of the all parameters (78.8% for Rater 1 and 50.9% for Rater 2). Within the features of the interactional management parameter, topic-related

features were the most salient for both raters (58.2% for Rater 1 and 38.9% for Rater 2). Regarding the features of turn-taking, they took the least importance following the features of using questions for both raters.

During the rating process of Task 2, the topic-related features were the most frequently perceived by Rater 1 as 71.8% while Rater 2 put relatively less weight on these features as in Task 1 (38.9% for the first task and 42.3% for Task 2). In terms of the interactional features of turn-taking, Rater 1 had more focus on them as 7.2% compared to Task 1 (4.2%) followed by the interactional features of using questions (2.6%). Rater 2 less perceived the features of turn-taking and those of using questions as 2.3% and 4.6% respectively. As in Task 1, both raters highly perceived the interactional features of topic management in Task 2 compared to the features of the other parameters of interaction.

Table 4.9 Descriptive statistics for the perceived interactional features of interactional management in the paired discussion tasks

		T1				T2			
		R1		R2		R1		R2	
		Freq.	%	Freq.	%	Freq.	%	Freq.	%
Topic management	MTA+	0	0.0	24	4.9	0	0.0	14	3.6
	MTA-	0	0.0	3	0.6	0	0.0	7	1.8
	MTD+	60	31.7	120	24.5	72	36.9	84	21.6
	MTD-	39	20.6	34	7.0	45	23.1	40	10.3
	MTC+	4	2.1	7	1.4	1	0.5	12	3.1
	MTO-	7	3.7	2	0.4	22	11.3	7	1.8
	Total	110	58.2	190	38.9	140	71.8	164	42.3
Turn management	MTT-	0	0.0	1	0.2	0	0.0	0	0.0
	MTI-	0	0.0	3	0.6	0	0.0	1	0.3
	MTL-	8	4.2	3	0.6	14	7.2	8	2.1
	Total	8	4.2	7	1.4	14	7.2	9	2.3
Using questions	MQA+	0	0.0	1	0.2	0	0.0	1	0.3
	MQC+	2	1.1	6	1.2	0	0.0	2	0.5
	MQO+	13	6.9	16	3.3	1	0.5	7	1.8
	MQO-	1	0.5	2	0.4	1	0.5	3	0.8
	MQI+	9	4.8	22	4.5	3	1.5	3	0.8
	MQI-	3	1.6	0	0.0	0	0.0	0	0.0
	MQF+	3	1.6	2	0.4	0	0.0	1	0.3
	MQF-	0	0.0	3	0.6	0	0.0	1	0.3
	Total	31	16.4	52	10.6	5	2.6	18	4.6
Total		149	78.8	249	50.9	159	81.5	191	49.2

In summary, the descriptive statistics for perceived interactional features indicates the following findings. First, both raters put the greatest importance on the features of interactional management. Among the features of this parameter, the topic-related features were most saliently perceived by both raters in both tasks. In addition, the portion of the perceived topic-related features increased when two raters assessed speech performances from Task 2. In terms of the features of using questions, the percentage of this subcategory dramatically decreased from Task 1 to Task 2. Finally, two raters' different perspectives on assessing interactional competence were captured as they differently weighed on the interactional features from different subcategories. Specifically, Rater 2 nearly equally perceived the features of both interactive listening and interactional management parameters in the two tasks while Rater 1 took more importance on the features of interactional management rather than those of interactive listening.

Table 4.10 summarizes the perceived interactional features with a frequency more than 1 across tasks and raters. The bolded features are the ones that were most frequently and commonly captured by both raters as the main three features salient to raters in a format of discussion task regardless of topic difference.

Table 4.10 A summary of the most frequently perceived interactional features

	Interactional features
Task 1	**Making relevant comments +**
	Making relevant comments −
	Topic development +
	Topic development −
	Length of turn −
	Opinion questions +
	Information questions −
Task 2	**Making relevant comments +**
	Making relevant comments −
	Topic development +
	Topic development −
	Length of turn −

Note. Frequency in all tasks > 1

4.5 Qualitative Analysis of Interlocutors' Response

In this section, fine-grained analyses on interlocutors' response elicited from paired discussion tasks are reported. On the basis of raters' comments, perceived interactional features at a micro-level during test tasks are analyzed in depth, illustrating how interlocutors interacted with each other to complete the given tasks. Excerpt 1 below presents raters' comments related to a feature of topic development. In the transcription of

the interlocutors of Pair 2, it seems that the two interlocutors kept their conversation well as using informative questions and demonstrating their comprehension on what the partner has already mentioned. However, as shown in Excerpt 1, Rater 2 pointed out the interlocutors' informal talk which was irrelevant to the given topic prompt. This implies that more elaborated specification of topic-related interactional features on the basis of raters' perception of elicited speech performance is needed to reflect what the interlocutors actually performed during the test in order to better understand interactional competence in the context of a discussion.

Excerpt 1 from Pair 2 in Task 2

Transcription of interlocutors		**Rater 2's comments**
2A:	… Uhm, do you like your major?	*...but considering the task prompt, nature of a formal discussion, they've spent too much time on irrelevant topics. Although they backchannels, used good conversational strategy, the contents of their conversation did not complete the requirement.*
2B:	Um, yeah, pretty much, I'm not going to use it, but I like my major	
2A:	What field of education you are in?	
2B:	Uhm uh Geography education	
2A:	Uhm one of my friends from the department education, they are…	

Among interactional features categorized by Ducasse and Brown (2009), Table 4.11 identifies operationalized interactional features which can

be included as sub-features in a topic development feature. As the most saliently perceived feature among entire interactional features observed in this study, raters made a number of comments on the topic-related features which can be further categorized into the six subcategories: organization of ideas, use of supporting resources, consistency of arguments, topic relevance, consideration of both sides and negotiation of differing ideas. Table 4.11 shows the subdivided categories of topic development feature with raters' comments.

Table 4.11 Interactional features related to topic development

Features	Positive	Negative
Organizations of ideas	*Clear layout of one's view* *Logical organization of ideas* *Smooth transition in stating one's ideas* *Provision of the equal amount of input for the partner*	*Disorganized ideas/reasons/examples* *Awkward transition to the next idea* *No clear conclusion of one's thoughts*
Use of supporting resources	*Provision of specific reasons and examples* *Provision of appropriate personal examples* *Provision of relevant information* *Provision of reliable information* *Provision of possible solutions/suggestions/alternatives* *Further expansion/detailed explanation of one's opinion*	*Unable to elaborate on one's ideas* *Non-specific and unclear reasons/examples* *Unrelated ideas/solutions/alternatives* *The use of personal information rather than general information and knowledge* *Questionable source of examples with no reference* *Redundant/repetitive elaboration/expression* *Insufficient support*

		Reiteration of what the partner has already mentioned
Topic relevance	*Defining/clarifying the meaning of a term in the prompt* *Talking about the prompt as focusing on the given topic* *Raising awareness on the terms in the prompt and setting an agreed guideline to talk about* *Narrowing the scope of the prompt* *Backing on the given topic/prompt*	*Easily getting sidetracked from the prompt* *Unnecessary and disturbing small talks* *Off-topic comments that are inappropriate for a formal discussion* *Comments irrelevant to the given topic* *Moving too directly to modifying the prompt* *Clear unawareness of the prompt*
Consistency of arguments	*No side-tracked opinions* *Indicating reasonable rebuttal points*	*Vague stance on the given topic* *Neutral stance on the given topic* *Unmatched opinion with what one has already mentioned* *Support the opposite side* *Lost on track*
Consideration of both sides	*Considering both sides with explaining differing views* *Providing supportive ideas to the partner's suggestion* *Providing ideas for the opposite side*	*Strongly stick to one point of view and elaborate on it*
Negotiation of differing ideas	*Negotiation to come to an agreement* *Finding a consensus on a certain point in the prompt*	*Strongly insisting one's views without considering the partner's response* *Ignoring/shutting down the partner's idea* *Unable to reach a consensus*

Considering that the paired discussion tasks required the interlocutors to clearly and logically state their opinion with supporting details, raters focused on not only how well the interlocutors organized their

ideas without any awkward transitions or conclusions but also whether they appropriately used supporting resources to back up their own arguments. In addition, the interlocutors' consistency of arguments, balanced perspective, and how they used negotiating strategies to argue against the opposing ideas were taken into the consideration when assessing their interactional competence.

