

**PAIRED LEARNER ASSESSMENT:
CAN IT SERVE AS A VALID MEASURE OF L2 PROFICIENCY
FOR DEVELOPMENTALLY EQUAL AND UNEQUAL LEARNER PAIRINGS?**

Jiyoon Lee

A DISSERTATION

in

Education

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

2011

Supervisor of Dissertation:

Teresa Pica, Professor of Education

Graduate Group Chair:

Stanton E.F. Wortham, Professor of Education

Dissertation Committee:

Teresa Pica, Professor of Education

Yuko Goto Butler, Associate Professor of Education

Christina Frei, Adjunct Associate Professor of Education

Paired Learner Assessment:

Can It Serve as a Valid Measure of L2 Proficiency

For Developmentally Equal and Unequal Learner Pairings?

COPYRIGHT

2011

Jiyoon Lee

“Your beginning will seem humble, so prosperous will your future be.”

Job 8:7

ACKNOWLEDGEMENTS

I would like to express my deepest thanks to those who have helped me along the way, and without whose support, encouragement, advice, and contributions, this dissertation would not have been possible. I would also like to thank the Korean Studies Program at the University of Pennsylvania and the Educational Testing Service for their generous funding of this study.

I have been extremely fortunate to have had the advice and guidance of three extraordinary scholars, each of whom has contributed to the development of my scholarship and my intellectual growth. I am grateful to Dr. Christina Frei for her advice from the beginning of the dissertation. Her insights about Oral Proficiency Interview and analysis of test-taker discourse helped me to form the idea of this study. I am also indebted to Dr. Yuko Goto Butler, who generously advised me on statistical analyses. Moreover, I greatly appreciate her advice and apprenticeship as a mentor throughout my study at Penn. The long but interesting discussions that we had about the many research studies we have prepared and conducted together allowed me to grow as a critical researcher, and I will never forget her appreciation of my ideas. My deepest gratitude goes to Dr. Teresa Pica, my dissertation committee chair, my mentor, and my teacher. Her insights, critical comments, love of teaching and research, and constant faith in me inspired me throughout my study. Our discussions about my study, along with her guidance, advice, and care, gave me hope and faith in my life as an emerging scholar and teacher. Her dedication to teaching and research will be a role model for me, and will always motivate me.

I am grateful to the study participants, friends, and colleagues who have helped me in many ways throughout this long process: Elaine Allard, Leslie Altena, Danielle Bergez and her little son Leo Bergez, Sungjung Cho, Bridget Goodman, Na-rae Han, Seunghee Hong, Karen Jury, Soyeon Kim, Amanda Kniepkamp, Kathy Lee, Genevieve Leung, Hoa Nguyen, Suzanne Oh, Xiaolin Peng, Shannon Sauro, Laura Sicola, Ming-Hsuan Wu, Taeim Yoon, and Wei Zeng. They were great friends, discussants, and study partners throughout my doctoral study. Our conversations motivated and inspired me. I especially enjoyed and appreciated the time that Julia Deak and I worked together at the Graduate Student Center. My special thanks go to my former neighbor, Matthew Schreibeis, who was a very supportive and critical study partner during this particularly challenging time. His support and patience helped to make the writing process more leisurely. I also want to thank Lynn and Bob Gruner who always welcomed my weary body and soul. They were an oasis in my life in the States.

Last but not least, my earnest gratitude goes to my beloved family, whose endless love, support, and faith, carried me through this long journey. Although we were separated by the Pacific with a thirteen-hour time difference, their cheerful voices, day and night, encouraged me and helped to sustain me. I appreciated Youngim's endless love, jokes, and funny stories. I cannot thank enough my one and only baby sister, Jiyoung, for her love and support. I always depended on her maturity and consideration. Finally, I am thankful for my parents: my father, for his vision, strong will for my future, and love, including his more than 30 letters; and my mother, for her unconditional love and support throughout my life.

ABSTRACT

PAIRED LEARNER ASSESSMENT: CAN IT SERVE AS A VALID MEASURE OF L2 PROFICIENCY FOR DEVELOPMENTALLY EQUAL AND UNEQUAL LEARNER PAIRINGS?

Jiyoon Lee

Professor Teresa Pica

Increasing attention has been given to Paired Assessment (PA), in which two second language (L2) learners work as status-equal interlocutors to demonstrate their L2 proficiency. Claims have been made that the status-equal format of a PA can provide useful data on a wide range of linguistic and sociolinguistic abilities. These abilities are more typically assessed through interviews, protocols, and questionnaires administered by a test provider who serves in a high authority capacity. PA research findings have been informative with respect to the characteristics of activities that can be used to provide valid and reliable performance data. However, the findings on interlocutor characteristics have been conflicting, a situation that has been attributed to methodological inconsistencies within and across relevant studies. Of critical concern is whether the a lower, higher, or equal L2 developmental level of a test-taker vis-a-vis that of the paired partner will yield consistent performance results. This concern is of theoretical importance with respect to the role of PA in tracking developmental change. It also holds practical importance, as PA is often carried out in classrooms, where there are often differences among learners in their developmental levels. These issues and concerns provided the impetus for this dissertation research. Results of the study revealed that 1) ETTs' ability to produce linguistically accurate utterances did not vary

regardless of their NETTs developmental stages. This result was consistent to that of their test-raters' evaluation. 2) ETTs' ability to interact in ways that are sociolinguistically appropriate and interactionally strategic did not vary regardless of their NETTs' L2 developmental stages. However, the test-raters' evaluation of ETTs' performance in this dimension revealed variation depending on NETTs' L2 developmental stages. These results shed light on 1) the extent to which there is variation in ETTs' linguistic and sociolinguistic performance in relation to NETTs' L2 development; 2) the role of PA in providing data that can contribute to a valid and reliable assessment battery; 4) the value of PA as a classroom assessment as well as high-stakes testing instrument.

TABLE OF CONTENTS

List of Tables	viii
List of Figures	xi
CHAPTER ONE: Rationale.....	1
Theoretical Framework of Interaction in Assessment	3
Practical Application of Interaction in Assessment	6
One-on-one Interview: Testers as test-interlocutors	9
Paired Assessment: Another test-taker as test-interlocutor	16
Theoretical and practical grounding of PA.....	17
Needs for Further Research	27
CHAPTER TWO: Literature Review	29
Language Ability	29
Language Ability Variation	31
Research Findings.....	34
Construct of PA	37
Methodological Shortcomings	40
Research Questions	46
CHAPTER THREE: Methodology.....	48
Participants and Setting	48
Variables	49
Operation of an Independent Variable	50

Operation of Dependent Variables.....	55
Materials	60
Advertisement/e-mail for Soliciting Participants.....	61
Screening Test	61
Background Questionnaire	64
Testing Tasks	65
Exit Questionnaire	67
Rubric	67
Procedure	69
Step 1	70
Step 2	71
Rating.....	73
Data Analysis	74
CHAPTER FOUR: Results.....	80
Overview of Research Questions	80
Results for Research Question One	82
Results for Research Question Two	85
Results for Research Question Three	87
Results for Research Question Four	88
Results for Exit Survey	93

CHAPTER FIVE: Discussion and Conclusion.....	97
Findings regarding test-takers' linguistic accuracy	100
Discussion of Results for Research Question 1	101
Discussion of Results for Research Question 3	103
Findings regarding test-takers' interaction ability.....	104
Discussion of Results for Research Question 2	105
Discussion of Results for Research Question 4	107
Implications for Pedagogy and Future Research	109
APPENDICES	
Appendix A: Participant Information	114
Appendix B: Recruitment advertisement & e-mail	115
Appendix C: Screening test	116
Appendix D: Tasks	122
Appendix E: Directions for the tests.....	129
REFERENCES	133

LIST OF TABLES

Table 1.1 Language assessments with test-interlocutors	7
Table 1.2 Subcategory of assessment with test-interlocutors	8
Table 1.3 Interaction features in language assessment	12
Table 1.4 Comparison of interaction in PA and Interviews.....	20
Table 2.1 Research questions and findings.....	34
Table 2.2 Evaluation criteria.....	39
Table 2.3 Uncontrolled variables	41
Table 2.4 Base and performance criteria	42
Table 2.5 Data analysis methods of the studies	45
Table 3.1 Description of stages of English question acquisition	51
Table 3.2 Independent Variable	52
Table 3.3 Tests of normality	53
Table 3.4 Analysis results	55
Table 3.5 Multiple Comparisons	55
Table 3.6 Operationalization of dependent variables	60
Table 3.7 Communication task types for L2 research and pedagogy analysis based on: Interactant (X/Y) relationships and requirements in communicating information (INF) to achieve task goals	65
Table 3.8 Testing tasks	67
Table 3.9 Rater rubric	68
Table 3.10 Data Collection Procedure	69
Table 3.11 an example of CRTTs' interaction in PA	70

Table 3.12 Example Sequence by CGTTs	71
Table 3.13 Example Sequence by Tests	71
Table 3.14 Test-rater rubric	73
Table 3.15 Inter-rater reliability.....	75
Table 3.16 Inter-rater reliability.....	76
Table 3.17 Inter-rater reliability	76
Table 3.18 Inter-rater reliability by group 1	76
Table 3.19 Inter-rater reliability by group 2	77
Table 3.20 Inter-rater reliability by group 3	78
Table 3.21 Summary of Statistical Tests by Research Questions.....	79
Table 4.1 Research Questions.....	80
Table 4.2 Total number of utterances	82
Table 4.3 Test for Normality – Total number of utterances	83
Table 4.4 One-way ANOVA of the total number of utterances	83
Table 4.5 Descriptive Statistics of the percentage of error-free utterances	84
Table 4.6 Friedman test results	84
Table 4.7 Descriptive statistics	85
Table 4.8 Friedman test.....	86
Table 4.9 Friedman test.....	87
Table 4.10 Descriptive Statistics.....	87
Table 4.11 Friedman test.....	88
Table 4.12 Descriptive Statistics of Sociolinguistic appropriateness	88
Table 4.13 Tests of Normality	89

Table 4.14 Multivariate Tests ^b	89
Table 4.15 Friedman test.....	90
Table 4.16 Descriptive Statistics of Interaction Strategies	90
Table 4.17 Tests of Normality ^b	90
Table 4.18 Friedman test.....	91
Table 4.19 Wilcoxon test	91
Table 4.20 Wilcoxon test	92
Table 4.21 Wilcoxon test	92
Table 4.22 Descriptive Statistics	93
Table 4.23 Descriptive Statistics	93
Table 4.24 Descriptive Statistics	94
Table 4.25 Tests of Normality	95
Table 4.26 ANOVA	95
Table 4.27 Ranks	95
Table 4.28 Test Statistics	95
Table 4.29 Summary of research questions and results	96

LIST OF FIGURES

Figure 1.1 Language Assessment Framework	4
Figure 1.2 Language Assessment Framework without Interlocutors	5
Figure 1.3 Distribution of language uses in the one-on-one interviews	13
Figure 1.4 Distribution of language uses in PA.....	18
Figure 3.1 Processes of the normality test	53
Figure 3.2 Processes of analysis	54
Figure 3.3 Scrambled questions task	62
Figure 3.4 Preference task.....	63
Figure 3.5 Production task	63
Figure 3.6 Background Questionnaire	64
Figure 3.7 Instruction.....	66
Figure 3.8 Moves of CGTTs during PA	72
Figure 3.6 Background Questionnaire	64
Figure 3.7 Instruction.....	66
Figure 5.1 Language assessment framework	98

CHAPTER ONE: Rationale

Introduction

This study was designed to better understand paired assessment¹ (PA) as an approach to evaluating L2 learners at process and outcome levels by analyzing the interaction between two non-native speaking test-takers (Csepes, 2002; Hughes, 2003; Nakatsuhara, 2006; Swain, 2001). This study was motivated by evidence of variation in the quality and quantity of L2 output and interaction in learner pairings that differ in L1, gender, ethnicity, or L2 developmental stages (e.g., see Gass & Varonis, 1989 for L1; Pica, Holliday, Lewis, & Morgenthaler, 1989 for gender; Beebe, 1977; Beebe & Zuengler, 1983 for ethnicity, and Watanabe & Swain, 2007 for language ability). As proposed, the study addressed questions as to whether the L2 samples obtained through PA are valid indicators of linguistic accuracy, sociolinguistic appropriateness, and interaction strategies for SLA across pairs of same and different language developmental stages. It also addressed questions about the extent to which PA offers unique information on L2 learning processes that have heretofore been difficult to assess.

My interests and concerns about this particular assessment format are deeply rooted in my experience as a language learner and language teacher. As a long time language learner from a predominantly exam-focused educational context, taking tests was one of the major concerns that I always had at school. My experience with test

¹ The terms *assessment*, *testing*, *evaluation*, and *measurement* are often used interchangeably despite their differences. Those subtle differences are as follows; *assessment* encompasses any procedure to collect information of individual or group of test-takers both qualitatively and quantitatively, the term *testing* usually implies a procedure to collect a specific type of information, *evaluation* usually involves decision making, and finally *measurement* entails quantification procedure of data collected (Allen & Yen, 2001; Davis, et al., 2002; Kizlik, 2008). Despite these technical differences, this study will use these terms interchangeably.

taking intensified when I enrolled at an innovative foreign language high school. The curricular innovations included new assessment approaches such as interviews with a teacher, interaction with other classmates, and individual presentations.

The unique characteristics of interaction with my classmates, in particular, caught my attention: unlike other assessment approaches, this approach required cooperating with other classmates and sharing responsibilities. I felt relieved to have another classmate, with whom I could work, yet at the same time it was a challenging experience as I realized that my performance was easily affected by my classmate. For instance, if the conversation between my classmate and me was based on cooperative interaction, I felt confident and comfortable during the assessment. On the other hand, if the interaction was argumentative and confrontational or proceeded to a direction that I did not expect, I lost confidence and often made mistakes.

My experience of PA as a language learner influenced my practice as a language teacher. The dynamics between two test-takers as well as their joint endeavor to negotiate meaning and manage interaction during communication breakdown strongly attracted my attention. The idea that I could save time by simultaneously assessing two students also urged me to keep using this assessment approach as an option. Nonetheless, my students expressed similar concerns that I had experienced as a language learner. They were worried about the other test-takers' lack of preparation, argumentative and confrontational manner, their dominance in interaction, and the possible influence of the other test-takers on their performance. In addition, it was challenging for me to assess individual students' performance on the tests since two test-takers jointly contributed to

test results. It was also complicated for me to decide how much attention I should pay to test-takers' language itself over their interaction.

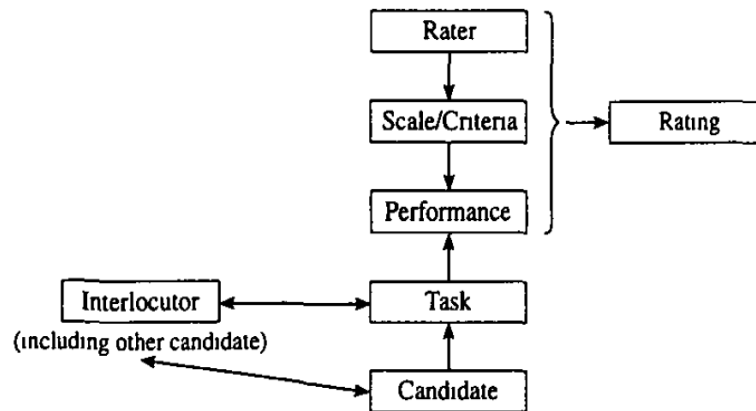
Through my academic training, I found that language testing researchers have shared similar concerns and investigated issues related to these experiences. Their research findings and discussion were helpful to me in systematizing my personal experience and situating it within the larger academic context.

This chapter will start with a general introduction of an assessment framework that informs understanding of the mechanism of aforementioned assessment approaches and then focus on two particular assessment types that follow this framework. It will provide theoretical and practical grounding to introduce PA as part of a testing battery. Finally the chapter ends with introducing unresolved issues of PA and calling for more systematic research.

Theoretical Framework of Interaction in Assessment

Figure 1.1 displays the theoretical framework suggested by McNamara (1996 & 1997). As such, it elaborates a multi-faceted procedure that includes test participant roles and contributions.

Figure 1.1 Language Assessment Framework (Candidate: Test-taker)



(Adopted from McNamara 1996: 86)

This framework illustrates each facet involved in testing such as test-takers, test-interlocutors, testing-tasks, test-raters, and scales/criteria: as such it helps understanding of their roles and embedded challenges and issues. According to this framework, test-takers display their ability to use target language while solving testing-tasks and interacting with test-interlocutors. Their performance is evaluated by test-rater(s) using pre-developed rating scales or criteria.

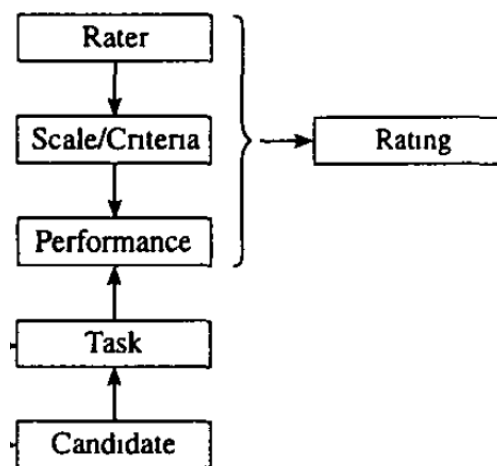
What is important in this framework is the presence of test-interlocutors in a testing setting: test-interlocutors can be either a) tester(s) who interact with and evaluate test-takers' performance or b) other test-takers who interact with test-takers to solve testing tasks and whose performance is evaluated by tester(s) who do not appear in testing settings but observe both test-takers and test-interlocutors' performance. Testers as test-interlocutors are to elicit ratable language samples from test-takers by asking a series of questions. In many cases, while asking questions and leading interaction, testers evaluate test-takers' task performance. In contrast, other test-takers as test-interlocutors

compete or cooperate with test-takers in order to solve testing-tasks, and their performance will be evaluated by test-raters who observe their performance without actively participating in interaction during assessment.