Table 4.12 Interactional features related to interaction at a macro level

Features	Positive	Negative
Attitude	*Eager/active to participate in the discussion* *Positive and open-minded attitude* *Setting a positive and comfortable environment to start a discussion* *Politeness (of asking a question)*	*Speaking in a quiet and passive way* *Triggering the partner to speak while being passive* *Unable to interact well* *Losing one's motivation to talk/negotiate with the partner* *Too defensive* *Lack of skills in attentive listening*
Management of time	–	*Unable to finish due to the lack of time* *Abrupt stop of discussion* *Awkward and poor time management* *Too many small talks during a limited time*

Table 4.12 above presents interactional features at a macro-level which were not addressed in both Ducasse and Brown (2009)'s illustration

of interactional features and the descriptions of rating scales used in this study. In the previous studies on interaction at a macro-level, researchers agreed to consider collaborative interaction as an ideal pattern in the context of communication. In this regard, Wang (2015) designed a rating scale for assessing interactional competence with two dimensions, pattern of interaction and task completion status, but there was no illustration of individual test takers' contribution to interaction which was distinctively perceived by raters in this study. To be more specific, raters perceived positively when a test taker contributed to making a comfortable environment to smoothly maintain their conversation or when they maintained a polite attitude while participating in discussions.

In addition, it is noteworthy that raters captured how well test takers managed their time given to discuss, which can be seen as one of essential aspects of leading a discussion. Although these interactional features at a macro-level are not the core components of interactional features, it is certain that the features of time management affected raters' perception as components that consist of a comprehensive set of interactional competence that needs to be assessed. Raters' comments on other interactional features will be provided in Appendix D.

4.6 Raters' Feedback on the Rating Scale for Interactional Competence

As mentioned in Section 4.5, raters perceived interactional features which were not described in the rating scales for assessing interactional competence and thereby causing problems concerning the validity of using the rating scale for the interactive tasks in this study. In this regard, raters' comments on the rating scale for interactional competence were collected and summarized. In the raters' comments on the used rating scale, pattern of interaction and task completion status were commented on by both raters regarding determining the priority between the two criteria. Additionally, raters made comments on multiple aspects of speech performance that should be addressed in the given rating scale. The researcher then classified the comments into five categories: prompt understanding, organization of discussion, use of supporting resources, topic relevance, management of discussion, and the contents of descriptors (See Appendix E). Excerpts 2 and 3 show one of raters' comments on the lack of levels of task completion status in the rating scale.

Excerpt 2 – Rater 1's Feedback on the Rating Scale

> *Interlocutors seem to partially understand the topic/prompt. Which category is it when the question was partially answered? Moreover, interlocutors easily get off topics. Good generation of questions and small talks, but mostly unrelated to the topic. How should these be rated and in which category? I think the scoring rubric makes very limited distinction between the two candidates. For example, word choice, cohesion, natural flow, etc. is not taken into consideration and the scoring rubric does not reflect how well the goal of the task was achieved.*

Excerpt 3 – Rater 2's Feedback on the Rating Scale

> *I think like what I've mentioned, there should be a degree of task completion for the participants. If I think about a black and white answer, I have to choose between a yes "task complete" or no "incomplete" or "partially" but this is very difficult to define. Having different sub features would help. (For example, 1) gave # of reasons, 2) gave support for reasons and 3) used examples)*

As presented in Excerpts 2 and 3, raters pointed out that more fine-grained levels of task completion status with its sub-features were needed to assess elicited speech performance. This limited aspect of the used rating scale caused a critical problem in that dichotomous distinction of the task completion status and the limited contents of description of either complete

or incomplete task status did not reflect what was actually performed by interlocutors and thereby lowering the validity of the obtained scores. In addition, as presented in Excerpt 2, raters argued that a wide range of interactional features and more detailed descriptors of them were required for the rating scale used in this study. These grave problems on the limited descriptions on the confined levels of criteria in the used rating scale were pointed out across all raters' comments (See Appendix E).

Chapter 5 Discussion

In this chapter, the major findings of this study are summarized and discussed in terms of the reliability of interaction scores, the relationships between scores from paired discussion tasks and other criterion measures, the types of salient interactional features identified in paired discussion tasks, raters' perception of interlocutors' response, and raters' feedback on the existing rating scale for interactional competence.

5.1 The Reliability of Interaction Scores and Their Relationships with Other Criterion Measures

5.1.1 The Reliability of Interaction Scores

The first research question posed for this study has to do with whether the interlocutors' interactional competence score from paired discussion tasks can achieve acceptable levels of inter rater and score reliabilities. First, various indices of inter rater reliability obtained in this study indicate that acceptable levels of inter rater reliability can be achieved for interaction scores. However, it should be noted that the inter rater reliability coefficients reached the acceptable levels only when the process

of the third rater's adjudication was employed. This might be partially due to the raters' different perspectives on the interlocutors' speech performance during the tasks.

As discussed in Sections 4.4 and 4.5, two native English raters' perspectives on rating interactional competence were quite different. The first major contributing factor that needs to be mentioned here is the lack of thorough and extended rater training. The raters in the current study received intensive rater training by using performance samples, but the duration of training was rather short-termed (about three hours). For this reason, the raters seemed to have great difficulty rating speech samples, which can be evidenced by several of their negative comments about the functioning and adequacy of the existing rating scale. In relation to this, another notable issue is that the two raters diverged in their view on different interactional features which were identified from the interlocutors' performance. In other words, the raters weighed various features from different interaction categories to justify their judgments on the interlocutors' interactional competence. This can also mean that the two different raters can have a different basis for the same scores assigned for the same interlocutors. Therefore, further studies are called for to investigate raters' perception on conceptualizing interactional competence operationalized in a particular communicative context, and much clearer and

elaborated construct definitions of interactional competence at varying levels need to be established and agreed upon among language testers. Based on the in-depth analyses of the raters' conceptualization and the definitions agreed-upon, a new rating scale should be constructed and used to guide raters in making valid and reliable judgments.

In terms of the differences between the measures of inter rater reliability for Task 1 and Task 2 respectively, the coefficients of inter rater reliability for Task 1 are relatively higher than that of Task 2. Although it is not statistically plausible to compare the calculated coefficients for the two discussion tasks, the different aspects of speech performance affected by different task topics can explain the tendencies of inter rater reliability for Task 1 having higher values. Specifically, considering the possibility that the topic of Task 2 contributed to more lengthy utterances with less turns and a relatively small number of perceived features of using questions than those during Task 1 (See Appendix F), the topic difference between the two tasks might affect the interlocutors' interaction hence making raters have different standards when evaluating interlocutors' interactional competence. This implies that when teachers or test developers create elicitation tasks, various factors of task characteristics that might change the quality of interaction should be identified and controlled prior to the carrying out of a task for making valid inferences of scores to be made.

5.1.2 The Relationships Between the Scores from Discussion Tasks and Other Criterion Measures

To answer the second research question, the relationships among interaction scores and other criterion measures, including general speaking proficiency scores and TEPS scores, were examined. First, it is reported that general speaking proficiency scores and TEPS scores were positively correlated at a weak level. This means that although TEPS scores indicate test takers' receptive areas of English proficiency such as listening and reading comprehension, they were turned out to be positively correlated with its productive dimension. On the other hand, interaction scores did not show any significant relationship with TEPS scores whereas a weak correlation between interaction scores and general speaking proficiency scores was found. Given that it is much more natural to assume that an ability to interact with others is relatively closely associated with speaking ability than the receptive ability of language, a low level of relationship between interaction scores and TEPS scores is not a surprising result. However, it should be clearly elucidated as to why a weak correlation between general speaking proficiency scores and interaction scores appeared.

First, components of constructs for general speaking proficiency and interactional competence were quite different. Interlocutors' speaking

proficiency was assessed in terms of individual interlocutors' pronunciation, use of grammar, vocabulary, fluency, and organization of contents. That is, raters held a linguistically grounded perspective to assess the speaking abilities of each test takers. On the other hand, to assess interlocutors' interactional competence, not only a linguistically grounded but also a socially grounded perspective was necessarily required since the elicited speech performance was co-constructed by two interlocutors through interaction between them. Therefore, beyond linguistic ability of language, much more focus was put on interlocutors' capability of how one led or was involved in a discussion on a particular topic while sharing and negotiating their ideas through interaction and using appropriate communicative strategies. This distinctive differences between general speaking proficiency and interactional competence as separately observable and assessable constructs might partially contribute to the low correlation between the two measures.

Second, as shown in Chapter 4, the test takers recruited for this study turned to be a truncated sample of EFL learners in terms of overall language proficiency, with their scores being highly skewed to the left. This means that their TEPS scores as the general indicator of English proficiency were not evenly distributed across all ranges of TEPS scores and thereby unable to show any clear and true relationship with either speaking proficiency or

interaction scores. Such a limitation in participant sampling for this study might have resulted in low correlations between interaction scores and other criterion measures. In addition, interlocutors' low motivation to engage in discussion tasks might also have affected their speech performance and obtained scores. Even though they were asked to actively participate in the given tasks, some test takers seemed to be shy of interacting with strangers and did not want to talk with their assigned partner. This passive attitude could have been perceived as negative evidence of interlocutors' interactional competence by raters.

Fourth, the low correlations might also be partially due to the different rater characteristics. In the current study, two native English raters took part in the rating process to assess interactional competence while two Korean ESL raters participated in the rating process for assessing general speaking proficiency. The effect of rater background variables on rating L2 performance has been heatedly debated for decades. For example, in Zhang and Elder's (2011) study, it was examined whether judgments of twenty non-native English teachers on thirty ESL test takers' oral English proficiency corresponded to those of nineteen native English teachers with using both quantitative and qualitative methodologies. The findings showed that there was no significant difference in holistic judgments of raters from the two different groups and their broad level of agreement on the speaking

proficiency. In this study, however, the collected raters' comments revealed that non-native and native raters seemed to put different weights on a range of interactional features to justify their decisions. Although different rater background variables have been viewed as a potential source of rating variation that can also affect the rating process, their magnitude of impact on resulting ratings do not seem to be statistically meaningful. In the current study, since two native English raters and two Korean ESL raters assessed two different speaking constructs, interactional competence and general speaking proficiency, with using different rating scales, different, it was not really possible to examine the effect of different rater characteristics (including the natives/non-native status) on the same construct thoroughly. In the future, it seems necessary to conduct more in-depth studies along this line using both quantitative and qualitative methods.

Lastly, the role of the rating scale for interactional competence matters. As briefly discussed in the previous section, there were problems in the existing rating scale regarding the lack of a wide range of interactional features displayed in the speech performance, the confined dimensions of a construct, and the limited levels of interaction patterns and task completion status. These clear limitations of the used rating scale might have not only confused raters but also made it difficult for them to form consistent and valid judgments on interlocutors' interactional competence.

5.2 Salient Interactional Features in Paired Discussion Tasks

5.2.1 Types of Interactional Features in Paired Discussion Tasks

To investigate whether the paired or group speaking tests have the potential to elicit better speech performance with abundant interactional features, numerous research has observed which types of interactional features appeared from such interactive tests (Brooks, 2009; Együd & Glover, 2001; Ikeda, 1999). In Brooks (2009), for example, 17 interactional features including seeking for information, asking a question, and clarification request were observed at a high frequency from a paired test. In addition to exploring features of interactional competence, Galaczi (2008), May (2009, 2011), and Ducasse and Brown (2009) studied how raters interpreted elicited interactional features as evidence of test takers' interactional competence. According to May (2009, 2011), raters tended to regard interlocutors' collaborative talk in a positive light, and similarly Gan (2009) also reported that the interlocutors who showed collaborative features were assigned more positive ratings. These findings are aligned with those of the current study. In the discussion tasks used in this study, 18 features of interactional competence such as features of making relevant comments and topic development were observed and perceived either positively or negatively by raters. Through investigating raters' perception

of observed interactional features, the need of deeper identification of a wide range of interactional features was captured in the context of a discussion task.

Furthermore, one of the raters made comments on an interlocutors' non-verbal interpersonal feature, gaze, which is not widely studied in the field of L2 speaking assessment. Although there was no description on non-verbal interpersonal features in the rating scale used in this study, one rater made a positive comment about the interlocutor that constantly faced the partner as to show support. On the other hand, features of correcting mistake, the number of turns, and turn overlapping that were frequently mentioned in the previous studies were not perceived by raters. These contradictory findings to those of previously conducted research studies show that it is necessary to investigate which interactional features are observed from various task types and whether observable features are fully described in a rating scale or not.

5.2.2 The Effect of Topic on Interlocutors' Response

With regard to the total frequency accounts of individual interactional features, similar patterns with differences were shown for two paired discussion tasks. Overall, raters mostly perceived topic-related

features the most, followed by features of interactive listening. The features of using questions were perceived as the least in both tasks. However, there were differences in terms of the frequency accounts between two tasks. First, the frequency accounts of features of topic management were higher in Task 2 than in Task 1. On the other hand, features of using questions were more frequently used in Task 1 than in Task 2. Considering that different topics were used for the two tasks, the findings in terms of the frequency of perceived interactional features suggest that even in the same type of task, topic as one of the variables of the task characteristics might influence the distribution of interactional features. The topic for Task 1 was related to living in a dormitory room with either a preassigned roommate or a roommate of one's own choice. To discuss one's own opinion, interlocutors needed to choose one option and clearly state their opinions with supporting reasons. Then, to figure out one's partners' point of view, interlocutors had to ask questions of opinion and information while making relevant comments. In addition, the number of turn and utterances indicate that interlocutors tended to make more turns which were relatively shorter than the turns made in Task 2 (refer to Appendix F). The topic of Task 2 was to discuss the necessity of university lectures to be given in English to improve students' global competence. Due to the intended vagueness of the definition of 'global competence' in the prompt, most of the test takers took

time to discuss and clarify the meaning of it. For the most part, it was once after the interlocutors reached an agreed definition, that they explained one's own opinion on the given prompt in a relatively longer turn, resulting in less use of question-related features to interact with the partner, compared to Task 1.

To conclude, the topic difference between tasks seemed to affect the distribution of interactional features to some degree. Such findings of the current study are contradictory to the results of Wang's (2015) research. In Wang (2015), four different types of tasks were used to examine any effects that task type might have on speech performance from intermediate-level interlocutors. The results showed that some interactional features were more frequently used in different speaking tasks. For example, in the decision-making task, the features of turn connection, agreement questions and opinion questions were frequently used. In a free discussion task, however, no features were significantly used. According to Wang (2015), these findings were due to the characteristics of the prompt for the discussion task which had no clearly-specified goals to be achieved by test takers. As a result, these characteristics would contribute to the reduction of pressure on test taker. However, the prompt used in this study explicitly instructed test takers to choose their position and exchange opinions with specific reasons and supporting examples. Even though there were no clear outcomes that

test takers were expected to meet, raters considered these elements for assessing the levels of task completion status. In addition, there were differences in terms of task topic and test takers' English proficiency between Wang's (2015) study and the current study. In Wang (2015), test takers at an intermediate level were asked to discuss their favorite electronic device and talk about its negative influence. They were given one minute for preparation and two and a half minutes for discussion. In the present study, test takers at an advanced level of English proficiency were asked to discuss two different topics related to a university setting. They were given three minutes for preparation and seven minutes for discussion to deliver their position on the given argument with specific supporting details. These differences in terms of test takers' English proficiency and characteristics of a discussion task, which includes given time, topic, and prompt, would contribute to the use of interactional features and interaction patterns. This is also in line with Leaper and Riazi's (2014) findings that showed the effect of the different prompts on interlocutors' speech performance during group oral tests. As discussed in the literature review, Leaper and Riazi (2014) indicated that the four prompts with different topics and questions had influence the length and complexity of turns and fluency. In terms of test takers' English proficiency, the finding of the present study is in line with the results of other studies conducted on how language learners at different

levels were able to engage in paired or group interactions. For example, Gan (2010) found out that test takers at a higher proficiency constructively and contingently participated in a discussion while using a wide range of speech functions such as (dis)agreement, explanation, competition for conversation floor, and turn overlapping. On the other hand, test takers at a lower level of proficiency showed peripheral participation, using minimal acknowledgement tokens and questions. In this regard, the findings of the current study imply that to develop elicitation tasks for both pedagogical and testing purposes, it should be kept in mind how multiple variables of a task influence test takers' speech performance and thereby affecting raters' perception and ratings.

5.2.3 Differences in Raters' Salience on Interactional Features

Based on the descriptive statistics of perceived interactional features, differences between raters in terms of their salience on interactional features were captured. First, although there was no description for non-verbal interpersonal features in the rating scale used for this study, Rater 2 positively perceived one of the features, gaze, in Task 1 while Rater 1 did not make any comments on non-verbal features. In terms of features of interactive listening, Rater 2 put much more focus on them in both tasks

than Rater 1 did. In addition, Rater 2 evenly perceived features of two subcategories of interactive listening: signaling comprehension and signaling support. On the other hand, Rater 1 made only a few comments on a requiring clarification feature of a signaling support subcategory. This suggests that salient features of interaction attended to by each of the raters can be different, and scores and scoring reasons given by each rater should be investigated with greater consideration. Lastly, within the category of interactional management features, topic-related features were the most frequently perceived by raters in both Task 1 and Task 2. Since the used task type was a free discussion, raters mainly focused how well interlocutors addressed and developed a given topic through interaction, resulting in putting considerable attention to topic-related features. In both tasks, making relevant comments to demonstrate comprehension, topic development, and length of turn features were saliently perceived (Frequency > 1) by both raters in both tasks.

5.3 Raters' Perception of Interlocutors' Response

The analysis of raters' perception of interlocutors' response has also provided in-depth insights into the need of elaborating interactional features to reflect actual speech performance in a better way. As discussed in Section

5.2, raters placed a great deal of importance on topic-related features to assess interactional competence in the context of paired discussion tasks. However, raters' comments showed that topic-related features, particularly a topic development feature, should be further specified into relevant subcategories including organization of ideas, use of supporting resources, consistency of arguments, topic relevance, consideration of both sides, and negotiation of differing ideas. This indicates that within the context of discussion tasks, topic development was linked not only to literally developing the move to extend a given topic but also to how test takers delivered one's ideas, what sources they used to effectively support their opinion, how well they discussed, how focused the interlocutors were on a given aspect of the assigned topic by consistently sticking to it, and how well they managed and organized their views on it. This finding resembles the results of Wang's (2015) study. According to Wang (2015), empirical evidence of four-factor model showed that interactional features were characterized into four communication functions, including argument, discussion, support, and connection. This result suggests that interactional features could also be grouped according to their communication functions, as done by Wang (2015), instead of the parameters and categories specified in the hypothesized model of interactional features proposed by Ducasse and Brown (2015).