Regardless of test-interlocutors' status and roles in testing situations, this framework emphasizes the presence of test-interlocutors in the testing settings and the interaction that they can generate. This framework makes it possible to observe: 1) test-takers' ability to use language and 2) their socio-cognitive processes of solving testing-tasks (Lantolf, 2000; McNamara, 1996 & 1977; Swain, 2001). The interaction between test-takers and test-interlocutors works as a device to elicit test-takers' samples of language that can be used to gain insight into a learner's linguistic ability.

Moreover, the importance of test-interlocutors in testing becomes even clearer when contrasted with the linear assessment process depicted in figure 1.2 that excludes test-interlocutors from the testing setting.

Figure 1.2 Language Assessment Framework without Interlocutors (Candidate: test-taker)



(Adopted and modified from McNamara 1996: 86)

Figure 1.2 only helps test-raters to observe the end product that test-takers can solve testing-tasks; furthermore, unless a think-aloud protocol is employed, the test-takers' cognitive process to solve testing-tasks is hardly observable, and their interaction strategies and sociolinguistic moves are difficult to identify.

McNamara's framework depicted in Figure 1.1 emphasizes that test-interlocutors' role is vital to externalize test-takers' language and interaction for assessment purposes. It also argues that language produced under this assessment framework can have a greater range of linguistic use in comparison to the one produced in the linear assessment process in Figure 1.2. The following section introduces two assessment types that employ test-interlocutors in the testing settings: the characteristics of the two different assessment types will be explained and research studies about these assessment types will be discussed subsequently.

Practical Application of Interaction in Assessment

Test-interlocutors in McNamara's framework can be either 1) testers or 2) other test-takers. These two distinctive statuses of test-interlocutors serve to determine different assessment types: a tester as a test-interlocutor as in a one-on-one interview, and a test-interlocutor as another test-taker as in a PA.

Testing organizations such as the American Council on the Teaching of Foreign Languages (ACTFL), the Educational Testing Service (ETS), the University of Cambridge ESOL Examination Center (Cambridge ESOL), the Center for Applied Linguistics (CAL), along with many others have developed a variety of assessments

employing either or both test-interlocutor types. Table 1.1 shows some of the examples of those assessments.

Table 1.1 Language assessments with test-interlocutors

Testing Organization	Testing name	Target population	Testing format
ACTFL	Oral Proficiency Interview (OPI)	Adults	Interview
British Council/IDP Education Australia	International English Language Testing System (IELTS)	Adults	Interview
	Key English Test (KET)	Adults	Interview
University of Cambridge ESOL Examination	Preliminary English Test (PET)	Adults	Interview
	First Certificate in English (FCE)	Adults	Interview/PA
	Certificate in Advanced English (CAE)	Adults	Interview/PA
	Certificate of Proficiency in English (CPE)	Adults	Interview/PA
	Certificates in ESOL Skills for Life	Adults	Interview
	Cambridge Young Learners English Tests (YLE)	Children	Interview
	Student Oral Proficiency Assessment (SOPA)	Grade 2 – 8	Interview
Center for Applied Linguistics (CAL)	CAL Oral Proficiency Exam (COPE)	Grade 5 – 8 immersion program	Interview
	Early Language Listening and Oral Proficiency Assessment (ELLOPA)	Young children (PreK – 2)	Interview
	Early Language Listening and Oral Proficiency Assessment (ELLOPA)	Young children (PreK – 2)	Interview
	Hong Kong Examinations and Assessment Authority (HKEAA)	Hong Kong Use of English A/S level Examination	Young adults

These assessments are roughly categorized into two different types depending on test-interlocutor roles. The first category is a one-on-one interview where a high-authority figure such as a teacher or an examiner examines a test-taker, and the second category is a PA where two or more test-takers work on testing materials together. The following table shows this sub-categorization in detail.

Table 1.2 Subcategory of assessment with test-interlocutors

Interlocutor Status	Testing Organization	Exams
Tester/higher authority test-interlocutor (Test-outsider)	American Council on Teaching Foreign Languages (ACTFL)	Oral Proficiency Interview (OPI)
	British Council/IDP Education Australia	International English Language Testing System (IELTS) Speaking Section
Another test-taker (Test-insider)	University of Cambridge ESOL Examination	First Certificate in English (FCE)
		Certificate in Advanced English (CAE)
	Hong Kong Examinations and Assessment Authority (HKEAA)	Certificate of Proficiency in English (CPE) Hong Kong Use of English A/S level Examination

The following sections will describe one-on-one interviews' history, characteristics, and criticisms, which will serve as rationale for implementing PA as part of a testing battery.

One-on-one interview: Testers as test-interlocutors

Format and History

A one-on-one interview is one type of assessment which employs a test-interlocutor. Testers as test-interlocutors are native or near-native speakers of a target language and often times they manage interaction during assessment by initiating, following up, or terminating utterances. This format allows testers to interact with one test-taker at a time, which helps them to customize language and interaction accordingly by modifying questions or interaction patterns. Testers ask questions in order to elicit ratable language samples from test-takers for approximately fifteen to 30 minutes depending on test-takers' language ability. Language samples are collected through a series of questions whose purpose may differ depending on the stages of interviews (i.e., warming-up, level-checking, level-probing, and winding down). For instance, testers use

the same content with different sentence structures or different content with the same sentence structures to confirm or challenge test-takers' level (ACTFL, 1999).

This assessment approach has been most widely employed in the field since 1940s, which was practiced due to a diplomatic and military necessity (Alderson & Banerjee, 2001; Barnwell, 1996; Carlsen, 2002; Fulcher, 1997; Luoma, 2004). As the United States participated in World War II, it was critically necessary for diplomats, officers, and civil servants to have good command of foreign languages to conduct their assignments abroad successfully. With these practical necessities in mind, in the 1950s the Foreign Service Institute (FSI) focused on improving their oral language ability and developing a one-on-one interview format as a component of its testing suite (Alderson & Banerjee, 2001; Bachman & Palmer, 1981; Fulcher, 1997). Later the ACTFL adopted this format as an academic version, which is known as the Oral Proficiency Interview (OPI).

Criticisms

Despite its popularity in the field, this format was not free from criticisms. Although one-on-one interviews aim to assess whether test-takers' abilities to produce and comprehend language and interact in ways that are sociolinguistically appropriate such abilities have not been measured due to limited social interaction possibilities of the format (e.g., ACTFL, 1999; Brown, 2003; Salabery, 2000). The language sample gathered in interviews is not a valid predictor of test takers' comprehensive ability and their ability to use language beyond interview setting. This claim has been used to prove that one-on-one interviews have construct as well as criterion validity problems. That is, as the format cannot assess what it is supposed to and cannot produce information about

test-takers' ability beyond a testing setting. Moreover, the asymmetrical relationship and one-directional interaction between test-takers and testers (test-interlocutors) has brought much criticism. This criticism will be discussed in the following sections.

1) Limited information on test-takers' language use

Validation studies of a one-on-one interview found that the interview format does not necessarily yield a wide range of language uses usually elicited in conversation between status equals (Bachman & Savignon, 1986; Lazaraton, 1996; Kitajima, 2009; Swender, 1999; van Lier, 1989). This finding did not comport with the claim by one-on-one interview test-developers that test-takers' ability can be comprehensively elicited and observed in the testing setting (i.e., a construct validity problem). In addition, the results of the interview play a limited role in identifying the test-takers' ability to use language in other contexts and pose a criterion validity problem.

As analyses of the discourse of interview data have revealed, usually only simple declarative sentences, isolated lexical items, or function words are produced. For instance, as noted in excerpt 1, only minimal level of test-takers' language production and interaction have been found in the one-on-one interview setting (Brooks, 2009).

Excerpt 1.1: T – Tester, A – Ami (test-taker)

- 1 T: So do you, do you mostly use uh the Internet, do you use the email, to keep in touch with
- 2 A: **Mm hm,**
- 3 T: With your friends
- 4 A: **Yeah**
- 5 T: Back
- 6 A: **Mm hm**
- 7 T: Back home in Japan?
- 8 A: **Yeah.**
- 9 T: And do you, do you use these abbreviations and these short forms of words when you talk to your friends in Japan?
- 10 A: **No never.**
- 11 T: Never?
- 12 A: **I've never done.**

- 13 T: Really?
14 A: **No.**

(Adopted from Brooks, 2009: 354, emphasis added)

As indicated in bold print, the language produced by the test-taker in excerpt 1.1 is limited to a simple declarative sentence (e.g., *sentence 13*) or yes/no (e.g., *sentences 4, 8, 11, & 15*). These utterances do not necessarily provide a full picture of this test-taker's ability to use language in the context; in particular, considering the fact that this test-taker was attending one of the highest levels in their language institution, the elicited language does not provide enough details regarding their language ability.

Another example taken from Brown (2003) also shows this pattern:

Excerpt 1.2. I - Interviewer (tester) & C – Candidate (test-taker)

- 1 I: so, you're from the Chinese community yourself is that [right?]
2 **C: [yes.]**
3 I: so do- Chinese people eat a lot of Indian food< or is it mainly (.)Chinese food.
4 **C: oh mainly Chinese food. (0.6)**
5 I: but sometimes you eat Indian
6 **C: e::r yeah sometimes (0.9)**
7 I: sometimes Malay.
8 **C: mmm:: (0.9)**
9 **C: yeah [hnhnhn] .hh hh. (.) not very often.**
10 I: [not often though]
11 I: (°I see.°) (1.0) erm now tell me your plans are w-when . . .

(Adopted from Brown, 2003:12, emphasis added.)

This interaction suggests that this test-taker comprehended the tester's questions; however, she did not elaborate her answers but she could manage to convey her message. More detailed information, whether this test-taker has ability to produce a range of language use, is hardly inferred from this utterance.

In addition, the information about test-takers' ability to initiate questions is rarely elicited in this assessment setting (Brooks, 2009; Johnson, 2001; Lazaraton, 2002; Young, 2002; Young & He, 1998). The instances of questions initiated by the test-taker carries important information about the stages of test-takers' Interlanguage (IL) development

(Pienemann & Johnston, 1984; Mackey, 1999) as well as their ability to take control of interaction by initiating and changing topics and gaining, holding, or yielding turns.

Table 1.3 taken from Brooks (2009) shows the percentage information regarding test-takers and testers' features of interaction.

Table 1.3 Interaction features in language assessment

Features of interaction	Individual format					
	Student			Tester		
	Freq.	Range	%	Freq.	Range	%
Seeking confirmation	38	6	31.7	39	7	11.9
Asking a question	6	2	5.0	162	21	49.4
Asking for agreement	20	5	16.7	10	2	3.0
Clarification request	23	6	19.2	14	3	4.3
Incorporating words	17	6	14.2	6	3	4.3
Prompting elaboration	2	1	1.7	36	7	11.0
Finishing sentences	4	2	3.3	12	4	3.7
Referring to partner's ideas	-	-	-	10	3	3.0
Paraphrasing	-	-	-	19	3	5.8
Eliciting opinions	4	1	3.3	7	1	2.1
Expressing incomprehension	2	1	1.7	1	1	.3
Managing topic	-	-	-	3	1	.9
Suggesting words	-	-	-	4	1	1.2
Other correcting	-	-	-	4	2	1.2
Responding to help	-	-	-	1	1	.3
Asking for help	-	-	-	-	-	-
Correction uptake	4	2	3.3	-	-	-
Total	120	-	100	328	-	100

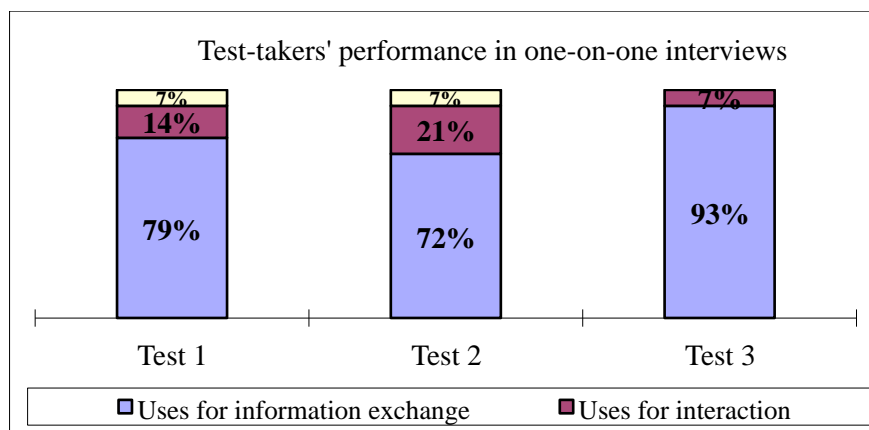
(Adopted from Brooks, 2009:352)

While comparing the percentage information of test-takers' and testers' features of interaction, Brooks (2009) found noticeable differences in the percentage of asking questions. Almost 50% of testers' utterances was dedicated to asking questions to test-takers who rarely ask questions to testers. More significantly, it is hardly observable that test-takers (students in the table) managing topics or providing corrective feedback to their test-interlocutors.

This finding confirms another earlier study that looked at the distribution of

language use in the one-on-one interviews. French (1999) analyzed three one-to-one interviews in order to see the relative proportion of the following three language uses, informational, interactional, and managing interaction. He found that the uses for exchanging information dominantly appeared in test-takers' language (79%), and there were some instances of uses for interaction, but very little information about uses of managing interaction.

Figure 1.3 Distribution of language uses in the one-on-one interviews



(Adopted from French, 1999)

This finding confirms Salaberry (2000)'s analysis of test-takers' utterance in an interview. While checking the six major categories of language uses featured in the conversation, he found that only one (i.e., imparting and seeking factual information) was fully elicited. Rather they did not take a risk to deal with advanced use of language such as inviting and eliciting, moving to different directions, or terminating the interaction with test-interlocutors; they only tried the basic language function which is transmitting factual information to their test-interlocutors.

This limited information about test-takers' language and interaction ability has led researchers to question the claim that a one-on-one interview format can measure test-takers' ability to use language comprehensively. Its construct validity was also challenged as the testing format does not measure what it is supposed to. Moreover, although a one-on-one interview attempts to adopt the features and structure of conversation, the fundamental structure and the language produced during the assessment do not go beyond the characteristics of interview formats (Lazaraton, 1992, 1996, & 1997, van Lier 1989, Young, 1995; Young & Milanovic, 1992).

2) *Asymmetrical relationship between test-takers and test-interlocutors*

Another major criticism of this testing format is the one-sided flow of information. That is, tester-initiated questions and the lack of test-takers' decision making possibilities characterize interaction between testers and test-takers "pseudosocial" as well as unilateral (Van Moere, 2006; Van Lier, 1989:501). The asymmetrical relationship limits the opportunity for test-takers to control the flow of interaction during the interview by limiting the possibilities of their decision making regarding topic initiation, persistence, shift, or termination. Lantolf and Ahmed (1989) and Perret (1990) provided evidence to support the claim that a rigid asymmetrical relationship was preserved during interviews. They also argued that it limited the opportunity that test-takers could actively participate in interaction and hindered observation of their ability to cope with a range of discourse situations in which they might play several different conversational roles or take different stances. This particularly rigid nature of the relationship also makes it challenging to

observe test-takers' ability to interact in ways that are sociolinguistically appropriate with a wide range of test-interlocutors.

Furthermore, researchers have shown concern that this asymmetrical setting may force test-takers to conform to the tester's socio-cultural standards (Salaberry, 2000; Van Moere, 2006). For instance, the OPI protocol asks testers to challenge test-takers at a probing stage in order to examine the ceiling of test-takers' language ability. However, it is often reported that test-takers misunderstand this prompt as a signal that testers want them to adjust their stance or thoughts to those of testers. These findings make us question that the rigid interaction environments and asymmetrical relationships may hinder test-takers from experimenting and trying discourse patterns required in a range of sociolinguistic situations, which eventually limit the opportunity to observe whether test-takers have communicative competence. Furthermore, as the testers highly rely on questions to elicit interaction, there is the risk that a one-direction pattern of interaction will ensue. This argument is further supported by other research findings that an interview format does not always allow test-takers to engage in the main features of a conversation such as turn initiation, termination, and the control of topics and talking time (e.g., van Lier, 1989).

A correlation study that compared a one-on-one interview and group assessment formats confirmed that there was little overlap in test-takers' language use in those two different testing formats (e.g., Shohamy et al. 1986). Moreover, the partial scope of test-takers' language use found in the aforementioned studies supported the need for another assessment type that would allow researchers to gather a richer sample of test-takers'

language. These findings collectively provide strong evidence that another type of assessment should be introduced, and many researchers have directed their attention to other forms of assessment formats. The sections that follow briefly explain basic information regarding PA, which is another form of assessment employing test-interlocutors.

Paired Assessment: Another test-taker as test-interlocutor

Format and History

The aforementioned dissatisfaction with the one-on-one interviews has drawn the attention of language teachers and test developers to another form of assessment coined paired assessment (PA), where another test-taker appears as a test-interlocutor in a testing setting. In contrast to the one-on-one interview, at least two test-takers who are both non-native speakers of a target language work on a series of testing tasks in this assessment (Hughes, 2003; Nakatsuhara, 2006; Swain, 2001). In this assessment type, instead of a tester who is a native or near native speaker of a target language and who has higher authority to initiate, persist, and terminate topics during assessment, PA involves another non-native test-taker as a test-interlocutor. Working on a series of testing-tasks together, these two test-takers can suggest, persist, and terminate topics at their discretion. In contrast to one-on-one interviews, PA presents symmetrical relationship between test-takers.