In addition, raters perceived interactional features in terms of test takers' attitude of engaging in the discussion and time management at a macro level, which was not discussed in the previous studies. Specifically, raters positively perceived interlocutors' contribution in creating a comfortable environment, when maintaining a smooth flow of discussion or holding polite attitude as engaging in conversation. Regarding the features of time management, raters considered inability to finish discussion due to the lack of time, abrupt stop of discussion, or awkward and poor time management as negative evidence of interactional competence, suggesting that features of time management should be included in one of the essential aspects consisting of interactional competence as a construct.

5.4 Raters' Feedback on the Rating Scale for Interactional Competence

The present study also provides raters' feedback on the rating scale for interactional competence. First of all, the dichotomous levels of interaction patterns and task completion status in the existing rating scale were noticeably pointed out by both raters. Both raters claimed that the several layers of cooperation during the interaction between the interlocutors should be examined to appropriately assess elicited speech

performance, but the rating scale merely provided brief descriptions of three types of interaction patterns without differing levels. Furthermore, the raters argued that in a collaborative talk, individual test takers differently contributed to their co-constructed performance, particularly in terms of activeness of participation in the discussion, and therefore more sophisticated scoring system is required to evaluate each interlocutor's varying levels of contribution to the interaction. This result is contradictory to what May (2009) found in her study on the rater's perspective on a paired speaking test. According to May (2009), raters viewed some primary features of interaction as mutual achievements, which suggests awarding the shared single score to the interlocutors of a pair as reflecting the inherently co-constructed nature of interaction. However, the raters in the current study insisted that the different levels of contribution to the interaction and task completion status achieved by each interlocutor were separately perceived by raters distinctively, enabling them to assign scores for individual test takers. Nevertheless, this aspect of the raters' perception of the elicited speech performance was not addressed in the existing rating scale. In addition, it was difficult for raters to assign scores when features from different interaction patterns were overlapped, requiring a more elaborate scoring system.

In regards to the task completion status, the raters pointed out the lack of in-depth explanation in the existing rating scale as to how well the goals were met at differing levels. Moreover, the raters raised some concerns about putting different weights on interaction patterns and task completion status. In the given rating scale, a talk with an incomplete task status can get a higher score than that with a complete task status when the former had a better interaction. Regarding this issue, the raters raised a question whether interaction patterns were considered to be more valuable feature than task completion status.

In addition to these limitations, the raters argued that the rating scale used in this study might need to contain more descriptions of additional criteria, including prompt understanding, organization of discussion, use of supporting resources, topic relevance, and management of discussion. This suggests that in order to help raters in making more valid judgments on elicited speech performance, a wider range of interaction features from various aspects of a construct at differing levels should be considered, researched, and implemented into rating scales to be used in the future.

Chapter 6 Conclusion

This study aimed to investigate the feasibility of using paired discussion tasks for assessing interactional competence and examine the function and adequacy of the existing rating scale for interactional competence developed by Wang (2015). First, the results of the current study clearly show that the paired discussion tasks not only achieved the acceptable levels of inter rater reliability, but also elicited a wide variety of interactional features from the interlocutors. In this study, 18 interactional features were identified at different frequencies. Among them, making relevant comments, topic development, and the length of turn were the three primary features that were most saliently perceived by raters in both discussion tasks. In addition, through the qualitative analysis of the interlocutors' response based on raters' comments, the need for further sophisticated investigation of interactional features was revealed. For example, the feature of topic development as the most frequently perceived feature by raters was further specified into the six subcategories including organization of ideas, use of supporting resources, consistency of arguments, topic relevance, consideration of both sides, and negotiation of differing ideas. Considering that the tasks used in this study were in the format of a

discussion, raters assessed the interlocutors' interactional competence by focusing on how the topics of the given tasks were developed in appropriate manners through their interaction. In this regard, the existing rating scale manifests clear limitations in terms of its functioning and adequacy. Though a wide range of interactional features were observed in the two discussion tasks, the descriptors of the used rating scale did not fully reflect the actual speech performance. Moreover, raters' feedback on the rating scale indicated that they had great difficulty making consistent and valid judgments on the interlocutors' interactional competence, since the rating scale merely described the dichotomous levels of interaction patterns and task completion status, respectively. In addition, they showed great concerns about whether a collaborative pattern of interaction needed to take priority over task completion status.

From these major findings, some implications can be drawn in methodological and practical perspectives. First, to explore the nature of interactional competence, many researchers have made efforts to identify the types of interactional features through theoretical analysis or conducting experiments. However, the previous studies on this construct are limited in that there has been no fine-grained empirical analysis on operationalized interactional features elicited from varying forms and structures of interactive tasks. In this regard, the results of the current study empirically

examined that not only abundant interactional features were drawn as evidence of the interlocutors' interactional competence from a paired discussion task, but also the types and distribution of interactional features were different from those in other research. To be more specific, in the context of a paired discussion task, raters paid scrupulous attention particularly to the topic-related features as to consider how well the interlocutors addressed the given topics through their interaction. In addition, within the same task type, topic difference appeared to influence on the quality of interaction. Therefore, the current study provided proof to investigate interactional competence as a construct with using both quantitative and qualitative methods, reflecting the various task characteristics to understand the construct in a better way.

Second, the findings of the current study provides useful suggestions for developing rating scales for assessing interactional competence through paired discussion tasks. For teachers and assessors, it is recommended to develop a rating scale with several levels of criteria including interaction patterns and task completion status. In addition, the raters' feedback on the rating scale proved the need for dividing interaction patterns and task completion status as a distinctive criterion as neither of the two criteria can take on an added importance.

Lastly, developing an adequate rating scale is not the only way to assess interlocutors' interactional competence in a valid and reliable way. The qualitative analysis of the interlocutors' response in the current study demonstrated that raters have had different perspectives in interpreting the interlocutors' response as evidence of their ability to interact with others. Therefore, it is equally crucial that raters receive a thorough and intensive rating training with using an empirically supported rating scale that guides them to make consistent and valid judgments.

Despite the major implications of the findings in this study, there are some limitations that can be addressed in the future research. First, the current study only included the interlocutors at an advanced level as participants. This means that test takers of the current study are inevitably a truncated sample of EFL learners in terms of overall language proficiency. Therefore, it is recommended for future studies to consider participants at different levels of English speaking proficiency to investigate whether other EFL learners at varying levels will show different types and distribution of interactional features in a paired discussion task.

Second, the hypothesized categories of interaction patterns designed by Ducasse & Brown (2009) were used to identify elicited interactional features. However, there were some features that were not listed in the categories but captured in the current study whereas some features listed in

the categories were not all used. Furthermore, some features can be further specified into distinctively subdivided categories. Since the current study used the interactive tasks in the format of a discussion, only the topic-related features were highlighted and analyzed into subcategories in depth. Therefore, future studies can investigate more elaborated and various interactional features by including different types of interactive tasks.

Lastly, the inter rater reliability needs to be further improved through a more thorough and intensive rating training with using a valid rating scale. As discussed in this study, the inter rater reliability without the third rater's adjudication reached a relatively lower level of inter rater reliability. Therefore, through the effective use of valid and credible rating scale by trained raters, the level of inter rater reliability should be increased.

To conclude, future studies will enrich and deepen the understanding of interactional competence as these limitations are considered. There is also helpful prospect of further research leading to successful practice of assessing interactional competence in the context of L2 paired speaking assessment.

References

Bachman, L. F., & Savignon, S. J. (1986). The evaluation of communicative language proficiency: A critique of the ACTFL oral interview. *The modern language journal, 70*(4), 380-390.

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford: Oxford University Press.

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

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

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

Brown, A. (2003). Interviewer variation and the co-construction of speaking proficiency. *Language Testing, 20*, 1–25.

Butler, F. A., Eignor, D., Jones, S., McNamara, T., & Suomi, B. K. (2000). *TOEFL 2000 Speaking Framework: A Working Paper* (TOEFL Monograph No. MS-20). Princeton, NJ: ETS.

Bygate, M., Skehan, P., & Swain, M. (Eds.). (2001). *Researching pedagogic task*. Harlow: Pearson Education.

Canale, M. (1983). From communicative competence to communicative language pedagogy. *Language and communication, 1*(1), 1-47.

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

Cohen, J. (1968). Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin, 70*(4), 213-220.

Council of Europe (2001) Common European Framework of Reference for Languages: *Learning, Teaching, Assessment*. Cambridge: Cambridge University Press.

Csépes, I. (2009). *Measuring oral proficiency through paired-task performance.* New York: Peter Lang.