This format is not new in language classroom. This format has been widely practiced as a setting for classroom activities and sometimes classroom assessment. Two

standardized tests, several suites of Cambridge ESOL assessment and Hong Kong Examinations and Assessment Authority (HKEAA), have also adopted this format. In the following sections, I will review the theoretical and practical grounding of PA as a component of a testing battery. PA allows researchers to 1) observe a variety of language uses, 2) examine test-takers' ability to interact in ways that are sociolinguistically appropriate, and 3) observe test-takers' ability to interact in ways claimed to be strategically useful for SLA. PA also helps to 4) connect teaching and testing. I will then conclude a section on the necessity of research to examine the variables that may affect test-takers' performance during PA.

Theoretical and practical grounding of PA

1) PA: an approach to assessment that reveals a variety of test-takers' language use

PA creates a setting where test-takers produce richer and more varied language use (Iwashita, 1999; Lazaraton, 2002; Taylor, 2000 & 2001). The quantity and richness of learners' output provided more samples that could be used to evaluate learners' development and performance with respect to form, function, and appropriateness. Test-takers' utterances are characterized by the frequency of question forms (excerpt 1.3).

- Excerpt 1.3 E: test-taker #1 M: test-taker#2
- | | | |
|----|----------|---|
| 1 | E | ok. Which one you (want) prefer? |
| 2 | M | A:::H in my opinion we – I wanna choose a hotel |
| 3 | E | I will choose a hotel too ((laughter)) |
| 4 | M | YEA:H? |
| 5 | E | yeah |
| 6 | M | why? |
| 7 | E | well (.3) basically: (.8) I like to (inform) I like to negotiate with people |
| 8 | M | Ah[a] |
| 9 | E | [%you] know% DEALing with people |
| 10 | M | yeah [me to-] |
| 11 | E | [talk] to people (you [know]) |
| 12 | M | [yea] because you know I like to go traveling everywhere so:: (.8) if I can working in hotel I can (.) see a lot of people's different – ah [their from different] fro- (.6) different from (countries) |

13 E [Yeah different people] yeah especially when (inside) =
 14 M = yes I: [we can]
 15 E [((laughter)) ya:h]
 16 M we can practice all English and ah maybe French [or Spa]nish with them =
 17 E [yeah] = French [Spa]nish
 18 M [yeah] [() benefits]
 19 E **[yeah you can get] () yeah (.6) SHOP?**
 20 M N[O:I] don't think so
 21 E [yeah] it's too boring =
 22 M = yes too boring I don't – (.3) I don't like to be suffer from (.3) boredom
 summer
 23 E yeah I do:n't (.6) know (.5) back and forth pick all those milk shelves [yeah]
 24 M **[how] about a (.3) farm?**

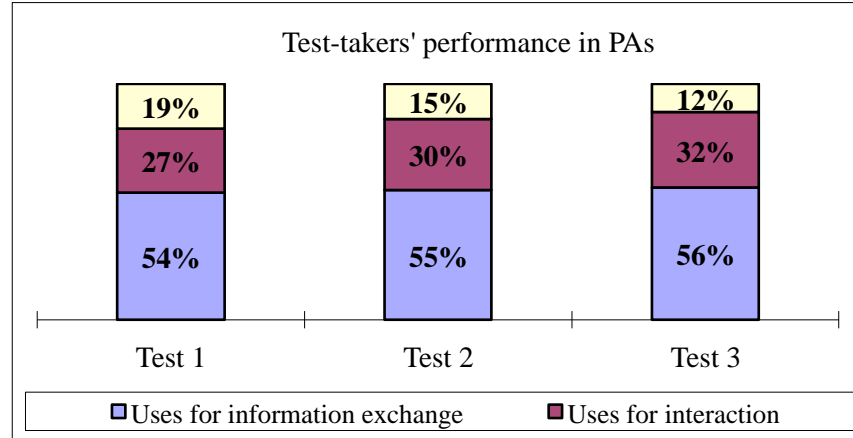
(Adopted from Galaczi, 2008:101, emphasis added.)

Those questions were used in order to initiate the interaction or invite the test-interlocutor to participate in the interaction (lines 1 and 24). This is clearly different from a passive respondent role that test-takers have been shown to play in a one-on-one interview.

Some of the questions are employed to clarify the test-interlocutor's utterance (line 4), and finally, they use questions to extend the interaction requesting elaboration of former utterance (line 6) or suggesting another choice during the task (line 19). Furthermore, syntactic complication with subordinating clauses in test-takers' language (line 12) and more advanced expressions such as phrasal verbs (Chen, 2007; Liao & Fukuya, 2004) (line 22) convey more information about test-takers' ability to use embedded constructions and complex sentences.

French (1999) offers another example of a wide range of language use. In contrast to his findings regarding the proportions of three language uses such as information, interactional, and managing interaction (see Figure 1.3), those from test-takers language use in PA was more diversified (Figure 1.4).

Figure 1.4 Distribution of language uses in PA



(Adopted from French, 1999)

Test-takers still used language for exchanging information around half of their utterances in average; however, the proportion of test-takers' language use for interaction and managing topics noticeably increased in PA. This finding shows that test-takers tried more complicated language use such as inviting, initiating, or terminating interaction.

2) *PA: As an approach to assessment that reveals test-takers' communicative competence*

The varied language use and a range of communication situations that can arise during PA provide insight into the learners' knowledge and use of sociolinguistically appropriate language with respect to functions such as conversational repair, agreement, disagreement, and strategies for seeking clarification and checking comprehension. PA thereby provides language samples that reveal the extent to which test-takers' have acquired important components of communicative competence.

Communicative competence is composed of grammatical (words and rules), sociolinguistic (social appropriateness), discourse (cohesion and coherence), and strategic competence (appropriate use of communication strategies), and thus a PA has the

potential to reveal the extent to which test-takers have not only grammatical knowledge but also the ability to use this knowledge appropriately and strategically depending on situations (Bachman, 1990; Canale, 1983; Canale & Swain, 1980; Hymes, 1972; Savignon, 1997).

The features of interaction in Table 1.4 show that test-takers were situated in interaction situations where their communicative competence could be challenged and assessed. Some of them are more complicated and require test-takers' advanced ability to handle.

Table 1.4 Comparison of interaction in PA and Interviews

Features of interaction	Paired format			Individual format					
	Student–student			Student			Tester		
	Freq.	Range	%	Freq.	Range	%	Freq.	Range	%
Seeking confirmation	50	7	20.7	38	6	31.7	39	7	11.9
Asking a question	33	6	13.7	6	2	5.0	162	21	49.4
Asking for agreement	26	7	10.8	20	5	16.7	10	2	3.0
Clarification request	22	5	9.1	23	6	19.2	14	3	4.3
Incorporating words	20	5	8.3	17	6	14.2	6	3	4.3
Prompting elaboration	18	5	7.5	2	1	1.7	36	7	11.0
Finishing sentences	15	5	6.2	4	2	3.3	12	4	3.7
Referring to partner's ideas	13	3	5.4	–	–	–	10	3	3.0
Paraphrasing	12	2	5.0	–	–	–	19	3	5.8
Eliciting opinions	10	1	4.1	4	1	3.3	7	1	2.1
Expressing incomprehension	8	2	3.3	2	1	1.7	1	1	.3
Managing topic	5	2	2.1	–	–	–	3	1	.9
Suggesting words	3	1	1.2	–	–	–	4	1	1.2
Other correcting	3	1	1.2	–	–	–	4	2	1.2
Responding to help	1	1	.4	–	–	–	1	1	.3
Asking for help	1	1	.4	–	–	–	–	–	–
Correction uptake	1	1	.4	4	2	3.3	–	–	–
Total	241	–	100	120	–	100	328	–	100

(Adopted from Brooks, 2009: 352)

These examples include features of interaction that distinctively appeared in PA such as managing topics (2.1 % vs. 0%) and referring to partners' ideas (5.4% vs. 0%).

Understanding an appropriate moment and an adequate way to join, interrupt, persist, and

terminate topics require knowledge of language and situations as well as ability to use the knowledge appropriately. Instances of referring to test-interlocutors' ideas also show whether test-takers can produce utterance cohesively and coherently. These features of interaction also reveal test-takers' attitudes toward their test-interlocutors and ability as well as willingness to sustain interaction. This information is important to assessing communicative competence because test-takers' attitudes and willingness play a role in actively and appropriately using the knowledge.

Symmetrical relationships between two test-takers also challenge test-takers' communicative competence as they allow interaction situations of cooperation as well as confrontation, disagreement, or competition (Együd & Glover, 2001; Galaczi, 2008; Iwashita, 1999; Kormos, 1999; Lazaraton, 1997 & 2002). Excerpt 1.4 presents two test-takers' interaction of disagreement and their strategies to persist their arguments.

Excerpt 1.4 TT 2: Test-taker #2, TT 3: Test-taker #3

- 1 TT 2: *but* I thought suburban, I mean, calm park was kinda boring for others, I mean, people who live in that city, it's good but, it can be good but, mean, when other visit, visit us and we need to take them uh to other places, it can be too boring, you know what I mean?
- 2 TT 3: Yeah, yeah
- 3 TT 2: but *whatever* yeah (*laugh*)
- 4 TT 3: *Yeah I agree but I mean some people* travel to se the park too, they, I mean if there's city, there's park you can do whatever you can do in the park, for (separation?) for go to... sometimes there is some park that is a leek or lake.

(Adopted from the pilot data, emphasis added.)

TT #2 expressed his opinion about what should exist in a city and disagreed with TT #1's idea of having a park in a city. Although he supported his argument well, he softened his strong stance in line 3. Saying "*whatever*", he moderated his argument in line 1 to alleviate the argumentative mood between them. Line 4 also shows TT #3's strategy to disagree with TT #2: first he expressed an affirmative expression (e.g., *Yeah I agree*) in

order to show his respect to TT #2's opinion; however, he proceeded to disagree with TT #2 by saying, "*but I mean some people...*" and provided his thoughts. The possibilities of disagreeing with other test-interlocutors require test-takers' a thorough understanding of linguistic expressions and ability to use them appropriately. In other words, while engaging in conflicting as well as cooperative interaction with their test-interlocutors, TTs are to present their strategies to solve communicative problems they encountered, which reveals TTs' ability to interact in ways that are sociolinguistically appropriate.

3) *PA: As an approach to providing information about test-takers' SLA*

Because PA allows for the use of communicative tasks such as jigsaw, information gap tasks, and decision making tasks, it has the potential to provide information on interaction and cognitive strategies that have been linked to successful L2 development and SLA. The interaction between non-native speaking test-takers in PA provides information about test-takers' ability to use interaction strategies for SLA and their cognitive processes and outcomes of solving testing-tasks. Numerous empirical studies in SLA have revealed that interaction assists L2 acquisition as it provides an opportunity for learners to receive input and feedback with which they can test their hypotheses, make changes in, adjust, and modify their L2 if necessary (Gass, 1997; Long, 1996; Mackey, 2002; Pica, 1994; Lyster, 2002 & 2007). Interaction also allows learners to notice differences in their IL and gaps in their already internalized grammar and lexicon. It is also argued that interaction facilitates areas in which learners are still in the process of developing for application to their communication of meaning; it reveals what they notice and do not notice and what difficulty they have in retrieving the use in

conversation (e.g., Leeser, 2004; Mackey, Oliver, & Leeman, 2003; Pica, Lincoln-Porter, Paninos, & Linnell, 1996; Pica, Young, & Doughty, 1987).

While being engaged in interaction, learners employ interaction strategies such as clarification requests, confirmation checks, and comprehension checks when they encounter communication breakdown (e.g., Gass & Mackey, 2007; Gass & Varonis, 1994; Long, 1980 & 1996; Mackey & Philp, 1998). Long (1980 & 1996) defined clarification requests as any expression to elicit clarification regarding preceding utterances, confirmation checks as any expression immediately following an utterance by the listener to elicit confirmation whether the utterance had been correctly understood or correctly heard by the speaker, and comprehension checks are used to confirm listener's understanding. For instance, the lines indicated in the bold types in excerpt 1.5 show test-taker Y's attempt to make a request of clarification.

Excerpt 1.5 E: Eun-mi (a test-taker) Y: Yang (a test-taker)

- 1 E: We have to see, we have to write English, right?
- 2 Y: [laughs] Yeah sure.
- 3 E: In English so maybe you know I think the character, how can I,
[whispers]
- 4 Y: Mm hm
- 5 E: The leETTer, um character is the most
- 6 Y: Yeah
- 7 E: Important part in the uh culture, I mean
- 8 Y: Yeah, you mean charac-**
- 9 E: Yeah
- 10 Y: Character?**
- 11 E: So that infl- that can be influenced on our culture
- 12 Y: Yeah
- 13 E: I just worry about that a little bit. If I can't
- 14 Y: Yeah
- 15 E: Uh prevent our
- 16 Y: You mean**
- 17 E: Yeah language
- 18 Y: Oh you mean**
- 19 E: Yeah maybe
- 20 Y: You mean if we learn another**
- 21 E: Yeah

- 22 Y: Foreign language too much,
- 23 E: Yeah right
- 24 Y: We will lost our culture.
- 25 E: Yeah.

(Brooks, 2009: 356, emphasis added.)

Test-taker Y used clarification requests in lines 8, 10, 16, 18, and 20 in order to elicit clarification from his test-interlocutor. With test-taker Y's interaction strategies, test-taker E modified her choice of lexicon and test-taker Y could avoid possible misunderstanding.

These interaction strategies can lead to more elaboration or modification of utterances for better understanding, which provide information regarding test-takers' current L2 ability. For instance, lines 1, 3, and 5 in excerpt 1.6 show that test-taker #1 tried several attempts to elaborate his utterances to make himself understood clearly.

Excerpt 1.6 E: test-taker #1 M: test-taker #2

- 1 E well (.3) basically: (.8) I like to (inform) I like to negotiate with people
- 2 M Ah[a]
- 3 E [%you] know% DEALing with people
- 4 M yeah [me to-]
- 5 E [talk] to people (you [know])

In comparison to those features in one-on-one interview settings test-takers approach their test-interlocutors in ways that enable them to elaborate their utterances more with greater frequency and detail. This in turn, enables them to modify their speech in ways that make input more accessible and more likely to provide information about their IL status and their attempt to integrate and test their current IL. These interaction strategies that are useful for SLA are presented more in learner-learner dyads, and it is also expected in PA. As revealed in Table 1.4, the high percentage of the instances of

prompting elaboration, finishing sentences, paraphrasing, expressing incomprehension, and other correcting behaviors strongly support this claim.

Test-takers' use of interaction strategies conveys important information about their IL status and their ability to negotiate meaning to understand and be understood. This information can help teachers to have a better understanding of their students' L2 development over time. Furthermore, as studies on learner negotiation in dyads and those using focused tasks have informed us about learners' socio-cognitive processes and outcomes, the PA can document what has been observations recorded informally and data collected under controlled conditions by using communicative tasks such as the information gap task as an assessment instrument (Pica & Lee, 2009).

4) *PA: As an approach to assessment that connects language teaching and testing*

PA is an effective approach to formative assessment of learning as it can be integrated in the teaching and testing well. The recent trend in language education emphasizes the formative assessment approach that is implemented during a course of instruction. The information collected in formative assessment shows to what extent learners understand content and how much they have progressed. It will also help teachers to develop and revise their curriculum accordingly. Analyzing assessment traditions in America, Falsgraf (2009) explained that there are views that consider the purpose of assessment as a device to improve teaching and learning and, argued not standardized testing but formative performance assessments improve education. Moreover, as the test result can be longitudinal, learning progress that is charted over time such as learners' mastery of question formation can be captured through this format.

As mentioned in the previous section, the dyadic format has been widely employed in language classroom with the benefits of facilitating interaction. Learners' familiarity of this format facilitates teachers' decision to use PA as a classroom assessment format. In particular, educational contexts where formative assessment and classroom assessment based on observation during teaching are emphasized, PA can be one of the assessment choices due to its efficiency in terms of time management.

In addition, PA allows test-takers to have more opportunities to talk, which eventually helps them to present their ability less stressfully (e.g., Fulcher, 1996), and testers to gather more information about test-takers. Moreover, classroom logistics such as a large number of students also attracts classroom practitioners' attention to PA. Interviews and surveys reveal the first concern that many classroom teachers have is the lack of time to assess students individually (Butler & Lee, 2004; Iwashita, 1999; Nevo & Shohamy, 1984). PA is more time and cost efficient. It lessens teachers' burdens to assess test-takers individually and shorten the time as teachers should pay attention to each test-taker (Bonk & Ockey, 2003; Ducasse & Brown, 2009; Folland & Robertson, 1976; Hilsdon, 1995; Robinson, 1995; van Moere, 2006). Finally, PA, including group formats, helps test-takers to reduce anxiety during tests. Fulcher (1996) investigated test-takers' reaction to different types of assessments and found that test-takers felt less anxious in carrying out the group discussion task. His finding confirms Berkoff (1985)'s argument that paired assessment is helpful to reduce test-takers' anxiety. The following closing section of this chapter will present unresolved issues and future research directions of PA.

Needs for Further Research

Despite the aforementioned theoretical and practical grounding of implementing PA as part of a testing battery, assessment researchers and practitioners are still cautious about implementing this format rigorously (e.g., Csepes, 2002; Foot, 1999). Their foremost concern is a test-taker's influence on the other test-taker's performance as it is possible that both test-takers may bring their own characteristics in testing settings, which may influence their performance. As cited in Swain (2001), Green (1998) emphasizes this potential influence.

The difficulty with paired reports is that the presence of another individual changes the way in which the task would be approached by an individual working alone on that task. Two individuals *working together on a task interact, and each modifies the behavior of the other*. The manner in which the task is solved by a pair may differ enormously from the way in which either individual might solve the task alone (Green, 1998:49, emphasis added).

Green's concern is also shared by many others and more specified in McNamara (1996).

In the case of a speaking test, for example, the candidate may be required to interact with an interlocutor, who may be another candidate, a trained native speaker, or a highly proficient non-native speaker. *The age, sex, educational level, proficiency/native speaker status and personal qualities of the interlocutor relative to the same qualities in the candidate are all likely to be significant in influencing the candidate's performance* (McNamara, 1996:86, emphasis added).

In spite of those concerns, only a small number of empirical studies are currently available, and only limited variables have been addressed for systematic research studies (Brooks, 2009; Davis, 2009; Galaczi, 2008; McNamara, 1996 & 1997; Swain, 2001; Watanabe, 2008). The variables that those studies have examined include interaction patterns (Galaczi, 2008), personality (Berry, 2007; Bonk & Van Moere, 2004); language ability (Iwashita, 1999; Nakatsuhara, 2006); or acquaintanceship (O'Sullivan, 2002) with respect to test-takers' performance.

Nevertheless, questions about variation remain with respect to test takers' language ability as a factor in obtaining a valid sample of their linguistic accuracy and sociolinguistic appropriateness, as there is a direct connection between these two dimensions of the learners' communicative competence and the learners' communicative language ability. Questions remain therefore as to whether language ability differences between test takers makes a difference in the linguistic accuracy and sociolinguistic appropriateness of their language samples. Theoretical disputes concerning test-takers' performance variation and methodological shortcomings regarding the way to best analyze and interpret the other test-taker's influence call for more rigorous and systematic research studies. I will thoroughly discuss the theoretical and methodological gaps in the previous studies and propose my research questions in the following chapter.

CHAPTER TWO: Literature Review

Introduction

This chapter will provide an overview of the discussion regarding PA as a valid approach to evaluating L2 learners at the process and outcome levels. Questions remain as to whether L2 samples obtained through PA are valid indicators of linguistic accuracy, sociolinguistic appropriateness, and interaction strategies for SLA, in particular, across pairs of same and different L2 developmental stages. The chapter will start with an overview of theoretical and practical concerns regarding language ability and move on to a discussion of studies which have examined the influence of language ability variation on test-takers' performance in PA. The chapter will conclude with research questions.

Language Ability

Language ability generally refers to learners' (test-takers') skills in or ability of speaking, listening to, reading, and/or writing, which are measured based on evaluation criteria (Leeser, 2004; Watanabe, 2008). Nevertheless, its interpretation and evaluation methods have varied. As a new paradigm regarding language learning and acquisition was introduced in the field, new terminology describing the concept has been coined. Those terms include language ability, language knowledge, language use ability, communicative competence, and communicative language ability. For instances, in the late 1970s, proficiency was considered a technical ability to produce language flawlessly with no accent or grammatical errors (Ingram, 1977; Sollenberger, 1978). While excluding individuals' sociocultural understanding, their knowledge of the functions,

content, and accuracy of language was believed to reveal language proficiency, and a quality measured in scales (Ingram, 1978; McNamara, 1999).

More recent model of language ability theorized to describe the state of individuals' language use ability is communicative competence, which is considered to encompass not only individuals' knowledge of language but also their ability to produce appropriately. The concept of communicative competence was suggested in order to advance Chomsky's limited distinction between competence (i.e., what people know about language) and performance (i.e., what people do with language).

The relationship between language ability and communicative competence has been interpreted in many ways, one of which is suggested by Savignon (1983). She equalized language ability and communicative competence, arguing that language ability should be delineated and evaluated as such. This argument was also supported by Bachman (1990), who explained that communicative competence should be the measurement of language ability, and situational information should be incorporated when evaluating test-takers' language ability.

Communicative competence is composed of four sub-competences: 1) grammatical competence of knowledge of words language rules, 2) sociolinguistic competence, which reveals the individuals' ability to use the knowledge appropriately in specific situations, 3) discourse competence, which gauges the level of consistency and cohesion in utterances, and 4) strategic competence, which shows the individuals' ability to employ adequate interaction strategies in given situations (Bachman, 1990; Canale & Swain, 1980; Canale, 1983; Hymes, 1972; Savignon, 1997).

Compared to the previous unitary understanding of language ability, which exceedingly emphasized grammatical competence, communicative competence gives equal values to the contextual and social environments where individuals use language (Bachman, 2009). This extended view of language ability has shed light on its evaluation. That is, more attention has been drawn to test-takers' understanding of sociocultural aspects of language use. This increased attention has led to the introduction of L2 assessment of pragmatics, aptitude, and implicature (McNamara & Roever, 2006), which have provided information about another angle of test-takers' language use ability.

Despite its contributions to the field of language assessment, communicative competence has been dually criticized for its disproportionate attention to individuals-in-isolation and its lack of attention to test-takers in interaction (McNamara, 1996:85). This criticism has grown as group and paired assessment, in which test-takers interact with other test-interlocutors, have been employed more frequently as a classroom assessment and a high-stakes testing tool. In particular, as revealed in the language assessment framework that values another person's presence, the interaction between the two test-takers and the potential variation in test-takers' performance need to be researched more systematically (Green, 1998; McNamara, 1996; Swain, 2001).

Language Ability Variation

The aforementioned new paradigm to view language ability challenges not only language testing researchers but also language teachers. In particular, its interpretation of multicomponential nature of language ability is realized in various ways in language classroom. Surveys and interviews with classroom teachers reveal that managing

different language ability learners who are at different language developmental stages is one of the greatest concerns that language teachers encounter in large classes (e.g., Iwashita, 2001). In conjunction with SLA findings, an example of different language ability of learners is their different L2 developmental stages. Activities and assessments that are too difficult or too easy for learners' current stages can discourage them, negatively influencing their motivation to learn language. Furthermore, when implementing group or pair work, teachers wonder whether different pairing in language development impacts the effectiveness of instructions and activities for learners' language learning.

Some SLA research findings support the claim that the differences between pairs can provide learning opportunities by increasing the quantity and quality of interaction. It is believed that interaction between learners can promote L2 acquisition by helping learners to notice linguistic forms and lexicons and to test their hypothesis about L2 during dyadic interaction (Lantolf 1996; Ellis 2000, 2003; Swain and Lapkin 2000; Skehan 2003). The findings of Gass and Varonis (1985) and Porter (1986) suggested that the discrepancy in L2 ability between learners helps learners to draw attention to their language use and to increase the quantity and quality of interaction and utterance. Learners' language ability also affects how well they resolve language problems encountered during interaction (Leeser, 2004).

In addition, researchers have found that the differences between pairs often define the relationship between learners, as well as their interaction patterns during tasks. While some research studies provide empirical evidence that learners in advanced level can play

a role in leading the interaction by helping a beginning level partner (van Lier, 1996), Kowal and Swain (1994, 1997) found that advanced learners usually dominate interaction in implementing tasks. Storch (2001) found that learners with the biggest differences in L2 developmental stages tended to collaborate more during tasks than those with little difference. Similarly, she found that pairs who are in the same developmental stages cooperated least. While supporting Storch's claim of the relationship between language level differences between pairs and interaction pattern, Watanabe and Swain (2007) added that language ability differences between pairs will promote more interaction, eventually benefiting both higher and lower learners' L2 acquisition. Yule and Macdonald (1990) examined learners' interaction in times of communication breakdown and found that as long as lower level learners have more information while conducting tasks, more interaction can be promoted. They are in general agreement that all the learners in different language ability dyads do not benefit equally, and the processes and outcomes of their L2 acquisition display differently.

The aforementioned SLA research findings are informative to teachers who need to develop tasks, group learners, and evaluate learners' performance. What is unclear, however, is the influence of language ability difference between pairs of learners in testing settings. In contrast to a learning process, testing requires more equal opportunities among test-takers because decision making such as advancement to next level or admission to higher institution is involved. It is also necessary to understand whether or to what extent the quantity and quality of test-takers' language as well as their interaction patterns vary depending on the other test-taker's language ability. Despite the

significance of these issues, this line of research is quite recent, and there is a notable scarcity of research studies in this area. The following sections will discuss some of important studies which examine these issues in language testing.

Research findings

In contrast to the mostly concurring findings in SLA research, there remain controversies over how the language ability differences between pairs influence their performance and scores. Moreover, the scarce body of literature on this topic hinders our understanding of the extent to which language ability differences influence test-takers' performance. The following Table 2.1 shows a number of studies in the field, and the following section will provide a brief summary of the studies.

Table 2.1 Research questions and findings

Studies	Research questions	Influence of the other test-taker influence on test-taker performance
Iwashita (1999)	1. Do test-takers' scores differ in relation to the proficiency of the speaking partner? 2. Does the test-takers' discourse differ according to the proficiency of the speaking partner?	+ (noticeable individual variation)
Csepes (2002)	1. What impact does the partner's proficiency level have on candidates' test scores? 2. Do candidates' scores vary if they have partners of different proficiency levels? If yes, what kind of variation characterizes test scores?	-
Nakatsuhara (2006)	1. Are conversation styles of dyads different between same proficiency-level pairs (SPL) and different proficiency-level pairs (DPL)? 2. Are dyadic interactions with different ability speakers asymmetrical? If so, how are they asymmetrical? To what extent are they asymmetrical?	-
Davis (2009)	1. Does interlocutor proficiency level influence average rating scores? 2. Does interlocutor proficiency level influence the amount of language produced? 3. Is the amount of language produced associated with average rating scores? 4. Is the proficiency level of one's interlocutor associated with the type of interaction produced in the task?	-/+

Test-takers in Iwashita (1999)'s study were paired with higher or lower test-takers and worked on three different tasks (1 two-way task and 2 one-way tasks). She examined test-raters' evaluation of their performance, test-takers' discourse, and questionnaires: test-raters evaluated a range of linguistic and interaction features, which were also examined in transcripts. Her findings showed that high language ability test-takers gained higher mean scores when they were paired with the same language ability test-takers, and low language ability test-takers gained higher mean scores when they worked with higher level test-takers. The analysis of discourse revealed that high language ability test-takers talked more when they were paired with the same language ability test-takers, while low language ability test-takers talked more with higher level test-takers. Nevertheless, there were noticeable individual performance variations and large standard deviations; due to the small number of participants, more rigorous statistical analysis (i.e., inferential statistics) was not conducted.

Running more rigorous statistical analysis with a larger number of subjects, Csepes (2002) did not find any statistically significant results. She investigated the potential influence of a test-taker's language ability variation on the other test-taker's performance. Test-participants were grouped into a core test-taker group and three different language ability non-core test-takers groups. The test-takers worked on three different tasks with three different test-takers, and their performance was rated based on a rubric. After determining there were no statistically significant differences of test-takers' performances across the different test-taker language ability groups, she confuted

concerns about a test-taker's influence on the other test-taker's performance and argued that PA is a fair and valid testing format.

In comparison to the emphasis on quantitative data analysis in the previous study, Naktsuhara (2006) adopted analysis of discourse to determine whether any patterns or differences in test-takers' performance existed. Her study closely examined the discourse pattern of test-takers when they interacted with same or different language ability test-takers in terms of interactional contingency, quantitative dominance, and goal orientation of their conversation. Her findings showed that test-takers' discourse was slightly more contingent when they were paired with same language ability test-takers, but not significantly so. Quantitative dominance and goal orientation tended to be skewed toward higher language ability test-takers. She showed that higher language ability test-takers talked more and initiated more topics when they were paired with lower language ability test-takers. However, her research findings did not reveal strong evidence regarding the influence of a test-taker's language ability variation on the other test-taker.

Davis (2009)'s findings also revealed no influence of a test-taker's language ability differences on the other test-taker's performance in terms of testing scores. He employed Rasch analysis to examine test-rater harshness and the differences among testing scores. This rigorous statistical analysis did not reveal whether a test-taker's language ability difference influences the other test-taker's performance. Nonetheless, as with Iwashita (1999), Davis also confirmed that the amount of talk increased as test-takers were paired with high language ability test-takers.

As revealed in Table 2.1, the findings regarding a test-taker influence on the other test-takers' performance are inconclusive. While Iwashita (1999) showed a difference in test-takers' scores, Csepes (2002) and Davis (2009) did not. Two studies show that the amount of talk increased as test-takers interacted with higher language ability test-takers (Iwashita, 1999; Davis, 2009); however, the performance differences of other features such as dominance or equality between test-takers were not noticeable (Nakatsuhara, 2006). Although these research findings are an informative first step in research, the inconclusiveness of their findings indicates more research on the theoretical framework and methodology they employed. The sections that follow will discuss the construct debate and several methodological shortcomings found in the aforementioned studies.

Construct of PA

One explanation for the mixed results of the abovementioned studies is a lack of thorough discussion of constructs of PA. A construct is defined as an attribute, trait, skill, or ability of a human being, which is "hypothesized in a theory of language ability (Hughes, 2003:31)". Defining constructs in tests is the first and foremost step of developing tests (McNamara, Hill, & May, 2002). Although the format itself is not completely new in the field of language education, PA is a relatively recent approach in assessment. In particular, new perspectives on language ability (e.g., communicative competence) and the influence from SLA research findings have led to an ongoing discussion of the construct of PA. While some researchers pay attention to the *equality* of both test-takers' contribution to solving testing-tasks, others value test-takers' performance *variation*, relying on the other test-taker as the major construct of PA.

Investigating test-raters' interpretation of constructs of PA, Ducasse and Brown (2009) proposed *interpersonal non-verbal communication*, *interactive listening*, and *interactional management skills* as constructs of PA. These constructs value test-takers' attention to the other test-takers and their willingness to sustain interaction. Interactive listening reveals listeners' attempts to show their comprehension of speakers' utterance by employing interaction strategies. Their findings are consistent with SLA research findings regarding negotiation of meaning at times of communication breakdown.

Another way to induce constructs of PA is to apply and revise the concept of communicative competence. As noted earlier, one of the criticisms regarding the current understanding of communicative competence is its strong emphasis on intrapersonal ability rather than interpersonal competence of test-takers. As Chalhoub-Deville (2003) suggested, "ability-in-individual-in-context" can be considered a construct of PA. This argument is similar to McNamara (2001)'s proposal of individuals-in-interaction. McNamara and Roever (2006) argued that traditional language assessments failed to measure test-takers' ability to interact with others. In other words, the significance of evaluating test-takers in interaction should be realized, and the social dimension involved in language performance should be targeted in language assessment. Checking test-takers' ability in ways that are sociolinguistically appropriate while they interact with another test-taker during testing settings should be a construct as well.

The aforementioned discussion regarding the possible constructs of PA led to the development of the following dimensions of constructs: linguistic dimension and interaction dimension. The construct in the linguistic dimension includes test-takers'

knowledge about language, such as structure, phonology, and lexicon. This dimension reflects the original emphasis and understanding of language ability and the roles of language assessment. The targeted features in the interaction dimension encompass test-takers' ability to interact in ways that are sociolinguistically appropriate and strategically useful for SLA. In this dimension, test-takers' ability to understand the situational appropriateness of their behavior and to use interaction strategies such as comprehension checks, confirmation checks, and clarification requests during the communication breakdown can be examined.

The research studies which examined the construct validity of PA did not fully cover the construct discussion. According to Ducasse and Brown (2009), examining evaluation criteria can reveal the interpretation of constructs of PA in each study. They also argued that the evaluation criteria affect the validity of construct (Brown, 2005; Ducasse and Brown, 2009). Analysis of evaluation scales and criteria shed light on the discussion about construct as it expands the validity claim from the content level to the construct level. The following Table 2.2 shows the research questions and target features in their evaluation criteria.

Table 2.2 Evaluation criteria

Studies	Evaluation criteria	Interactional dimension		
		Linguistic dimension Linguistic accuracy	Sociolinguistic appropriateness	Interaction strategies for SLA
Iwashita (1999)	Grammar & expression, fluency, pronunciation, vocabulary, communication strategies, & task fulfillment slowdown, display questions, lexical simplification, comprehension check, fronting, clarification request, grammatical simplification// C-units, turns, and	+	+	+

	ungrammatical utterance			
Csepes (2002)	Communicative impact, grammar and coherence, vocabulary, & sounds, stress, and intonation	+	+	-
Nakatsuhara (2006)	Interactional contingency, goal orientation, and quantitative dominance	-	+	-
Davis (2009)	Grammar & vocabulary, pronunciation, fluency, discourse management, and amount of talk	+	+	-

As revealed in Table 2.2, the evaluation criteria of these studies failed to incorporate the major constructs of PA. They did not provide a full picture of test-takers' performance under the influence of the variation in the other test-takers' language ability.

Methodological shortcomings

The methodological shortcomings in the aforementioned studies are twofold. The first problem arises with regard to test-takers' profile. The second problem resides in their data analysis phases. The following section will describe the shortcomings in detail.

1) Study design: test-takers' profile

The methodological shortcomings that have arisen in the designs of the aforementioned studies are revealed in 1) failing in controlling compounding variables and 2) operationalization of language ability variation. A range of test-taker variables have been chosen for research studies. Test-takers' L1, gender, age, and acquaintanceship have been researched whether they caused any variation in the other test-takers' performance. While some of them showed its influence on test-takers' performance, for instance, gender and L1, other variables such as acquaintanceship did not affect test-takers' performance. First, as shown in Table 2.3, the aforementioned studies did not control the independent variables regarding test-takers such as age, gender,

acquaintanceship, and first language. These uncontrolled variables may have allowed for mixed results.

Table 2.3 Uncontrolled variables

Uncontrolled variables	Age	Gender	Acquaintanceship	L1
Iwashita (1999)	+	-	N/A	-
Csepes (2002)	+	-	+	+
Nakatsuhara (2006)	+	-	N/A	-
Davis (2009)	+	-	N/A	+

Despite the research findings that revealed that gender is one of the most influential factors on test-takers' performance (e.g., Lazaraton & Davis, 2008), as the studies above did not control test-takers' gender, it is challenging to determine whether the variation in test-takers' performance was caused by their gender or language ability. However, it is acceptable that the aforementioned studies ignored the acquaintanceship between test-takers: as revealed in O'Sullivan (2002), acquaintanceship between test-takers and test-interlocutors did not affect test-takers' performance. In terms of L1, robust findings in SLA studies on the influence and roles of learners' L1 on their IL development convince us that L1 should be controlled in testing setting as well. Potential influence from those variables may help to explain the controversial findings of the language assessment studies.