Damon, W., & Phelps, E. (1989). Critical distinctions among three approaches to peer education. *International Journal of Educational Research*, *58*, 9–19.

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

Derwing, T., Munro, M., & Thomson, R. (2008). A longitudinal study of ESL learners' fluency and comprehensibility development. *Applied Linguistics 29*, 359–380.

Ducasse, A. M., & Brown, A. (2009). Assessing paired orals: raters' orientation to interaction. *Language Testing*, *26*(3), 423-443.

Együd, G., & Glover, P. (2001). Oral testing in pairs: a secondary school perspective. *ELT Journal*, *55*(1), 70-76.

Ellis, R. (2000). Task-based research and language pedagogy. *Language Teaching Research*, *4*(3), 193−220.

Foster, P., & Tavakoli, P. (2009). Native speakers and task performance: comparing effects on complexity, fluency and lexical diversity. *Language Learning*, *59*(4), 866–96.

Fulcher, G. (1987). Tests of oral performance: The need for data-based criteria. *ELT Journal*, *41*(4), 287-291.

Fulcher, G. (1995). Variable competence in second language acquisition: a problem for research methodology? *System*, *23*(1), 25-33.

Fulcher, G. (1996). Does thick description lead to smart tests? A data-based approach to rating scale construction. *Language Testing*, *13*(2), 208–238.

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

Fulcher, G., Davidson, F., & Kemp, J. (2011). Effective rating scale development for speaking tests: Performance decision trees. *Language Testing*, *28*(1), 5-29.

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

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

Galaczi, E. D. (2013). Interactional competence across proficiency levels: How do learners manage interaction in paired speaking tests? *Applied Linguistics*, *35*(5), 553-574.

Gan, Z. (2010). Interaction in group oral assessment: A case study of higher- and lower- scoring students. *Language Testing*, *27*, 585-602.

Gan, Z., & Davison, C. (2011). Gestural behavior in group oral assessment: A case study of higher- and lower-scoring students. *International Journal of Applied Linguistics*, *21*(1), 94–120.

He, L., & Dai, Y. (2006). A corpus-based investigation into the validity of the CET-SET group discussion. *Language Testing*. *23*(3), 370–401.

He, A. W., & Young, R. (1998). Language proficiency interviews: A discourse approach. In Young, R., & He, A. W. (Eds.), *Talking and testing: Discourse approaches to the assessment of oral proficiency* (pp. 1-24). Amsterdam: John Benjamins.

Henning, G. (1987). *A guide to language testing: Development, evaluation, research*. Rowley, MA: Newbury House.

Hirai, A., & Koizumi, R. (2013). Validation of Empirically Derived Rating Scales for a Story Retelling Speaking Test. *Language Assessment Quarterly*, *10*(4), 398–422.

Hunston, S., Francis, G., & Manning, E. (1997). Grammar and vocabulary: Showing the connections. *ELT Journal*, *51*(3), 208-216.

Hymes, D. (1972). On communicative competence. In Pride J. B., & Holmes, J. (Eds.), *Sociolinguistics* (pp. 269-293). Harmondsworth: Penguin.

IBM (2011). IBM SPSS Statistics Brief Guide. Armonk, New York: Author.

Ikeda, K. (1998). The paired learner interview: A preliminary investigation applying Vygotskian insights. *Language, Culture, and Curriculum*, *11*(1), 71–96.

Iwashita, N. (1996). The validity of the paired interview format in oral performance assessment. *Melbourne Papers in Language Testing*, *5*, 1–65.

Jamieson, J., Eignor, D. R., Grabe, W., & Kunnan, A. J. (2008). Frameworks for a new TOEFL. In Chapelle, C. A., Jamieson, J., & Enright, M. K. (Eds.), *Building a validity argument for the Test of English as a Foreign Language* (pp. 55–95). New York: Routledge.

Johnson, M. (2001). *The art of non-conversation: A re-examination of the validity of the oral proficiency interview*. New Haven, CT: Yale University Press.

Jungheim, N. O. (2001). The unspoken element of communicative competence: Evaluating language learners' nonverbal behavior. In Hudson, T., & Brown, J. D. (Eds.), *A focus on language test development: Expanding the language proficiency construct across a variety of tests* (pp. 1-34). Honolulu: University of Hawai'i, Second Language Teaching and Curriculum Center.

Kang, O., & Wang, L. (2014). Impact of different task types on candidates' speaking performances and interactive features that distinguish between CEFR levels. *Research Notes 57*, 40–49.

Kim, S. (2006). Academic oral communication needs of East Asian international graduate students in non-science and non-engineering fields. *English for Specific Purposes*, *25*(4), 479-489.

Knoch, U. (2007). Do empirically developed rating scales function differently to conventional rating scales for academic writing? *Spaan Fellow Working Papers in Second or Foreign Language Assessment*, *5*, 1–36.

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

Leaper, D. A. & Riazi, M. (2014). The influence of prompt on group oral tests. *Language Testing*, *31*(2), 177–204.

Lennon, P. (1990). Investigating fluency in EFL: A quantitative approach. *Language Learning*, *3*, 387–417.

Lumley, T., & O'Sullivan, B. (2005). The effect of test-taker gender, audience and topic on task performance in tape-mediated assessment of speaking. *Language Testing*, *22*(4), 415–437.

Luoma, S. (1997). *Comparability of a tape-mediated and a face-to-face test of speaking: a triangulation study* (Unpublished Licentiate Thesis). Centre for Applied Language Studies, Jyväskylä University, Finland.

Luoma, S. (2004). *Assessing speaking*. Cambridge, UK: Cambridge University Press.

Markee, N.P. (2000). *Conversation analysis*. Mahwah, NJ: Lawrence Erlbaum.

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

May, L. (2011). Interactional competence in a paired speaking test: Features salient to raters. *Language Assessment Quarterly*, *8*(2), 127-145.

McNamara, T. (1996). *Measuring second language performance*. New York, NY: Longman.

McNamara, T. F. (1997). ‘Interaction’ in second language performance assessment: Whose performance? *Applied Linguistics*, *18*, 444- 446.

McNamara, T. F., & Lumley, T. (1997). The effect of interlocutor and assessment mode variables in overseas assessments of speaking skills in occupational settings. *Language Testing*, *14*(2), 140-156.

Michel, M. C., Kuiken, F., & Vedder, I. (2007). The influence of complexity in monologic versus dialogic tasks in Dutch L2. *International Review of Applied Linguistics, 45*(3), 241–259.

Nakatsuhara, F. (2006). The impact of proficiency-level on conversational styles in paired speaking tests. *Cambridge ESOL Research Notes*, *25*.

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. (2010). *Interactional competence measured in group oral tests: how do test-taker characteristics, task types and group sizes affect co-constructed discourse in groups?* Paper presented at the Language Testing Research Colloquium (LTRC), April, 2010, Cambridge, UK.

Nakatsuhara, F. (2011). Effects of test taker characteristics and the number of participants in group oral tests. *Language Testing*, *28*(4), 483–508.

Nuevo, A. (2006*). Task complexity and interaction: L2 learning opportunities and interaction* (Unpublished doctoral dissertation). Georgetown University, Washington, DC.

Nunan, D. (2004). *Task-based language teaching*. Cambridge: Cambridge University Press.

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), 161–86.

Ockey, G. J. (2014). The potential of the L2 group oral to elicit discourse with a mutual contingency pattern and afford equal speaking rights in an ESP context. *English for Specific Purposes*, *35*, 17-29.

Ockey, G. J., & Li, Z. (2015). New and not so new methods for assessing oral communication. *Language Value*, *7*(1), 1-21.

O' Loughlin, K. (1995). Lexical density in candidate output on direct and semi-direct versions of an oral proficiency test. *Language Testing*, *12*(2), 217-237.

O'Loughlin, K. (2002). The impact of gender in oral profi ciency testing. *Language Testing*, *19*(2), 169–92

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

Payne, J. S., & Whitney, P. J. (2002). Developing L2 oral proficiency through synchronous CMC: Output, working memory, and interlanguage development. *Calico Journal*, 7-32.

Pinget, A. F., Bosker, H. R., Quené, H., & de Jong, N. H. (2014). Native speakers' perceptions of fluency and accent in L2 speech. *Language Testing*, *31*(3), 349-365.

Putter, J. (1955). The treatment of ties in some nonparametric tests. *The Annals of Mathematical Statistics*, *26*(3), 368-386.

Rovai, A. P., Baker, J. D., & Ponton, M. K. (2013). *Social science research design and statistics: A practitioner's guide to research methods and IBM SPSS*. Chesapeake, VA: Watertree Press LLC.

Sato, M. (2014). Exploring the construct of interactional oral fluency: Second language acquisition and language testing approaches. *System*, *45*(1), 79–91.

Shohamy, E. (1994). The validity of direct versus semi-direct oral tests. *Language testing*, *11*(2), 99-123.

Shohamy, E. (1995). Performance assessment in language testing. *Annual review of applied linguistics*, *15*, 188-211.