Unclear operationalization of the independent variable (i.e., language ability) in their studies also makes us question their research findings. In the studies, as indicated in Table 2.4, some of the base criteria of test-takers' language ability employed in order to differentiate their levels are not exclusive of what they have used for their performance

evaluation. For example, the evaluation results of interviews in Iwashita's study and monologues in Davis' study what they measured in test-takers' performance in relation to the other test-takers' language ability.

Table 2.4 Base and performance criteria

Study	Screening test format	Test format	Base criteria	Performance criteria
Iwashita (1999)	Interview Survey of teachers' comments Length of target country experience	PA	Fluency Grammar Listening comprehension ²	Grammar & expression, fluency, pronunciation, vocabulary, communication strategies, & task fulfillment slowdown, display questions, lexical simplification, comprehension check, fronting, clarification request, grammatical simplification// C-units, turns, and ungrammatical utterance
Csepes (2002)	A cloze test, a self-assessment questionnaire, a teacher-assessment questionnaire	PA	Grammar reading comprehension Self evaluation of their language ability	Communicative impact, grammar and coherence, vocabulary, sounds, stress, and intonation
Nakatsuhara (2006)	Participants' self-report of their testing scores and Cambridge common scale for speaking test (CPE, CAE, and FCE levels)	PA	Grammar and Vocabulary Discourse Management Pronunciation Interactive Communication ³	Interactional contingency, goal orientation, and quantitative dominance
Davis (2009)	Self-reported scores on National Matriculation English Test & Monologue test	PA	Self-evaluation of their language ability Grammar	Grammar & vocabulary, pronunciation, fluency, discourse management, and amount of talk

² As Iwashita did not specify the base criteria in her study in 1999, these were inferred from Iwashita (2001) which used the same data.

³ These criteria were taken from the Cambridge common scale for speaking test as Nakatsuhara stated in her article.

Shown above, the screening tests and target test (i.e., PA) were distinctive; however, the base criteria and performance criteria are not exclusive. Unless the two descriptors of language ability are clearly operationalized and exclusive, it is challenging to avoid tautological arguments, which will eventually question the necessity of implementing PA. In other words, 'language ability' to identify the different groups of test-takers and to describe the test-takers' transcripts and scores should carry unique information.

Language developmental stages

2) *Incomplete data analysis*

The next potential explanation for the mixed results in the aforementioned studies is the incomplete data analysis. Early language testing validation research has mainly examined test-takers' scores and ratings in order to understand the patterns of test-taker performance, as well as psychometric qualities of a test, such as reliability and validity (Bachman, 1990; Banerjee & Luoma, 1997; Lazaraton, 2008; Lumley & Brown, 2005). This quantitative approach usually adopts statistical procedures such as *correlation* for examining similarities in test-takers' performance in different situations. In addition, the *ANOVA/MANOVA (or t-test)* procedure is employed to examine whether test-takers' performance on several occasions is different (Lumley & Brown, 2005). These quantitative data analyses help to capture general trend of data and is relatively straightforward to run using statistical analysis suites. However, these quantitative data analyses do not always show the detailed or individualized information of test-takers' performance. Empirical studies have shown that despite receiving the same scores, in the

nature of test-takers' performance and test-raters' rationale in assigning particular scores (Douglas, 1994; Douglas & Selinker, 1993).

Recently qualitative data analysis approaches have been adopted in order to examine test-takers' performance from another angle (Banerjee & Luoma, 1997; Lazaraton, 2008; McNamara, Hill, & May, 2002). Originally the qualitative approach was employed reluctantly because of apprehension related to its subjectivity. However, it is revealed that this approach can provide information about the content of test-takers' language and the processes of their language development in detail. Furthermore, the relationship between test-takers' performance and the scores that they receive can be revealed through qualitative approach (Galaczi, 2008). Among a range of qualitative data analysis methods such as verbal protocol, observations, questionnaires, and analysis of discourse (i.e., text, discourse, and conversation analysis), analysis of discourse is discussed to be a fairly suitable method to analyze the nature and variation of test-takers' language produced during PA (Banerjee & Luoma, 1997; Shohamy et al., 1993).

Understanding the advantages of employing both quantitative and qualitative data analysis approaches mentioned earlier, those four studies which look at the other test-takers' language ability variation and its influence on test-takers' performance were revisited. As revealed in Table 2.5, two studies that did not find any test-taker influence on test-takers' performance showed incomplete data analysis.

Table 2.5 Data analysis methods of the studies

Studies	Data analysis		Influence of the other test-taker language ability on test-takers' Performance	
	Test scores	Analysis of discourse	Test scores	Analysis of discourse
Iwashita (1999)	+	+	+	+
Csepes (2002)	+	-	-	N/A
Nakatsuhara (2006)	-	+	N/A	-
Davis (2009)	+	+	-	+

(+ indicates the given information is available and - means the information is not available.)

Studies which employed either quantitative or qualitative data analysis methods did not prove that test-takers' performance varied depending on the other test-takers' language ability (i.e., Csepes, 2002 & Nakatsuhara, 2006). Their conclusions are questionable due to incomplete nature of data analysis. Because the test-raters' evaluation in Csepes (2002) was not normally distributed, she ran non-parametric analysis (Chi-square analysis), which did not show statistically significant results. Furthermore, the choice of data analysis method does not provide convincing information. For instances, Csepes (2002) argued that she had to run non-parametric analysis (i.e., Chi-square analysis) since her data was not normally distributed. The shortcomings of her analysis method are the low level of power and inadequate choice of method. That is, the result of nonparametric analysis usually shows a weak statistical relationship among variables. In addition, as her data was not frequency based, Chi-square analysis was not adequate. Another statistical analysis can be employed to see whether the results may be different, and analysis of discourse in test-takers' performance may provide another aspect of the data. In comparison, Nakatsuhara (2006)'s study only looked at the transcriptions of her test-

takers to examine interactional features such as interactional contingency, goal orientation, and quantitative dominance. She reported no influence of the other test-takers' language ability differences on test-takers' performance in terms of those interaction features was found. However, testing scores and the relationship between testing scores and interaction pattern may provide more adequate information in terms of test construction and interpretation.

Although both quantitative and qualitative analysis methods were employed and similar evaluation criteria were used, the studies conducted by Iwashita (1999) and Davis (2009) showed contradictory results in testing scores. However, Iwashita's concerns regarding noticeable individual variations in test-takers' performance might be more related to her approaches to data analysis. As she overlooked the large standard deviation in test-takers' scores as well as in depth statistical analysis of testing results, her conclusions, therefore, are worthy of reconsideration.

Based on the aforementioned theoretical and empirical concerns regarding test-takers' performance in PA in relation to the other test-takers' language ability variation, the following four research questions have been developed.

Research questions

1. Does PA test-takers' use of grammatically accurate L2 utterances vary in relation to the developmentally-equal and unequal status of their pairing?
2. Does PA test-takers' use of sociolinguistically appropriate and interactionally strategic L2 utterances vary in relation to the developmentally-equal and unequal status of their pairing?
3. Does PA test-raters' rating of linguistic accuracy vary in relation to the developmentally-equal and unequal status of their pairing?

4. Does PA test-raters' rating of sociolinguistic appropriateness and interaction strategies vary in relation to the developmentally-equal and unequal status of their pairing?

The following chapter will provide detailed information regarding the methods of data collection and analysis that were developed and used to answer these research questions.

CHAPTER THREE: Methodology

Introduction

This chapter describes the methodology used in collecting and analyzing linguistic and interaction dimensions of condition-receiving test-takers (CRTT)' performance in paired-assessment (PA) in order to examine its validity.

Participants and Setting

Test participants: Condition-receiving test-takers and condition-giving test-takers

Test-participants, comprised of the condition-receiving test-takers (CRTT) and condition-giving test-takers (CGTT) in this study, were 30 adult female Mandarin Chinese speakers who were recruited from a large university in the United States. The age of the test participants spans from 21 to 36. Each participant has completed at least ten years of formal English language study prior to beginning university studies, typically beginning in the third year of primary school (approximately ages nine to ten) and continuing to the sixth year of secondary school (approximately ages fifteen to seventeen). The average number of years of prior formal English instruction was between ten and twenty years. Potential compounding variables, such as test-participants' gender, age, nationality, and social status, were controlled at the recruiting stage. Detailed demographic information on each participant is available in Appendix A. The test-participants were divided into two groups: condition-receiving test-takers (CRTTs) and condition-giving test-takers (CGTTs), as described below.

Condition-receiving test-takers (CRTTs)

Among the total number of 30 test-participants, fifteen were *CRTTs* whose performance in the PA setting was evaluated. Their level was determined by the L2 developmental stages of their English question acquisition revealed in the screening tests (Spada & Lightbown, 1993; Pienemann & Johnston, 1987; White, Spada, Lightbown, & Ranta, 1991). More detailed information about their testing scores will be explained later.

Condition-giving test-takers (CGTTs)

The remaining fifteen test-participants were *CGTTs*, whose performance during PA was not evaluated. However, the results on the screening tests classified them into three different groups: five *CGTTs* at higher developmental stage (*CGTTH*), five at same developmental stage (*CGTTS*), and five at lower developmental stage in relation to that of the *CRTTs*. Each test taker's performance was analyzed for behavioral similarities and differences when she was working with a *CGTTH*, a *CGTTS*, and a *CGTTL* interlocutor.

Test-raters

Two test-raters who were native speakers of North American English evaluated the *CRTTs*' performance. These test-raters had at least three years of experience teaching English as a foreign or L2 at university-based language institutes in the United States and abroad.

Variables

The *independent variable* was the developmental stages of *CGTT*'s question formation. Their L2 developmental stages were determined by their acquisition of

English question formation through screening tests. The *dependent variable*, sub-divided into linguistic dimension and interaction dimensions, was the CRTT's performance in response to three different CGTT's developmental stages. The linguistic dimension examined the global accuracy of CRTT's language, and the interaction dimension assessed CRTTs' ability to interact in ways that are sociolinguistically appropriate and strategically useful for SLA. Each variable will be explained in the following sections.

Operationalization of an Independent Variable

Condition-giving test-takers' developmental stages

The developmental stages of English question formation was chosen as an independent variable. It was chosen due to the robust and linear nature of its acquisition order. Empirical research findings have shown that the developmental stages of English question formation are invariable. That is, the development of question formation is linear in both English as a Second Language and English-as-a-Foreign Language settings. Furthermore, it shows linear developments in instructed as well as natural language learning settings, which makes the development of question formation an accurate indicator to discriminate learners depending on their levels (e.g., Pienemann & Johnston, 1987; Pienemann, Johnston, & Brindley, 1988, Mackey, 1999). Table 3.1 shows the stages of question formation and the examples.

Table 3.1 Description of stages of English question acquisition

Stage	Description of stage	Examples
2	SVO Canonical word order with question intonation	<i>It's a monster?</i> <i>Your cat is black?</i> <i>You have a cat?</i> <i>I draw a house here?</i>
3	Fronting: <i>Wh/Do/Q-word</i> Direct questions with main verbs and some form of fronting	<i>Where the cats are?</i> <i>What the cat doing in your picture?</i> <i>Do you have an animal?</i> <i>Does in this picture there is a cat?</i>
4	Pseudo Inversion: Y/N, Copula In yes/no questions an auxiliary or modal is in sentence-initial position. In <i>wh</i> -questions the copula and the subject change positions.	(Y/N) <i>Have you got a dog?</i> (Y/N) <i>Have you drawn the cat?</i> (Cop) <i>Where is the cat in your picture?</i>
5	Do/Aux-second Q-word → Aux/modal → subj (main verb, etc.) Auxiliary verbs and modals are placed in second position to <i>wh</i> -questions (and Q-words) and before subject (applies only in main clauses/direct questions).	<i>Why (Q) have (Aux) you (subj) left home?</i> <i>What do you have?</i> <i>Where does your cat sit?</i> <i>What have you got in your picture?</i>
6	Cancel Inv, Neg Q, Tag Q (Canc Inv) <i>Can you see what the time is?</i> Cancel Inv: <i>Wh</i> -question inversions are not present in relative clauses. Neg Q: A negated form of <i>do/Aux</i> is placed before the subject. Tag Q: An Aux verb and pronoun are attached to end of main clause.	(Canc Inv) <i>Can you tell me where the cat is?</i> (Neg Q) <i>Doesn't your cat look black?</i> (Neg Q) <i>Haven't you seen a dog?</i> (Tag Q) <i>It's on the wall, isn't it?</i>

(Taken from Mackey, 1999. p.567)

Test-participants took screening tests that were composed of a scrambled questions task, a preference task, and a picture-cued task. As taken from Spada and Lightbown (1993 & 1999) and White, Spada, Lightbown, and Ranta (1991), these tasks were used to elicit test-participants' knowledge and production of the English question formations. The first two tests were mainly used for the purpose of screening the participants. That is, the third task, a picture-cued task, was only used to confirm participants' L2 developmental stages. The screening test tasks were cross-checked by four native speakers of North American English, and two items which were controversial among them and also required cultural background were withdrawn. It will be explained

more in detail later in this chapter. The Cronbach's alpha of the screen test of these two combined tests was = .88.

Based on the results from the screening test, test-participants who scored lower than 71% correct were classified into lower developmental stage group, 75 – 95% was in the same developmental group, and 98 – 100% was the higher developmental group as shown in table below.

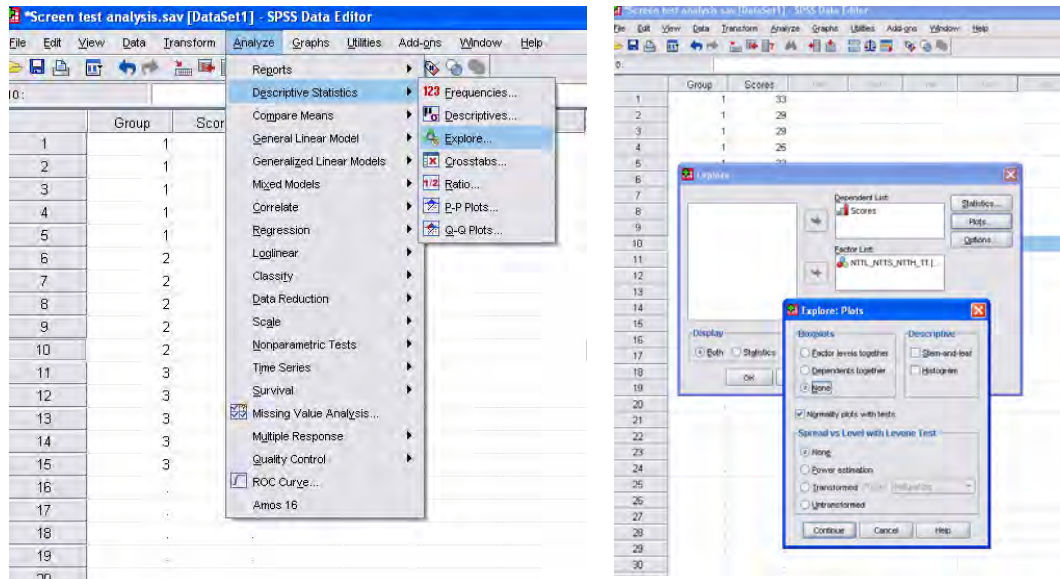
Table 3.2 Independent Variable

CGTT groups	
CGTTH (Higher Developmental stage Group)	98 – 100 % correct
CGTTS (Same Developmental stage Group)	95 – 78% correct
CGTTL (Lower Developmental stage Group)	71 – 53% correct

These three CGTT groups were distinctive in terms of their developmental stages diagnosed by three types of tasks which examined test-participants' developmental stages of forming English questions; however, other potential compounding variables such as their social status, age, nationality, and gender were controlled at the screening stage.

Statistical analysis of CGTT groups' performance on the screening test revealed that the three CGTT groups are distinctive to each other. Before ANOVA was run, the normality of distribution of data was performed. This procedure was necessary since normality is one of the assumptions in order to run ANOVA. The normality test was taken in the SPSS program. As shown in figure 3.1, the normality test option found in the descriptive statistics (explore) was run.

Figure 3.1 Processes of the normality test



Then the test yielded the following result.

Table 3.3 Tests of normality

Tests of Normality^b

		Kolmogorov-Smirnov ^a			Shapiro-Wilk		
		Statistic	Df	Sig.	Statistic	df	Sig.
Scores	1	.221	5	.200*	.953	5	.758
	2	.267	5	.200*	.939	5	.656

a. Lilliefors Significance Correction

*. This is a lower bound of the true significance.

b. Scores is constant when CGTTL_CGTTTS_CGTTTH_CRTT = 3. It has been omitted.

The Shapiro–Wilk test examines the null hypothesis that data is a normally distributed.

The test result was not statistically significant ($p > .05$). That is, the null hypothesis (i.e.,

the data is normally distributed.) is not rejected. As this data met the normality

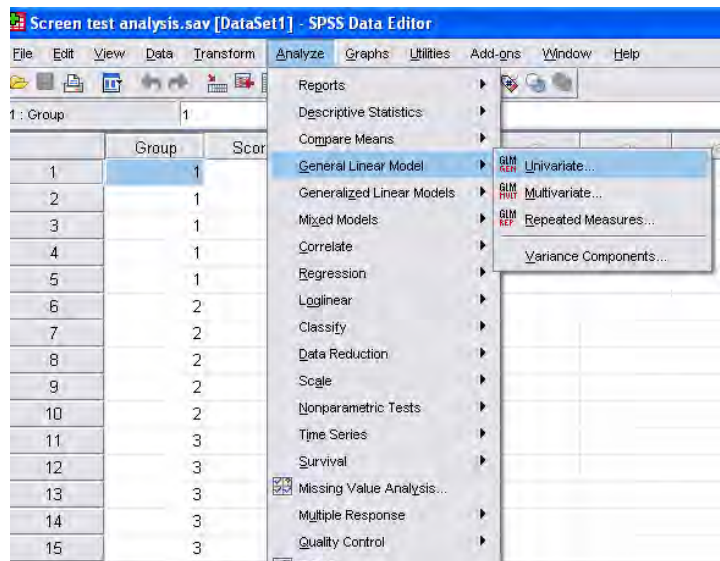
assumption, the following steps were taken to run ANOVA. First, the test-participants

were roughly grouped into CGTTHs, CGTTs, and CGTTLs depending on their raw

scores. In the EXCEL program, each group was assigned a number (e.g., CGTTLs – 1,

CGTTSs – 2, and CGTTHs – 3). Second, the scores that test-participants in each group obtained were entered in the program. Third, this information in EXCEL file was transferred to the SPSS spread sheet. Third, a one-way ANOVA was run; *groups* was chosen as a fixed factor and *scores* was chosen for dependent variable. The following Figure 3.2 shows the number of groups, scores, and analysis taken.

Figure 3.2 Processes of analysis



As shown below, $F(2, 12) = 24.7, p < 0.01$. The differences among the three CGTT groups were statistically significant. The effect size (Partial Eta Squared) is large (.8).

Table 3.4 Analysis results

Levene's Test of Equality of Error Variances

Dependent Variable: Scores

F	df1	df2	Sig.
7.338	2	12	.008

Tests of Between-Subjects Effects

Dependent Variable: Scores

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Corrected Model	377.733 ^a	2	188.867	24.742	.000	.805
Intercept	17681.667	1	17681.667	2.316E3	.000	.995
Group	377.733	2	188.867	24.742	.000	.805
Error	91.600	12	7.633			
Total	18151.000	15				
Corrected Total	469.333	15				

a. R Squared = .805 (Adjusted R Squared = .772)

Post-hoc analysis was conducted to determine whether the differences were meaningful.

Table 3.5 Multiple Comparisons

Dependent Variable: Scores	(I) Group	(J) Group	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Tukey HSD	CGTTLs	CGTTSs	-7.40*	1.747	.003	-12.06	-2.74
		CGTTHs	-12.20*	1.747	.000	-16.86	-7.54
	CGTTSs	CGTTLs	7.40*	1.747	.003	2.74	12.06
		CGTTHs	-4.80*	1.747	.043	-9.46	-.15
	CGTTHs	CGTTLs	12.20*	1.747	.000	7.54	16.86
		CGTTSs	4.80*	1.747	.043	.15	9.46

As shown in Table 3.5, the differences among the three CGTT groups are statistically significant and all the groups are different to each other.

Operationalization of Dependent Variables

The dependent variables in this study have been selected and operationalized based on the previous studies on PA as well as SLA. As such, they were categorized into linguistic and interaction dimensions: in the linguistic dimension, global grammatical

accuracy was examined. The interaction dimension was sub-divided into sociolinguistic appropriateness and strategies for SLA.

These variables (i.e., linguistic accuracy, sociolinguistic appropriateness, and interaction strategies for SLA) studied in each dimension provided opportunities to observe CRTTs' ability to produce grammatically appropriate language and interact with CGTTs in ways that are sociolinguistically appropriate as well as strategically useful for SLA. They have been also chosen as they are major components of communicative competence (e.g., Bachman & Palmer, 1996; Canale, 1981; Canale & Swain, 1980; Hymes, 1972; Savignon, 1997) and have shown to vary under different conditions of spoken interaction (e.g., Young, 1995; Tarone, 1985 & 1988). These dimensions will be further explained in the following sections.

Linguistic dimension

The linguistic dimension of the CRTTs' performance in this study focused on the global grammatical accuracy of their language production. Although examining global grammatical accuracy is the most comprehensive approach to detect all the errors that learners make, it has not been able to support or guarantee high consistency of rater evaluation (Iwashita, 2001; Iwashita et al., 2008). Hence in this study, target grammatical features of the study were determined in advance to increase the consistency of rating between two test-raters' evaluation. In light of the findings of Iwashita et al (2008), the target grammatical features included were 1) morphological features such as verb tense, third person singular, and plural markers and 2) syntactical features such as prepositions, article use, and word order. These foci prevented raters from being

distracted by phonology or word choices of the CRTTs. Moreover, CRTTs' mastery of some of these features is known to be developmentally determined (e.g., Bailey, Madden, & Krashen, 1974; Dulay & Burt, 1974; Goldschneider & DeKeyser, 2001; Pica, 1983; Pienemann & Johnston, 1987); however, some studies show that CRTTs' ability to use other features can vary depending on the context of interaction as well (e.g., Tarone, 1985, 1988, & 1999; Tarone & Liu, 1995). In other words, some features are sensitive to L2 development and have been shown to be sensitive to variation in several linguistic and psychosocial areas.

The quantification of linguistic dimension is adopted from Foster and Skehan (1996) and Skehan and Foster (1999). In addition, in order to make the quantification more applicable to spoken data, Crookes (1990) was also consulted. In particular, the nature of spoken data (e.g., fragments, short idea units, incomplete sentences, etc.) supported the adoption of *utterance* as a unit of analysis (Crookes, 1990; Long, 1980; Luoma, 2004; Sato, 1985), which was defined as "a sequence of speech produced under a single intonation contour bounded by pauses" (Sato, 1985: 83-4). By including an utterance as a unit of analysis in the linguistic dimension, the quantification of the linguistic dimension in this study was the proportion of error-free utterances (maximum value of 1). In this study, an error-free utterance was defined as an utterance in which there was no error in obligatory contexts for its suppliance and/or no error of over-suppliance of any grammatical features in contexts where suppliance was not appropriate. When more than an error was detected in an utterance, only an error was counted.

Interaction dimension

The interaction dimension examined CRTTs' ability to interact in ways that are 1) sociolinguistically appropriate so that contribute to sociolinguistic competence and 2) strategically useful for SLA. More specifically, this study operationalized sociolinguistic appropriateness as the degree of interactional consistency (e.g., Jones & Gerald, 1967; Young & Milanovic, 1992; van Lier, 1989). Interactional consistency is the degree of explicit cohesiveness of CRTTs' utterances which are related to what was previously produced by their CGTTs. Examples include using expressions produced by CGTTs or expressions which contain explicit connotation of agreement such as "as you said~", "I agree", "I disagree", "I don't think so", "I think so" or using the same words.

Interactional consistency was quantified in this study as follows: the proportion of utterances initiated by CRTTs that were used in the subsequent utterances by CGTTs in relation to the total utterances produced by the CRTTs. As these features have been defined and used in a range of ways in previous studies, the following example is presented to explain the feature of explicit coherence used in this study.

Excerpt 3.1 Explicit coherence

1. CRTT#6: Yea:h it's like a fairy tale. (3) Beautiful girl!?=
2. CGTTL#2: Mmhmm?
3. CRTT#6: Princess, prince, and then and then cute some dogs!
4. CGTTL#2: **Yeah I agree with you?** I I would choose this too? And I think maybe we have the thing interesting.

(Dissertation data. Emphasis added.)

CGTTL#2 develops her utterance related to the previously produced utterance using explicit expressions such as "as you said~", "I agree", "I disagree", "I don't think so", "I think so" or using the same words.

Interaction strategies for SLA were operationalized into the degree of CRTTs' uses of interaction strategies such as clarification requests, confirmation checks, or comprehension checks in order to assist the SLA process. CRTTs' use of these interaction strategies shows their ability to assist the SLA processes as well as their willingness to sustain interaction with their CGTTs, regardless of their CGTTs' language ability. Excerpt 3.2 shows an instance of clarification request.

Excerpt 3.2 CRTT #3 and CGTT#2: same language ability CGTT #2

1. CRTT#3: some people need see the natural
2. CGTT#2: Mm hm
3. CRTT#3: and like to live the natural and uhh they use it like the every time for him to go the park, to walk, to do sport
4. **CGTT#2: You mean this picture?**
5. CRTT#3: Yeah.
6. CGTT#2: OK.
7. CRTT#3: This is what I... yeah. so also you can take your dogs to there,
8. CGTT#2: right

(Data from the pilot study)

In this example, the CRTT #3 and the same level CGTT#2 were asked to choose two facilities that can improve a city's living condition. In line 4, the CGTT#2 employed clarification request (*i.e.*, you mean this picture?) with an indication of "you mean". It led to more elaboration or modification of utterances by the CGTT in line 7. The CRTT#3 elaborated the point he made formerly in line 3. These interactional strategies yield modification of CRTTs' language and assist learners to draw their attention to their own language use. It eventually leads to their L2 acquisition. That is, it is assumed that CRTTs will not try to request, make confirmation checks, or comprehension checks, if they think it is not necessary to employ these strategies to push their interlocutors. This will show CRTTs' judgment regarding the necessity of those strategies and their willingness to make interaction flow easily. These features were quantified as follows:

the proportion of CRTTs' utterances of clarification requests, confirmation checks, or comprehension checks in relation to the total utterances produced by CRTTs. Table 3.6 shows a summary of operationalization, quantification, and target features in each dimension.

Table 3.6 Operationalization of dependent variables

Dimensions	Linguistic dimension CRTTs' ability to produce grammatically accurate language	Interaction dimension CRTTs' ability to interact in ways that are 1) sociolinguistically appropriate and 2) strategically useful for SLA	
	Linguistic accuracy	Interactional contingency	Interaction strategies for SLA
Operationalization & Quantification	The degree of global grammatical accuracy in CRTTs' language - The proportion of error-free utterances in relation to total utterance (In total correct/total contexts for suppliance)	the degree of cohesiveness in CRTT's utterance in relation to that of the preceding utterance produced by her CGTTs	The degree of interaction strategies claimed to contribute to successful SLA. -The proportion of CRTTs' interaction strategies in relation to the total utterances produced by CRTTs
Rationale to choose the variable	Linguistic accuracy is used to judge CRTTs' ability to sustain linguistic accuracy regardless of CGTTs' ability.	Sociolinguistic appropriateness is used to judge CRTTs' ability and willingness to develop coherent utterances regardless of CGTTs' ability.	Strategies for SLA are used to judge CRTTs' ability and willingness to sustain interaction regardless of CGTTs' ability.
Target features	Global grammatical features	Expressions such as "as you said~", "I agree", "I disagree", "I don't think so", "I think so" or using the words produced by CGTTs earlier.	Clarification requests Confirmation checks Comprehension checks

The section following will explain the materials used in this study in detail.

Materials

The following sections will sequentially describe the materials used in this study.

The materials included the advertisement flyers and e-mail for soliciting participation, the information regarding the screening test, the web-based background questionnaire, testing tasks, and the exit questionnaire.

Advertisement/e-mail for Soliciting Participants⁴

Brief information regarding this study was included in the recruitment advertisement and e-mail (Appendix B), which were posted in a large urban university campus, a university-affiliated language institute, and common areas for the language institute students. In addition, recruiting e-mails were sent out to the head of the language institute and the Chinese students and scholars association at the University to solicit prospective test-participants

When prospective test-participants responded, an e-mail which includes information about the screening test was sent out.

Screening Test

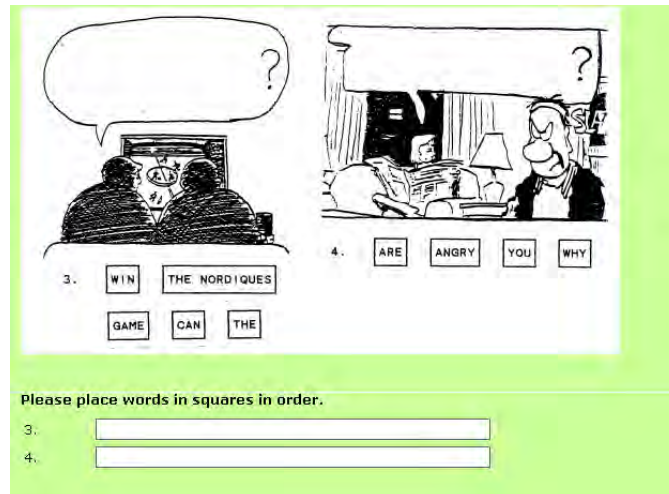
The screening test included a scrambled questions task, a preference task, and a picture-cued production task. These three tasks are widely used to examine the L2 developmental stages in terms of English question formation (Spada & Lightbown, 1993, 1999; White, Spada, Lightbown, & Ranta, 1991). The screening test was computerized and posted on the surveymonkey.com website (<http://www.surveymonkey.com/s/LeeDissertation>). Prospective test-participants received an e-mail with the link and were asked to take the test. The results of their performance were available as soon as they completed the test. The screening test was scored 1 when the prospective test-participants got the question correct and 0 when they got it wrong. The raw scores and their developmental stages were reported in Appendix A.

⁴ These advertisement and e-mail were submitted for the approval by the Institutional Review Board of the university where the researcher attends, and was approved on October 9, 2008.

Scrambled questions task

The scrambled questions task asked test-participants to re-arrange words to make interrogative sentences which explained cartoons. There were 20 questions on this section. Figure 3.3 shows an example of this task.

Figure 3.3 Scrambled questions task

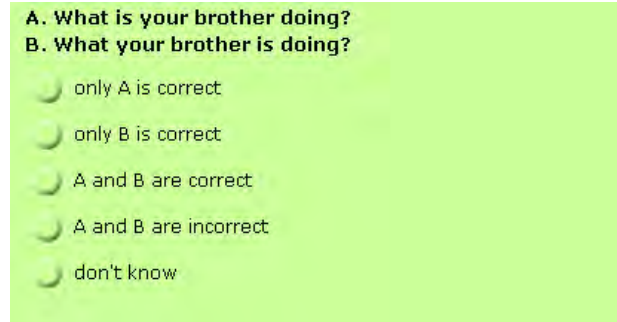


There were words in boxes as shown in the picture. Test-participants were asked to re-arrange the words into an interrogative form. For instance, in question number 3, they were tested whether they could create a sentence, “can the Nordiques win the game?” with the given words.

Preference task

Preference task asked test-participants to choose a grammatically correct sentence among choices. Figure 3.4 shows an example of this task.

Figure 3.4 Preference task



Production task

The final portion depicted in Figure 3.5 is a production task. Test-participants were asked to write eleven interrogative sentences to describe a picture. Test-takers' performance on this task was quantified based on the stages to which they belong. For instance, when a test-participant wrote, "why are you crying?", number 5 was given. The average of the assigned developmental stages on each sentence determined test-participants' final developmental stage.

Figure 3.5 Production task



Participants completed the test on their own; however, they were instructed not to refer to any outside sources such as dictionaries, grammar books, or people. Once they

completed the test, their answers were stored online, and the researcher downloaded and graded the tests. Four native speakers of English and two non-native speakers of English were consulted for the most appropriate answers for the test. If there were any discrepancies, more thorough discussion was done and an answer was chosen. A question (#19) in the scrambled task was thrown out since it required cultural background which many of the participants did not have (Appendix C).

Background Questionnaire

Consistent with Gass and Mackey (2007), a web-based background questionnaire (<http://www.surveymonkey.com/s/LeeDissertation>) was distributed along with their screening test.

The following figure (3.6) is an example of this portion.

Figure 3.6 Background questionnaire



How long have you studied English?

- 1 - 3 years
- 4 - 6 year
- 7 - 9 years
- 10 - 15 years
- 15 - 20 years

How long have you lived in an English speaking country?(including current residence in America.)

- 1 - 6 month
- 7 - 12 months
- 18 month - 2 years
- 3 - 5 years
- 6 - 8 years
- more than 10 years

This survey was used to collect potential test-participants' biographic, linguistic, and language learning information and was used to secure homogeneity among the participants at the screening stage.

Testing Tasks

In order to elicit CRTTs' performance, three decision-making tasks were employed. As noted in Table 3.4, a decision-making task allows a two way exchange of information between task-participants and a flexibility of interaction obligation between them. The possibility of observing CRTTs' ability or willingness to gain, maintain, and yield the conversational floor and cooperate with their interlocutor is a particular strength of this type of task (Doughty & Pica, 1986; Pica & Doughty, 1985; Pica, Kanagy, & Falodun, 1993). Moreover, decision-making tasks can create situations in which task-participants question, clarify, and modify their utterances (Duff, 1986).

Table 3.7 Communication task types for L2 research and pedagogy analysis based on: Interactant (X/Y) relationships and requirements in communicating information (INF) to achieve task goals

Task type	INF holder	INF requester	INF supplier	INF requester-supplier relationship	Interaction requirement	Goal orientation	Outcome options
Jigsaw	X & Y	X & Y	X & Y	2 way (X to Y & Y to X)	+ required	+ convergent	1
Information Gap	X or Y	X or Y	X or Y	1 way > 2way (X to Y/ Y to X)	+ required	+ convergent	1
Problem-solving	X = Y	X = Y	X = Y	2 way > 1 way (X to Y & Y to X)	- required	+ convergent	1
Decision-making	X = Y	X = Y	X = Y	2 way > 1 way (X to Y & Y to X)	- required	+ convergent	1+
Opinion Exchange	X = Y	X = Y	X = Y	2 way > 1 way (X to Y & Y to X)	- required	- convergent	1+/-

(Adopted from Pica, Kanagy, & Falodun, 1993: 19)

The tasks used in this study were adopted from the Cambridge ESOL First Certificate of English, third section. Shown in Figure 3.7, test-participants were given the following pictures with instructions (Appendix D & E).

Figure 3.7 Instruction

Test 1
The film club at your college has asked you to choose two films which would be interesting for the students to watch and then discuss. Here are the films they are considering. First, talk to each other about how interesting these different types of film would be. Then decide which two would be the best for students to discuss. Feel free to take notes of your thoughts and expressions you want to use during the speaking assessment.



Partner Name: _____

Your choices: 1. _____ 2. _____

Reason for choice 1:

Reason for choice 2:

Expressions or words you want to use during the test:

Go to the next page

Test-participants listened and read the instructions and looked at the pictures. During the given time, they thought about expressions and contents they were going to say later. As presented in table 3.8, the three decision making tasks were equivalent in terms of the procedure, the number of prompts, and the allotted time.