Sim, J., & Wright, C. C. (2005). The kappa statistic in reliability studies: use, interpretation, and sample size requirements. *Physical therapy*, *85*(3), 257-268.

Skehan, P. (1996). A framework for the implementation of task-based instruction. *Applied Linguistics*, *17*(1), 38–62.

Skehan, P. (1998). *A cognitive approach to language learning*. Oxford: Oxford University Press.

Skehan, P. (2009). Modelling second language performance: Integrating complexity, accuracy, fluency, and lexis. *Applied Linguistics*, *30*(4), 510-532.

Storch, N. (2001). *An investigation into the nature of pair work in an ESL classroom and its effect on grammatical development* (Unpublished doctoral dissertation). The University of Melbourne, Melbourne, Australia.

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

TEPS. (2017). *TEPS score band system*. Retrieved, April 12, 2007, from http://www.teps.or.kr/Info/Teps#

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

Van Moere, A., & Kobayashi, M. (2003). *Who speaks most in this group? Does that matter?* Paper presented at the annual meeting of the Language Testing and Research Colloquium (LTRC), July, 2003, Reading, UK.

Van Moere, A. (2013). Paired and group oral assessment. In C. A. Chapelle (Ed.), *The encyclopedia of applied linguistics* (pp. 1-4). Oxford, England: Wiley-Blackwell.

Vygotsky, L. S., & Cole, M. (1978). *Mind in Society: Development of Higher Psychological Processes*. Cambridge, MA: Harvard University Press.

Wang, L. (2015). *Assessing interactional competence in second language paired speaking tasks* (Unpublished doctoral dissertation). Northern Arizona University, Flagstaff, Arizona.

Young, R. F. (2008). *Language and interaction*: An advanced resource book. New York: Routledge.

Young, R. F. (2011). Interactional competence in language learning, teaching, and testing. *Handbook of research in second language teaching and learning*, *2,* 426-443.

Zhang, Y., & Elder, C. (2011). Judgments of oral proficiency by non-native and native English speaking teacher raters: Competing or complementary constructs? *Language Testing*, *28*(1), 31-50.

Appendices

Appendix A-1

Paired Discussion Task 1

Task 1

Directions **It has been recently announced that dormitory rooms at your university must be shared by two students. Would you rather have the university assign a student to share a room with you, or would you rather choose your own roommate?** Exchange opinions with your partner, and provide specific reasons and examples supporting your position on the announcement.

Appendix A-2

Paired Discussion Task 2

Task 2

Directions During the English discussion class, you and your partner are going to talk about the topic provided below.

"Lectures in university should be given in English to help students improve their global competence."

Exchange opinions with each other, and provide specific reasons and examples supporting your position on the argument.

Appendix B

Rating Scale for General Speaking Proficiency

Rater: _____ / Task: _____ **Candidate: _____ Total Score _____**

SCORE	PRONUNCIATION	GRAMMAR	VOCABULARY	FLUENCY	ORGANIZATION
7 EXCELLENT	Speech is consistently clear. It may include minor lapses or pronunciation problems, which do not interfere with listener's comprehension.	**ACCURACY** Speech maintains consistent and effective control of both the morphological and syntactic structures, and the grammatical control is manifested while attention is otherwise engaged (e.g. listening to or responding to other interlocutors' utterance). **RANGE** The speaker uses the wide range of	Speech demonstrates a wide range of lexical choices to effectively convey and express meaning and idea. The word choices are almost always accurate and appropriate.	The speaker effortlessly and skillfully communicates with a well-paced flow. Pauses, false starts, and reformulations are used for searching right words and expressions or finding an appropriate example/explanation, but not predominant. The speaker contributes to joint discourse with natural use of back-channeling and turn-	**ORGANIZATION** The speaker clearly identifies topic and develops the topic through distinguishing and appropriately sequencing its main points. Supporting details are fully elaborated and well connected to main points. The speaker smoothly concludes expressed ideas. **COHESION** Appropriate means for coordination or reference are used to

		morphologic and syntactic structures with a good command of idiomatic expressions and colloquialisms. Speech may include few minor errors in used structures which do not obscure the speaker's message.		taking and has the capacity of scaffolding conversation with another speaker.	effectively connect utterances and mark relationships among ideas without errors.
5-6 VERY GOOD	Speech is generally clear and intelligible with minor pronunciation problems. At times listener effort is needed, but overall intelligibility is not largely affected.	**ACCURACY** Speech demonstrates a fairly automatic and effective control of both the morphological and syntactic structures with few grammatical errors. Errors are non-systematic and are generally corrected by the speaker when they occur. **RANGE** The speaker uses both basic and complex morphologic and syntactic structures,	Speech demonstrates a fairly broad range of lexical choices to convey the speaker's message and idea. It may exhibit some imprecise or awkward use of vocabulary but may not interfere with the overall comprehension of the message.	The speaker communicates almost effortlessly with a fairly well-paced flow. Pauses appear at appropriate places for searching appropriate words, examples and/or explanations, not for grammatical problems. Speech may include short hesitations and frequent repetitions. The speaker fairly properly uses back-channeling and acknowledgment.	**ORGANIZATION** The speaker clearly identifies topic and it is generally well developed through distinguishing and appropriately sequencing its main topics. The speaker provides relevant ideas/information to support the main topics, but they may lack elaboration or specificity. **COHESION** Connectors and cohesive devices are

		though the range of structures used is somewhat limited. Speech may include imprecise or inaccurate use of grammatical structures, but it does not seriously obscure the speaker's message.			appropriately and fairly accurately used for marking relationships among utterances and ideas.
3-4 MODERATE	Speech is basically intelligible, but listener effort is needed due to pronunciation problems including stress and intonation.	**ACCURACY** Speech shows a relatively reasonable control of grammar. It may include some grammatical errors that obscure the speaker's message and interfere listener's comprehension at times. **RANGE** The speaker uses basic and limited range of morphological and syntactical structures and it sometimes prevent full expression	Speech demonstrates some variety of lexical choices, but some words are inaccurately used as obscuring the meaning at times.	The speaker sometimes communicates slowly. The speaker seems to be aware of what s/he is talking about, but the speaker's used pauses indicate s/he needs time to think about grammar and/or what to say. The use of back-channeling and turn-taking is sometimes inappropriate or unnatural. The speaker passively engages in	**ORGANIZATION** The speaker identifies topic, but the number of main topics or the development of ideas is limited and may sometimes inappropriately sequenced. Supporting details for main topics are mostly basic ideas with limited elaboration or explanation. **COHESION** Connectors and cohesive devices are sometimes adequately

		of ideas.		conversation with relatively low degree of scaffolding.	used, but confusing relationships among utterances or ideas sometimes interfere with listener's comprehension.
1-2 POOR	Pronunciation problems including stress and intonation are numerous and cause huge listener effort to understand.	**ACCURACY** Speech shows severely limited control of morphological and syntactic structures with numerous and various types of errors which prevent expression of ideas and connections among ideas. **RANGE** The speaker uses only some simple morphological and syntactic structures and some low-level structures may heavily rely on memorized and formulaic expressions.	Speech demonstrates the limited range of lexical choices with frequent repetition of some words, which severely limit or prevent the speaker's meaning and ideas.	The speaker obviously struggles to communicate with a highly slow flow. Frequent and long pauses are used and impede communication. The speaker's utterances rely on heavily formulaic expressions. The use of back-channeling is not clearly identified and the speaker hesitates about taking turns.	**ORGANIZATION** Main topics are vaguely distinguished and inappropriately sequenced. Supporting details lack elaboration or explanation with no connections to ideas. The speaker is unable to complete a task and may considerably rely on repetition of the written prompt. **COHESION** Connectors and cohesive devices are inadequately used as focusing relationships among utterances and ideas.
0	Pronunciation	Speech shows almost	The size of vocabulary	The speaker cannot	The speaker's

NO EVIDENCE	problems make speech virtually unintelligible and incomprehensible, or the the speaker speaks to little to judge.	no control of morphological and syntactic structures and errors highly frequently occur.	is too limited to a few words and phrases. This limitation severely prevents expression of meaning and ideas as seriously interfering with listener's comprehension.	manage communication and it is entirely dependent on repetition of words in the given prompt. No use of appropriate back-channeling and turn-taking is found at all. The speaker is incapable of scaffolding conversation.	utterance is too short to judge or almost no relevant content is expressed.

Appendix C

Rating Scale for Interactional Competence

The CAP Scale: Open Task Outcome + Divergent Negotiation Results

Rater: _____ / Task: _____ **Candidate: _____ Total Score _____**

<table>
<tr><th>Scale</th><th>Level of quality</th><th colspan="2">Description</th></tr>
<tr><td>6</td><td>+ collaborative*
+ complete*</td><td colspan="2">In a collaborative pattern:
Both interlocutors contribute equally to the conversation and interact cooperatively.</td></tr>
<tr><td>5</td><td>+ collaborative
- complete</td><td colspan="2">They generally demonstrated an active, balanced use of features such as making comments and back-channeling.</td></tr>
<tr><td>4</td><td>+ asymmetric
+ complete</td><td colspan="2">In an asymmetric pattern:
One interlocutor assumes a more dominant role and the other a more passive role.</td></tr>
<tr><td>3</td><td>+ asymmetric
- complete</td><td>The more dominant role
➢ contributes more to the task but shows limited engagement with the partner
➢ The interlocutor frequently often holds the conversation floor,</td><td>The more passive role
➢ speaks less, mostly reacting to the dominant interlocutor
➢ The interlocutor does not actively provides his or her own opinions, sometimes shows support to the dominant</td></tr>
</table>

		interrupts the partner's utterance or takes over the floor by means of making comments and providing back-channeling signals, does not actively respond to the partner	one by making brief comments and providing back-channeling signals without intruding on the partner
2	+ parallel + complete	In a parallel pattern: The interlocutors have equal access to the conversational floor and the development of the interaction, but do not work cooperatively. They generally demonstrate a fairly limited use of features such as making comments and back-channeling.	
1	+ parallel - complete		

Notes. * Descriptors of interaction patterns are provided in the boxes above. Descriptors of *task completion* are listed as below.

1. Commonly required components in an open outcome, divergent negotiation task:

a. provide opinions or choices in respond to the prompt

b. do not need to reach an agreement after a negotiation

2. **+ complete**: the interaction accomplishes all the required components in the task prompt.
   – complete: the interaction partially accomplishes the required components in the task prompt.

Appendix D-1

Raters' Comments on Interactional Features

Related to Interactive Listening

Features	Positive	Negative
Filling a silence	*Helping out the partner to find/recall a word* *Providing an appropriate word for the partner*	–
Making relevant comments	*Relevant comments to compromise their ideas* *Reiteration for checking one's idea* *Feedback on the partner's comments* *Rephrasing what the partner says* *Complimenting on the partner's opinion*	*Absence of the comments reflecting the listener's comprehension on the speaker's response* *Low degree of understanding on the partner's question* *Irrelevant/inappropriate comments* *No feedback on the partner's comment* *Neglecting the partner's response and adding onto one's opinion* *Trying to rebut without validation* *Slow response* *Response with too many pauses*
Agreeing/ disagreeing	*Agreement on the partner's opinion* *Agreement on the partner's opinion by validating the partner's opinion* *Partial acceptance with rebuttal* *Validating the partner's opinion but disagreeing with supportive opposing ideas* *Disagreement to the partner's opinion with opposing reasons*	–
Back-channeling	*Frequent back-channels* *Decent back-channels showing*	–

	one's engagement to a discussion	
Requesting clarification	*Asking for clarification/elaboration*	–

Appendix D-2

Raters' Comments on Interactional Features

Related to Topic Management

Features	Positive	Negative
Initiation of a new topic	*Initiation of a new topic helping to continue interlocutors' discussion* *Initiation of a new topic with personal examples*	*Sudden shift to a new topic* *Initiation of a new topic that is irrelevant to the given topic*
Organizations of ideas	*Clear layout of one's view* *Logical organization of ideas* *Smooth transition in stating one's ideas* *Provision of the equal amount of input for the partner*	*Disorganized ideas/reasons/examples* *Awkward transition to the next idea* *No clear conclusion of one's thoughts*
Use of supporting resources	*Provision of specific reasons and examples* *Provision of appropriate personal examples* *Provision of relevant information* *Provision of reliable information* *Provision of possible solutions/suggestions/alternatives* *Further expansion/detailed explanation of one's opinion*	*Unable to elaborate on one's ideas* *Non-specific and unclear reasons/examples* *Unrelated ideas/solutions/alternatives* *The use of personal information rather than general information and knowledge* *Questionable source of examples with no reference* *Redundant/repetitive elaboration/expression* *Insufficient support* *Reiteration of what the partner has already mentioned*
Consistency of arguments	*No side-tracked opinions* *Indicating reasonable rebuttal points*	*Vague stance on the given topic* *Neutral stance on the given topic* *Unmatched opinion with what one has already mentioned*

		Support the opposite side *Lost on track*
Topic relevance	*Defining/clarifying the meaning of a term in the prompt* *Talking about the prompt as focusing on the given topic* *Raising awareness on the terms in the prompt and setting an agreed guideline to talk about* *Narrowing the scope of the prompt* *Backing on the given topic/prompt*	–
Consideration of both sides	*Considering both sides with explaining differing views* *Providing supportive ideas to the partner's suggestion* *Providing ideas for the opposite side*	*Strongly stick to one point of view and elaborate on it*
Negotiation of differing ideas	*Negotiation to come to an agreement* *Finding a consensus on a certain point in the prompt*	*Strongly insisting one's views without considering the partner's response* *Ignoring/shutting down the partner's idea* *Unable to reach a consensus*

Appendix E

Raters' Feedback about the Rating Scale for Interactional Competence

Category	Raters' comments
Pattern of interaction	*The degree of cooperation at different levels* *Need to add the section of "attitude" to participate in discussion* *How are candidates differentiated in terms of their score when both worked cooperatively but one person demonstrated more active participation?* *Difficult to make a decision when features were overlapped (between parallel vs. asymmetric)*
Task completion status	*The scoring rubric does not reflect how well the task was achieved.* *The interlocutors, as described, "generally demonstrate a fairly limited use of features". However, this action does not play a positive role in reaching the goal of the task.* *The scoring rubric does not indicate in depth how well the objectives of the task were met.* *Are extra points awarded when candidates come up with solution or prompt fully?* *Which category is it when the question was partially answered?* *Decent comments and back-channeling happen, but the participants did not fulfill the required task's assignment.* *Is it enough to have two options- +complete, -complete-in terms of accomplishing the task?* *Need of elaborative outline of the "completion" of task or "partial" or "incomplete"* *Both participants don't initiate any talk on global competence. Does this lead to a score of '0'?*
Prompt understanding	*How will a candidate be graded if s/he is unable to understand the prompt fully?*

	No specific mention of global competence, but shows a very little understanding of it. *What happens if only partial understanding of terminology in the prompt is evident?*
Organization of discussion	*Logical flow* *Which is better organization? Back and forth reasoning or one participant lays out all her reasons and then the next person lays out?* *The organization and the quality of the discussion should be reflected on the rubric.* *Need of adding "organization" to the scoring rubric since organization of ideas affects the actual quality of the interaction* *How to measure depth of the discussion?* *How are candidates rated on the depth of the conversation? Do they still get a high mark for just having cooperated well?*
Use of supporting resources	*Candidates did a decent job in completing the given task. However, they did not provide enough examples to ...support their opinion. Though this is a minor mistake, ...I, as a rater, had to give them "complete", when ...actually they are closer to "complete."* *Seemingly the participants have completed the task, but when thoroughly listing their reasons and supports it lacks details, explanation and credibility.*
Topic relevance	*Easily gets off topic. Good generation of questions and small talks, but mostly unrelated to the topic. How should these be rated and in which category?*
Management of discussion	*Is there separate category or extra points awarded on someone who clearly initiates the conversation, but is not dominant nor disturbing the other from engaging?* *Time management should be reflected on the rubric.* *Being unable to wrap up should be reflected on the rubric.*
Descriptors	*The scoring rubric makes very limited distinction between the two candidates. (E.g. Word choice, cohesion, natural flow, etc. is not taken into consideration.)* *What is considered as more valuable? Dominance in conversation or completion of the task?*

	Vaguely defined and no specification of what kind of behavior would fall into "cooperative" and "not cooperative" to have something in between "cooperative" and "not cooperative"
	More specific description is need for a dominant and passive role.
	How to differentiate a person between a dominant vs. not a cooperative candidate?
	A more descriptive form of "parallel" is needed.
	Unclear definition of "cooperative" and "not cooperative"
	Need of elaborative description of "cooperative" and "not cooperative"
	Extracted some criteria that differentiate participants
	Need of mid points-cooperative discussion including conversational tactics and strategies, which leads to the additional points
	Specification of back-channeling methods-verbal use vs. physical interaction
	Need of specific areas to give an examinee a plus point-made of eye contacts, used variety of features.