Table 3.8 Testing tasks

	Testing task 1	Testing task 2	Testing task 3
Type	Decision-making task	Decision-making task	Decision-making task
Topic	Choosing a movie which would be most interesting for the students at school	Suggesting an event to a local café which attracts people most	Choosing two things which can please people in a city most
Prompt	7 pictures	7 pictures	7 pictures
Suggested interaction time	3 minutes	3 minutes	3 minutes

There were three versions of testing materials in which the sequences of the three tests were alternated in order to rule out any sequence effect of testing-tasks.

Exit Questionnaire

After completing each PA, test-participants were asked to fill out the exit questionnaire. This questionnaire gauged CRTTs' reaction regarding factors including the difficulty of testing tasks, their own language ability level, CGTTs' language ability level, and familiarity with CGTTs. The data from this questionnaire did not play a major role in analyzing the results of this study; however, it provided a basis for understanding the potential influence of these factors on CRTTs' performance.

Rubric

Table 3.9 is the rubric that the test-raters used, which was composed of linguistic accuracy, sociolinguistic appropriateness, and interaction strategies for SLA.

Table 3.9 Rater rubric

Linguistic accuracy	Sociolinguistic appropriateness – interaction consistency	Interaction strategies
<p>5 – Very rare errors/most of the time no errors</p> <p>4 – Errors are rare, mostly morphological including seemingly pronunciation issues (e.g., choose vs. chose)/ Sometimes errors</p> <p>3 – Errors are common, but mostly morphological including seemingly pronunciation issues (e.g., choose vs. chose)/usually errors</p> <p>2 – Errors are usual, mix of morphological and syntactical such as word order including phrasal level, articles, or prepositions/often errors</p> <p>1 – Many errors, mostly syntactical such as word order including phrasal level, articles, or prepositions</p>	<p>4 – The speaker develops her utterance related to the previously produced utterance most of the time using explicit expressions such as “as you said~”, “I agree”, “I disagree”, “I don’t think so”, “I think so” or using the words produced by CGTTs in the preceded utterances.</p> <p>2 –The speaker develops her utterance related to the previously produced utterance one or two times using explicit expressions such as “as you said~”, “I agree”, “I disagree”, “I don’t think so”, “I think so” or using the words produced by CGTTs in the preceded utterances.</p> <p>1 – The speaker develops her utterance related to the previously produced utterance none of the time using explicit expressions such as “as you said~”, “I agree”, “I disagree”, “I don’t think so”, “I think so” or using the words produced by CGTTs in the preceded utterances.</p>	<p>5 – Test taker was able to sustain interaction in responses using confirmation checks, clarification requests, and comprehension checks all of the time (more than 4 examples)</p> <p>3 – Test taker was able to sustain interaction in responses using confirmation checks, clarification requests, and comprehension checks some of the time (1 – 2 examples)</p> <p>1 – Test taker was able to sustain interaction in responses using confirmation checks, clarification requests, and comprehension checks none of the time</p>

Scores of each dimension was decided based on rater training and discussion with raters.

Originally each dimension had one through five score scales; however, pilot rating revealed that the fine distinctions among each score threatened the consistency inter- and intra-rater reliability. Raters did not express any difficulty to follow five score scales for the linguistic dimension; however, interaction dimension was challenging to apply the detailed scoring system. Therefore, the abovementioned scoring system was used and yielded decent inter-rater reliability.

Procedure

This study was conducted over two weeks and consisted of three major steps of data collection. These three phases and the timeline of the data collection procedure are summarized in Table 3.10 and will be explained in detail in the following sections.

Table 3.10 Data Collection Procedure

Procedure	Day	Description	Time
Step 1	Week 1	The screening test and background questionnaire was distributed to the participants.	30 minutes per participant
The screening test Consent Form Background Questionnaire			
Step 1-2		Each group (CRTTs, CGTTLs, CGTTSs, and CGTTHs) was formed, and each test-participant was notified via e-mail of their id number, the location, and time of testing.	
Step 2	Week 2	Day 1: CRTTs 1 through 6 took PA with CGTTH#1 through CGTTH#2, CGTTS#1 through CGTTS#2, and CGTTL#1 through CGTTL#2. Upon completing tasks, test-participants answer the exit questionnaire.	Each PA lasts 5 minutes
PA1 – 3	Day 1 – 3	Day 2: CRTTs # 7 through # 12 took PA with CGTTH#3 through CGTTH#4, CGTTS#3 through CGTTS#4, and CGTTL#3 through CGTTL#4. Upon completing tasks, test-participants answered the exit questionnaire.	15 minutes per set of PA
		Day 3: CRTTs # 13 through # 15 took PA with CGTTH#5, CGTTS#5, and CGTTL#5. Upon completing tasks, test-participants answered the exit questionnaire.	1 hour per day

Step 1 (Week 1)

When test-participants responded, they received an instruction on taking the screening test via e-mail.

Step 1-2 (Week 1)

Based on the information regarding their L2 developmental stages examined in the screening tests, test-participants were grouped into two groups: CRTT group and TI group.

After the grouping was determined, test-participants were informed of their ID numbers, testing time, and testing location. To prevent an awareness of test level results, thereby controlling CRTTs' perceptions of their CGTTs, pseudonyms were used. When the data collection was over, a group identification number was assigned to each test-participant. For instance, CRTT#1 was assigned to CRTTs; a higher language ability CGTT was CGTTH#1; same language ability CGTT was CGTTS#1; finally, CGTTL#1 was a lower CGTT.

Table 3.11 illustrates the sequence of tests and TIs.

Table 3.11 an example of CRTTs' interaction in PA

CRTT	Test 1	Test 2	Test 3
CRTT#1	CGTTH#1	CGTTS #1	CGTTL #1
CRTT#2	CGTTS#1	CGTTL #1	CGTTH #1
CRTT#3	CGTTL#1	CGTTH #1	CGTTS #1

In order to rule out any possible effect, related to the sequence of TIs, the sequence of interaction and testing materials were systematically managed as shown in Table 3.11. CRTT # 1 started the PA 1 with a CGTTH# 1, PA 2 with a CGTTS#1, and, finally, PA 3 with a CGTTL #1. CRTT#2 started the test with a CGTTS. The

advantages of this design could be seen in the efficiency of time management as well as prevention of attrition of test-participants.

Table 3.12 Example Sequence by CGTTs

	CGTTS#1	CGTTS#1	CGTTL #1
CRTT#1	Test 1	Test 2	Test 3
CRTT#2	Test 2	Test 3	Test 1
CRTT#3	Test 3	Test 1	Test 2

Participants in CGTT groups (e.g., CGTTH#1, CGTTS#1, and CGTTL#1) took part in three different tests with three different CRTTs. Each TI took tests 1 through 3 only once so as to exclude any possible practice effect. As noted in Table 3.13, CGTTH#1 worked on test 1 with CRTT #1, test 2 with CRTT #2, and test 3 with CRTT #3.

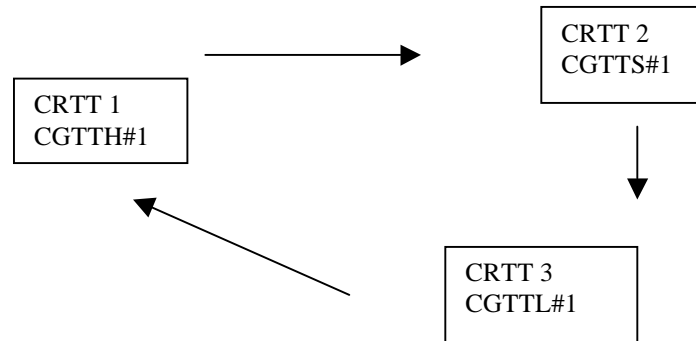
Table 3.13 Example Sequence by Tests

CRTT	Test 1	Test 2	Test 3
CRTT#1	CGTTH#1	CGTTS #1	CGTTL #1
CRTT#2	CGTTS#1	CGTTL #1	CGTTH #1
CRTT#3	CGTTL #1	CGTTH #1	CGTTS #1

Step 2 (Week 2, Day 1 - 3)

During step 2, each PA session has 3 CRTTs, one CGTTH, one CGTTS, and one CGTTL. Figure 3.8 illustrates a possible configuration of testing seating. As the arrows indicate, when a test was over, the CGTT group moved to another desk to work on the next test.

Figure 3.8 Moves of CGTTs during PA



Once CRTTs and CGTTs took their seats, audio-recorded information about the testing procedure and the evaluation criteria was played. In the beginning of the tests, each pair received a set of testing-tasks and written instructions. Each pair of test-participants introduced themselves to each other with the following formulaic introduction script in order to control the amount of information they provide to each other.

Introduce yourself to your partner. Say only the following information. Do NOT mention your real name, age, major, or job.

- I'm _____. (Say the name you were given earlier.)
- I'm learning English here.
- I'm glad to take this test with you.

This formulaic introduction was necessary to control any possible disclosure of information which might affect participants' perception towards their partners. The first sentence informed the researcher of the test-participants' identification without revealing their identities to each other. The second sentence set their social status as ESL students. The final sentence meant to create an amicable atmosphere among test-participants.

Following this introduction stage, each participant had 2 minutes to prepare for a given task (Foster & Skehan, 1999; Wigglesworth, 1997). During this time, they read the task instructions once more and developed strategies for the interaction in the given test, including preparing their responses and reasons for their decisions. Each pair of CRTT

and CGTT then had 3 minutes to work together to complete a task. Each test lasted no more than 10 minutes including introduction, preparation, task-engagement, and switch of their CGTTs. It took approximately 30 minutes for an individual CRTT to complete a set of three PAs. While CRTTs and CGTTs engaged in the test, their interaction was audio and video-recorded.

Rating

Test-raters' evaluation

CRTTs' performance in each test was evaluated by two independent test-raters. In order to avoid, any bias regarding video sequences, numbers were randomly assigned to each PA video clip. In addition, instead of indicating who were CRTTs and CGTTs, the test-raters were given the following grading sheet.

Table 3.14 Test-rater rubric

File #	Test-participants	Linguistic Accuracy					Sociolinguistic Appropriateness			Interaction Strategies		
		1	2	3	4	5	1	2	4	1	3	5
1	Left	1	2	3	4	5	1	2	4	1	3	5
	Right	1	2	3	4	5	1	2	4	1	3	5
2	Left	1	2	3	4	5	1	2	4	1	3	5
	Right	1	2	3	4	5	1	2	4	1	3	5
3	Left	1	2	3	4	5	1	2	4	1	3	5
	Right	1	2	3	4	5	1	2	4	1	3	5

In order to clarify their understanding of the rubric, test-raters and the researcher had offline and online meetings and several e-mail exchanges in which they evaluated 3 sample performances together. Once training was over, they watched video clips independently and evaluated CRTTs' performance using the given rubric. They rated 45 PA instances, both CRTTs' and CGTTs' performance in the first round (i.e., 90 rating

instances total). This approach and the previous version of the rubric which gave more freedom of interpretation to the raters caused the low inter-rater reliability (.6). Based on the test-raters' feedback and observation, the rubric was revised. Another test-rater training was run to understand the new rubric.

The second round of rating, the test-raters were asked to evaluate only CRTTs' performance. The first test-rater evaluated CRTTs' performance first and the rating information was given to the second test-rater. The second test-rater evaluated CRTTs' performance with focusing on her agreement to the first rater's evaluation. Both test-raters wrote notes regarding each CRTT's performance in addition to the rubric.

Performance Data

A native speaker of English, who has several years of teaching ESL and EFL, was hired to transcribe the test-participants' performance. The transcriptions were double-checked by two other people who have experience in non-native speaking data. Information about pauses and intonation was included in transcriptions in order to count utterances accurately.

Two native speakers of North American English were hired to code linguistic accuracy. As the inter-coder agreement was lower than .7, another native speaker of North American English was hired to double-check coding, which yielded higher inter-coder agreement and consistency.

Interaction dimension was coded by two people who have extensive training to deal with L2 learner data. The inter-coder agreement was .95.

Data Analysis

The test-raters' evaluation and transcription data was analyzed statistically. Excel was used to organize the scores and quantified language data, which were then imported into and analyzed using the statistical software package SPSS version 16 on an Acer Aspire laptop with an Intel Core 2 Duo processor running Microsoft Windows 7.

The data included CRTTs' performance rated by two independent test-raters and quantified language data such as the degree of linguistic accuracy, interaction contingency, and interaction strategies.

The test-rater inter-rater reliability was calculated using Spearman's rho (). The following steps were taken in order to calculate the inter-rater reliability. First, each test-rater was numbered (i.e., Rater G: 1 and rater D: 2). Second, their evaluation of CRTTs' performance was inserted in the EXCEL program. Third, the EXCEL file was transferred into the SPSS program. Finally, the bivariate correlation function was selected to calculate the inter-rater reliability. The inter-rater reliability was calculated twice: once the CRTTs' scores were treated as independent observation and the inter-rater reliability was calculated as if there were 45 participants. Second time, the CRTTs' scores were organized based on their CGTTs, and the inter-rater reliability was calculated.

Table 3.15 - 3.17 show the results of the first case.

Table 3.15 Inter-rater reliability

			Rater1_Accu_CGTT	Rater2_Accu_CGTT
Spearman's rho	Rater1_Accu	Correlation Coefficient	1.000	.901**
	_CGTT	Sig. (2-tailed)	.	.000
		N	45	45
	Rater2_Accu	Correlation Coefficient	.901**	1.000
	_CGTT	Sig. (2-tailed)	.000	.

N	45	45
---	----	----

** . Correlation is significant at the 0.01 level (2-tailed).

Table 3.16 Inter-rater reliability

			Rater1_Socio_CGTT	Rater2_Socio_CGTT
Spearman's rho	Rater1_Socio_CGTT	Correlation Coefficient	1.000	.935**
		Sig. (2-tailed)	.	.000
		N	45	45
	Rater2_Socio_CGTT	Correlation Coefficient	.935**	1.000
		Sig. (2-tailed)	.000	.
		N	45	45

** . Correlation is significant at the 0.01 level (2-tailed).

Table 3.17 Inter-rater reliability

			Rater1_Strat_Whole	Rater2_Sstrat_Whole
Spearman's rho	Rater1_Strat_Whole	Correlation Coefficient	1.000	.601**
		Sig. (2-tailed)	.	.000
		N	45	45
	Rater2_Strat_Whole	Correlation Coefficient	.601**	1.000
		Sig. (2-tailed)	.000	.
		N	45	45

** . Correlation is significant at the 0.01 level (2-tailed).

The inter-rater reliability was statistically significant for both the linguistic (i.e., Linguistic accuracy) and interaction dimensions (i.e., Sociolinguistic appropriateness and Interaction strategies).

The inter-rater reliability was calculated for CGTT groups. Table 3.18 – 3.20 show the result of this analysis.

Table 3.18 Inter-rater reliability by group 1

			Rater1_Accu_CGTT	Rater2_Accu_CGTT
Spearman's rho	Rater1_Accu_CGTT	Correlation Coefficient	1.000	.886**
		Sig. (2-tailed)	.	.000
		N	15	15

	Rater2_Accu_CGTTH	Correlation Coefficient	.886**	1.000
		Sig. (2-tailed)	.000	.
			Rater1_Accu_CGTTS	Rater2_Accu_CGTTS
	Rater1_Accu_CGTTS	Correlation Coefficient	1.000	.964**
		Sig. (2-tailed)	.	.000
		N	15	15
	Rater2_Accu_CGTTS	Correlation Coefficient	.964**	1.000
		Sig. (2-tailed)	.000	.
		N	15	15
			Rater1_Accu_CGTTL	Rater2_Accu_CGTTL
	Rater1_Accu_CGTTL	Correlation Coefficient	1.000	.867**
		Sig. (2-tailed)	.	.000
		N	15	15
	Rater2_Accu_CGTTL	Correlation Coefficient	.867**	1.000
		Sig. (2-tailed)	.000	.
		N	15	15

**Correlation is significant at the 0.01 level (2-tailed).

Table 3.19 Inter-rater reliability by group 2

			Rater1_Socio_CGTTH	Rater2_Socio_CGTTH
Spearman's rho	Rater1_Socio_CGTTH	Correlation Coefficient	1.000	.911**
		Sig. (2-tailed)	.	.000
		N	15	15
	Rater2_Socio_CGTTH	Correlation Coefficient	.911**	1.000
		Sig. (2-tailed)	.000	.
		N	15	15
			Rater1_Socio_CGTTS	Rater2_Socio_CGTTS
Spearman's rho	Rater1_Socio_CGTTS	Correlation Coefficient	1.000	1.000**
		Sig. (2-tailed)	.	.
		N	15	15
	Rater2_Socio_CGTTS	Correlation Coefficient	1.000**	1.000
		Sig. (2-tailed)	.	.

		N	15	15
			Rater1_Socio_CGTTL	Rater2_Socio_CGTTL
Spearman's rho	Rater1_Socio_CGTTL	Correlation Coefficient	1.000	.873**
		Sig. (2-tailed)	.	.000
		N	15	15
	Rater2_Socio_CGTTL	Correlation Coefficient	.873**	1.000
		Sig. (2-tailed)	.000	.
		N	15	15

** . Correlation is significant at the 0.01 level (2-tailed).

Table 3.20 Inter-rater reliability by group 3

			Rater1_Stra_CGTTH	Rater2_Stra_CGTTH
Spearman's rho	Rater1_Strat_CGTTH	Correlation Coefficient	1.000	.637**
		Sig. (2-tailed)	.	.011
		N	15	15
	Rater2_Strat_CGTTH	Correlation Coefficient	.637**	1.000
		Sig. (2-tailed)	.011	.
		N	15	15
			Rater1_Stra_CGTTS	Rater2_Stra_CGTTS
Spearman's rho	Rater1_Strat_CGTTS	Correlation Coefficient	.	.
		Sig. (2-tailed)	.	.
		N	15	15
	Rater2_Strat_CGTTS	Correlation Coefficient	.	.
		Sig. (2-tailed)	.	.
		N	15	15
			Rater1_Stra_CGTTL	Rater2_Stra_CGTTL
Spearman's rho	Rater1_Strat_CGTTL	Correlation Coefficient	1.000	.531*
		Sig. (1-tailed)	.	.025
		N	15	15
	Rater2_Strat_CGTTL	Correlation Coefficient	.531*	1.000
		Sig. (1-tailed)	.025	.
		N	15	15

**Correlation is significant at the 0.05 level (2-tailed)

*. Correlation is significant at the 0.05 level (1-tailed).