Appendix F

The Number of Turns and Utterances

		Task 1		Task 2	
Pair No.	**Interlocu-tor**	**No. of Turns**	**No. of Utterances**	**No. of Turns**	**No. of Utterances**
1	A	8	426	21	596
	B	6	387	17	447
2	A	15	428	22	309
	B	13	319	16	533
3	A	8	386	14	384
	B	9	401	16	513
4	A	24	529	15	498
	B	19	388	25	533
5	A	17	581	18	466
	B	16	302	24	501
6	A	5	446	4	323
	B	12	417	5	420
7	A	11	308	5	302
	B	10	440	4	461
8	A	7	413	6	490
	B	7	455	8	433
9	A	14	395	2	241
	B	16	408	3	632
10	A	30	672	18	571
	B	32	446	18	436
11	A	12	584	11	313
	B	8	544	9	311
12	A	23	358	9	347
	B	31	586	18	534

13	A	16	169	6	201
	B	24	318	10	287
14	A	8	668	3	555
	B	8	230	4	194
15	A	18	452	6	514
	B	19	244	7	278
16	A	20	294	17	424
	B	35	370	18	433
17	A	9	221	12	211
	B	13	389	10	397
18	A	4	110	3	196
	B	4	891	4	714
19	A	15	291	31	286
	B	18	592	29	601
20	A	22	585	10	572
	B	18	331	11	430
MEAN		15.1	419.4	12.2	422.2
SD		7.76	146.54	7.48	129.23

Appendix G

Research Consent Form for Participants

You are being asked to take part in a research study of how university students manage their conversation during a paired English speaking test. Please read this form carefully and ask any questions you may have before agreeing to take part in the study.

Purpose of the study:

The purpose of the study is to investigate how students manage their conversation during a paired English speaking test, which provides insight into the nature of interactional competence and the way to assess it.

Procedure and Time:

If you agree to take part in this study, you will be expected to participate in a paired English speaking test in which two discussion tasks are provided. Topics to be discussed during the test are issues related to the university setting. During the test, you will be given 2 minutes for preparation of each task and 7 minutes for each discussion. The whole process of the test will take about 25 minutes. With your permission, we will video-record the whole process of the test.

Risk and benefits:

I do not anticipate any risks to you participating in this study. Rather, you can benefit from using and practicing English during discussion and receiving scoring reports which provide information about your strong and weak points in interactional competence. I hope that your participation in

this study gives you an opportunity to practice English while managing a conversation in English.

Compensation:

You are going to earn 10,000 won for your participation in this study. Additionally, you will be provided with scoring reports representing your level of interactional competence and description of the level.

Confidentiality:

All recorded material of this study will be kept private and in a locked file. Only researchers will have access to them.

Voluntary participation:

Taking part in this study is voluntary. If you decide not to participate in the study, you are free to withdraw at any time.

Contact information:

The researcher conducting this study is Heemin Park, a graduate student at Seoul Nat'l University. Please ask any questions you have now. If you have any questions or concerns later regarding your participation in this study, you may contact Park at heemineng@snu.ac.kr.

You will be given a copy of this from to keep for your records.

Your Signature _____

Date _____

Your name _____

In addition to agreeing to participate, I also consent to having the tests video-recorded.

Your Signature _____

Date _____

Signature of person obtaining consent _____

Date _____

Name of person obtaining consent _____

Date _____

This consent form will be kept by the researcher beyond the end of the study.

국문 초록

2인 영어 토론 과제를 활용한 영어 능력 수준 상위권 한국 학생들의 상호작용 능력 평가 연구

본 논문의 목적은 2인 영어 토론 과제를 활용하여 제2 외국어로서 영어를 사용하는 화자의 영어 말하기 능력을 평가하고, 이를 통해 2인 말하기 평가 방식의 타당도(validity)와 실행 가능성(feasibility)을 검증하는 것이다. 2인 말하기 평가 방식은 실제 의사소통이 이루어지는 상황과 유사한 시험 과제(test tasks)를 제공한다는 점에서 실제성(authenticity)과 타당도가 강조되는 말하기 평가 방식으로 주목받고 있으며, 말하기 능력을 구성하는 여러 핵심 구인(constructs) 중 하나인 상호작용 능력(interactional competence) 평가를 가능하게 한다는 점에서 매우 중요한 의미를 가진다. 이러한 말하기 평가 방식의 장점에 주목하며 많은 연구자들은 2인 또는 그룹 구술시험을 통해 상호작용 능력의 본질과 시험 응시자들의 상호작용 패턴, 그리고 화자의 특성(interlocutor characteristics)이 상호작용에 미치는 영향 등을 연구해왔다. 그러나, 상호작용을 바탕으로 한 말하기 평가 방식은 2명 이상의 화자가 참여한다는 측면에서 시험의 공정성(fairness)과 신뢰도(reliability), 타당도와 관련한 많은 문제점을 끊임없이

일으켜왔고, 이에 대한 채점자들의 인식 조사에 대한 연구는 매우 부족한 상황이다. 더불어, 말하기 평가에서 중요한 역할을 차지하는 시험 과제의 특성(task characteristics)이 응시자 간 이루어지는 상호작용과 형성되는 담화의 특징에 미치는 영향에 관한 이전 연구는 매우 적을 뿐만 아니라 상충하는 결과를 도출하기도 했다. 게다가, 이에 대한 채점자들의 인식과 그들이 일정하고 타당한 판단을 내릴 수 있도록 돕는 평가표에 대한 연구 또한 매우 부족한 실정이다. 따라서 본 연구는 토론 과제를 활용한 2인 말하기 평가 상황에서 나타나는 상호작용 특질(interactional features)과 상호작용 능력 평가를 위해 Wang(2015)이 고안한 채점표의 적합성, 이에 대한 채점자들의 인식을 중점적으로 살펴보고자 하였다.

실험을 위해 2개의 토론 과제가 개발되었고, 영어 능력 수준이 상위권에 해당하는 40명의 학생이 모집되었다. 이 학생들은 2인 1조로 나뉘어 개발된 2개의 토론 과제에 같은 응시자와 함께 참여하였다. 응시자들의 상호작용 능력을 평가하기 위해 영어가 모국어인 2명의 채점자가 모집 및 훈련되었으며, 채점자들은 Wang(2015)이 고안한 6 척도의 통합적(holistic) 평가표를 사용했다. 응시자들의 상호능력 평가와 더불어 질적 연구를 통해 2인 영어 토론 과제에서 영어 수준 상위권 학습자가 사용하는 상호능력 특질과 그 가운데 채점자들이 보다 중요하다고 인식하는 특질들, 또 사용된 채점표에 대한 채점자들의 의견이 조사되었다.

실험 결과, 2인 영어 토론 과제의 사용이 적합한 수준의 채점자 간 신뢰도를 달성하였을 뿐만 아니라, 18개의 다양한 상호능력 특질을 이끌어냄으로써 타당하고 실제적(authentic)인 의사소통 평가 방식이라는 것을 뒷받침하였다. 나아가, 도출된 실험 결과의 질적 분석을 통해 시험 과제의 특성에 따라 다양한 상호작용 특질이 나타난다는 것을 확인하였다. 토론 형태인 본 연구의 시험 과제에서 채점자들은 응시자가 상호작용을 통해 얼마나 토론 주제를 논리적이고 매끄럽게 발전시키는지에 대해 중점적으로 평가하였고, 이는 이전 연구에서 다루어졌던 토론 주제와 관련된 상호작용 능력 특질(feature of topic development)이 더욱 세분화되어야 할 필요성을 보여주었다. 이와 관련하여, 채점자들은 사용한 평가표에 묘사된 응시자들의 상호작용과 과제 달성 정도가 실제 이루어진 상호작용과 담화 내용을 충분히 반영하지 못하는 한계점을 가지고 있으며, 상호작용과 과제 달성 정도는 더욱 세분화된 수준(levels)으로 이루어져야 한다고 기술하였다.

본 연구의 결과는 다음과 같은 함의를 지닌다. 첫째, 이 연구는 2인 영어 토론 과제가 시험 응시자의 상호능력 평가를 위한 증거로 사용되는 다양한 상호능력 특질을 이끌어낸다는 것을 확인하며 2인 영어 말하기 평가 방식의 타당도와 실현 가능성을 검증했다. 앞으로, 다양한 주제를 다룬 여러 유형의 시험 과제 상황에서 어떠한 상호작용 특질이 나타나는지, 화자의 영어 능력 수준에 따라 어떻게 다르게 나타나는지 등에 대해 계속해서 조사해야 할 것이다. 둘째, 이 연구는 상호능력 평가를 위한

평가표 개발에 유용한 지침을 마련한다. 교사나 평가자들은 평가표 개발에 앞서 실증적인 실험을 바탕으로 타당하고 자세하게 응시자들의 수행(speech performance)을 기술하는 평가표를 개발해야 한다. 또한, 상호작용 패턴과 과제 달성 정도는 여러 수준으로 구성된 구체적인 세부 항목으로 나누어 각 응시자가 상호작용과 담화 형성에 기여한 바를 채점자가 더욱 정확하게 평가할 수 있도록 만들어야 한다. 마지막으로, 상호작용 능력에 대한 채점자들의 개념화(conceptualization) 과정과 응시자들의 수행에 대한 채점자 인식이 지속해서 연구되어야 한다. 이는 채점자들이 평가하고자 하는 응시자의 상호작용 능력에 대해 일정하고(consistent) 타당한 판단을 내릴 수 있도록 도와주며 시험의 공정성을 보장하는 데 기여할 것이다.

주요어: 상호작용 능력, 2인 말하기 평가, 제2언어 평가, 과제 특성, 평가표, 채점자 인식, 한국 영어 학습자

학번: 2015-20050