This strong correlation reveals that the ratings between the two independent test-raters were consistent and reliable. As it was confirmed that the two test-raters' ratings showed

strong correlation, the next level analysis was possible. Table 3.21 summarizes the statistical tests that were used to answer each of the research questions.

Table 3.21 Summary of Statistical Tests by Research Questions

Research Question	Dependent Variable	Data to be compared	Statistical Test
RQ1: Does PA test-takers' use of grammatically accurate L2 utterances vary in relation to the developmentally-equal and unequal status of their pairing?	Performance data in transcriptions	Linguistic dimension data in test 1, 2, & 3	Friedman test
RQ2: Does PA test-takers' use of sociolinguistically appropriate and interactionally strategic L2 utterances vary in relation to the developmentally-equal and unequal status of their pairing?	Performance data in transcriptions	Interaction dimension data in test 1, 2, & 3	Friedman test
RQ3: Does PA test-raters' rating of linguistic accuracy vary in relation to the developmentally-equal and unequal status of their pairing?	Test-raters' evaluation	Linguistic dimension data in test 1, 2, & 3	Friedman test
RQ4: Does PA test-raters' rating of sociolinguistic appropriateness and interaction strategies vary in relation to the developmentally-equal and unequal status of their pairing?	Test-raters' evaluation	Interaction dimension data in test 1, 2, & 3	Friedman test

Effect sizes was calculated using the partial η^2 and ranged from 0 to 1. While 0 refers to no relationship between the repeated measure ANOVA and the dependent variables, 1 means the strongest possible relationship (Green & Salkind, 2005).

Chapter Four will report the results of the analysis by research questions.

CHAPTER FOUR: Results

Introduction

This chapter reports the results of the statistical analysis conducted on test-takers (CRTTs)' performance in pairing with developmentally equal or unequal condition-giving test-takers (CGTTs) (i.e., higher level (CGTTHs), same level (CGTTSs), and lower level (CGTTLs)). The statistical results presented in this chapter are organized according to the four research questions.

Overview of Research questions

Table 4.1 presents the four research questions leading this study.

Table 4.1 Research Questions

Research Question	Dependent Variable	Data to be compared	Statistical Test
RQ1: Does PA test-takers' use of grammatically accurate L2 utterances vary in relation to the developmentally-equal and unequal status of their pairing?	Performance data in transcriptions	Linguistic dimension data in test 1, 2, & 3	Friedman test
RQ2: Does PA test-takers' use of sociolinguistically appropriate and interactionally strategic L2 utterances vary in relation to the developmentally-equal and unequal status of their pairing?	Performance data in transcriptions	Interaction dimension data in test 1, 2, & 3	Friedman test
RQ3: Does PA test-raters' rating of linguistic accuracy vary in relation to the developmentally-equal and unequal status of their pairing?	Test-raters' evaluation	Linguistic dimension data in test 1, 2, & 3	Friedman test
RQ4: Does PA test-raters' rating of sociolinguistic appropriateness and interaction strategies vary in relation to the developmentally-equal and unequal status of their pairing?	Test-raters' evaluation	Interaction dimension data in test 1, 2, & 3	Friedman test

These research questions were developed to better understand paired assessment (PA) as an approach to evaluating L2 learners at process and outcome levels by analyzing the interaction between CRTTs and CGTTs. These research questions also attempted to examine the reliability as well as variability issues in terms of CRTTs' performance in different interaction situations, namely with different CGTTs. In order to answer those issues, this study addressed questions as to whether the L2 samples obtained through PA are valid indicators of linguistic accuracy, sociolinguistic appropriateness, and interaction strategies for SLA across pairs of same and different language developmental stages.

Research questions one and two provided information about CRTTs' performance while ruling out any external judgment or evaluation as transcriptions of CRTTs' performance will be analyzed. Research questions three and four were developed to examine CRTTs' performance from another aspect. That is, raters evaluated CRTTs' performance using a pre-developed rubric, as the evaluating procedures will be similar to those conducted by classroom practitioners.

CRTTs' performance transcriptions were analyzed to answer the research questions one and two. Research questions three and four were answered through the analysis of test-raters' evaluation. The data was analyzed in the following orders: the descriptive analysis of the data was conducted. This analysis provides information about mean, median, and standard deviation of the data. Then, the inferential statistics, repeated measures ANOVA, was conducted. Before repeated measures ANOVA was conducted, an assumption of ANOVA, normality of data, was examined. If the data was

normally distributed, repeated measures ANOVA was conducted. If not, Friedman test, which is the nonparametric equivalence to repeated measures ANOVA, was conducted.

Research Question One: *Does PA CRTTs’ use of grammatically accurate L2 utterances vary in relation to the developmentally-equal and unequal status of their pairing?*

The first research question asked whether CRTTs’ ability to produce grammatically accurate utterances would be influenced by depending on their CGTTs’ L2 developmental stages. This research question was answered through transcription analysis. First, the total number of utterances was counted. Second, utterances which did not have errors in morphemes and syntax chosen were counted. Finally, the percentage of error free utterances was calculated.

The total number of utterances is reported in the following table.

Table 4.2 Total number of utterances

CGTTs	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Min	Max
					Lower Bound	Upper Bound		
wCGTTLs	15	40.3333	11.12	2.87131	34.18	46.49	23.00	62.00
wCGTTSs	15	40.1333	12.82	3.31068	33.03	47.23	17.00	68.00
wCGTTHs	15	42.9333	8.472	2.18755	38.24	47.63	30.00	57.00
Total	45	41.1333	10.79	1.60668	37.90	44.37	17.00	68.00

The mean number of utterances that CRTT produced while they were engaged with CGTTLs was 40.33, with CGTTSs was 40.13, and with CGTTHs was 42.93. That is, CRTTs produced more utterances while they interacted with CGTTHs compared to CGTTLs or CGTTSs.

Prior to the analysis of the accuracy level, the total number of utterances was examined for its normality. The test shows that the total number of utterances was normally distributed.

Table 4.3 Test for Normality – Total number of utterances

	CGTTs	Kolmogorov-Smirnov ^a			Shapiro-Wilk		
		Statistic	Df	Sig.	Statistic	df	Sig.
Total Utterances	1	.139	15	.200*	.960	15	.701
	2	.130	15	.200*	.975	15	.922
	3	.186	15	.170	.936	15	.334

a. Lilliefors Significance Correction

*. This is a lower bound of the true significance.

As explained earlier, if the p-value of the Shapiro-Wilk test does not indicate that it is not statistically significant, it is assumed that the data is normally distributed ($p > .05$). Since an assumption for ANOVA was met, one-way ANOVA was run. In order to find out whether the mean differences are statistically dissimilar, ANOVA was conducted (refer to table 4.4).

Table 4.4 One-way ANOVA of the total number of utterances

Total Utterances	Sum of Squares	Df	Mean Square	F	Sig.
Between Groups	73.200	2	36.6	.305	.739
Within Groups	5038.000	42	119.95		
Total	5111.200	44			

This test yielded $F(2, 42) = .31$ and $p > .05$. That is, the mean differences among the total number of utterances produced with CGTTLs, CGTTSs, or CGTTHs are not statistically different. This result is contrary to those in the previous research result reported by Iwashita (1999) and Davis (2009), who found the amount of talk increased as test-takers were paired with high language ability test-takers.

In the next step, the mean percentage of grammatically accurate utterances (i.e., the mean percentage of error-free utterances) was calculated. As shown in Table 4.5, the

mean percentage of error-free utterances produced while CRTTs interacted with CGTTSs ($M = 86.34$) was higher than that with CGTTLs ($M = 80.83$) or CGTTHs ($M = 84.88$).

Table 4.5 Descriptive Statistics of the percentage of error-free utterances

	Mean	Std. Deviation	N
PerErrFreeUtterance_wCGTTLs	80.83	15.96210	15
PerErrFreeUtterance_wCGTTSs	86.34	5.75177	15
PerErrFreeUtterance_wCGTTHs	84.88	9.05163	15

Compared to the large standard deviation of the percentage of CRTTs' ability to produce grammatically accurate utterances while they were engaged in interaction with CGTTLs or CGTTHs ($SD = 15.96$ and $SD = 9.05$ respectively), the small standard deviation of the error free utterances percentage of CGTTSs is also noticeable ($SD = 5.75$). That is, there were less individual differences in the percentage of grammatically accurate utterances while CRTTs interacted with CGTTSs than CGTTLs or CGTTHs.

Once this data was transferred to SPSS for further analysis, a Friedman test, a non-parametric equivalent analysis, was also conducted to evaluate differences in the mean percentage of grammatically accurate utterances. As shown in Table 4.6, the test result was not significant, $\chi^2(2, N=15) = 1.67, p < .5$, and the Kendall coefficient of concordance of .055 indicated almost no differences among the three mean percentage of grammatically accurate utterances produced by CRTTs.

Table 4.6 Friedman test results

	Mean Rank
PerErrFreeUtter_wCGTTLs	2.07
PerErrFreeUtter_wCGTTSs	2.00
PerErrFreeUtter_wCGTTHs	1.93

N	15
Kendall's W ^a	.055
Chi-Square	1.655
Df	2
Asymp. Sig.	.437

a. Kendall's Coefficient of Concordance

Research Question Two: *Does PA CRTTs’ use of sociolinguistically appropriate and interactionally strategic L2 utterances vary in relation to the developmentally-equal and unequal status of their pairing?*

The second research question examined CRTTs’ ability to produce sociolinguistically appropriate and strategically useful interaction. This research question was also answered through transcription analysis. In this study, sociolinguistically appropriate utterances were operationalized as utterances which contain the same words or explicit expressions that link to the previously produced utterances by their CGTTs. Those expressions include “as you said~”, “I agree”, “I disagree”, “I don’t think so”, and “I think so”. Strategically useful utterances were operationalized as three ways of eliciting their CGTTs’ re-utterances. Those strategies were confirmation checks, comprehension checks, and clarification requests. The data used to answer the second research question was analyzed through transcriptions.

Utterances which contained the abovementioned information were counted. Then the percentage of sociolinguistically appropriate and strategically useful utterances was calculated. As the first step, mean percentage information is reported in the following table.

Table 4.7 Descriptive statistics

	N	Mean	Std. Deviation	Minimum	Maximum
SociolxApproUtter_wCGTTLs	15	10.74	10.23	.00	38.46
SociolxApproUtter_wCGTTSs	15	7.76	8.11	.00	22.45
SociolxApproEtter_wCGTTHs	15	10.33	7.40	.00	24.32
InterStrat_wCGTTLs	15	1.89	3.72	.00	11.54
InterStrat_wCGTTSs	15	1.13	3.48	.00	13.51
InterStrat_wCGTTHs	15	1.039	2.55	.00	9.26

While CRTTs were engaged in interaction with CGTTLs ($M = 10.74\%$) and CGTTHs ($M=10.33\%$), they used more explicit expressions such “as you said~”, “I agree”, “I disagree”, “I don’t think so”, and “I think so” than with CGTTSs ($M = 7.76\%$). The mean percentage of utterances including clarification request, confirmation checks, and comprehension checks was 1.89% with CGTTLs, 1.13% with CGTTSs, and 1.04% with CGTTHs.

As the data was percentage data, nonparametric analysis was run to examine the mean percentage differences among CRTTs’ performance. A Friedman test was conducted to evaluate differences in the mean percentage of sociolinguistically appropriate utterances.

Table 4.8 Friedman test

	Mean Rank
SociolxApproUtter_wCGTTLs	2.13
SociolxApproUtter_wCGTTSs	1.73
SociolxApproUtter_wCGTTHs	2.13

N	15
Kendall's W ^a	.055
Chi-Square	1.655
Df	2
Asymp. Sig.	.437

a. Kendall's Coefficient of Concordance

As shown in Table 4.8, the test result was not significant, $\chi^2(2, N=15) = 1.66, p < .5$, and the Kendall coefficient of concordance of .06 indicated that the differences among the three mean percentage of sociolinguistically appropriate utterances produced by CRTTs was minimal.

A Friedman test for interaction strategies is reported in the following section.

Table 4.9 Friedman test

	Mean Rank	N	15
InterStrat_wCGTTLs	2.07	Kendall's W ^a	.035
InterStrat_wCGTTSs	1.87	Chi-Square	1.043
InterStrat_wCGTTHs	2.07	Df	2
		Asymp. Sig.	.593

As shown in Table 4.9, the test result was not significant, $\chi^2(2, N=15) = 1.04, p > .05$, and the Kendall coefficient of concordance of .04 indicated almost no differences among the three mean percentage of strategically useful utterances produced by CRTTs.

Research Question Three: *Do PA CRTTs' scores in linguistic accuracy vary in relation to the developmentally-equal and unequal status of their pairing?*

The third research questions asked whether test-raters' evaluation of CRTTs' performance of linguistic accuracy varied depending on CGTTs' status. This research question was answered through statistical analysis of test-raters' evaluation.

The descriptive statistics of CRTTs' linguistic accuracy showed that CRTTs' linguistic accuracy ratings slightly increased as they interacted with higher level CGTTs (refer to Table 4.13). For instance, the mean rating which CRTTs got while interacting with CGTTLs was lowest ($M = 4.27$) and that with CGTTHs was highest ($M = 4.37$).

Table 4.10 Descriptive Statistics

	Mean	Std. Deviation	N
Rating_wCGTTLs	4.267	.7761	15
Rating_wCGTTSs	4.300	.8619	15
Rating_wCGTTHs	4.367	.6935	15

A Friedman test was conducted to examine the differences in medians among the ratings of CRTTs' performance with a focus on grammatical accuracy.

Table 4.11 Friedman test

	Mean Rank
Rating_wCGTTLs	2.03
Rating_wCGTTSs	1.93
Rating_wCGTTHs	2.03

N	15
Kendall's W ^a	.004
Chi-Square	.130
Df	2
Asymp. Sig.	.937

a. Kendall's Coefficient of Concordance

As shown in Table 4.11, the test result was not significant, $\chi^2(2, N=15) = .13, p > .5$, and the Kendall coefficient of concordance of .004 indicated almost no differences among the three mean percentage of grammatically accurate utterances produced by CRTTs.

Research Question Four: Do PA CRTTs' scores in sociolinguistic appropriateness and interaction strategies vary in relation to the developmentally-equal and unequal status of their pairing?

Research question four examined the test-raters' evaluation of CRTTs' performance with respect to CRTTs' ability to interact in ways that are sociolinguistically appropriate and strategically useful for SLA. The evaluation of CRTTs' interaction ability was focused on sociolinguistic appropriateness and interaction strategies. This research question was answered through the repeated measures ANOVA of test-raters' evaluation. Table 4.12 shows the information about the descriptive statistics of CRTTs' sociolinguistic appropriateness. It showed that CRTTs produced more cohesive utterances while they were interacting with CGTTSs (i.e., $M = 2.57$). They were least cohesive while they were engaged in the interaction with CGTTHs ($M = 2.27$).

Table 4.12 Descriptive Statistics of Sociolinguistic appropriateness

	Mean	Std. Deviation	N
Rating_wCGTTLs	2.47	1.1255	15
Rating_wCGTTSs	2.57	1.1782	15
Rating_wCGTTHs	2.27	1.2373	15

Before conducting repeated measures ANOVA, the data was examined whether it is normally distributed.

Table 4.13 Tests of Normality

	CGTTLs1_ CGTTSs2_ wCGTTHs3	Kolmogorov-Smirnov ^a			Shapiro-Wilk		
		Statistic	df	Sig.	Statistic	Df	Sig.
Ave_Sociolinguistics	1	.202	15	.101	.902	15	.101
	2	.269	15	.005	.776	15	.002
	3	.220	15	.050	.783	15	.002

a. Lilliefors Significance Correction

The results require both parametric and non-parametric analyses. Tables that follow show repeated measures ANOVA results.

Table 4.14 Multivariate Tests^b

Effect	Value	F	Hypothesis df	Error df	Sig.	Partial Eta Squared	
CGTTs	Pillai's Trace	.042	.282 ^a	2.000	13.000	.759	.042
	Wilks' Lambda	.958	.282 ^a	2.000	13.000	.759	.042
	Hotelling's Trace	.043	.282 ^a	2.000	13.000	.759	.042
	Roy's Largest Root	.043	.282 ^a	2.000	13.000	.759	.042

a. Exact statistic

b. Design: Intercept
Within Subjects Design: CGTTs

The results of repeated measures ANOVA did not indicate any significant CGTT influence on CRTTs' performance. Moreover, the effect size was minimal. $F(2, 28) = .36$, $Wilks' \lambda = .96$, $p > .5$, and partial $\eta^2 = .042$.

Table 4.15 Friedman test

	Mean Rank		N	15
PerErrFreeUtterance_ wCGTTLS	1.97		Kendall's W ^a	.002
PerErrFreeUtterance_ wCGTTSs	2.03		Chi-Square	.054
PerErrFreeUtterance_ wCGTTHs	2.00		Df	2
			Asymp. Sig.	.973

a. Kendall's Coefficient of Concordance

As shown in Table 4.15, the test result was not significant, $\chi^2(2, N=15) = .05, p > .05$, and the Kendall coefficient of concordance of .002 indicated almost no differences among the three mean percentage of ability to interact in ways that are sociolinguistically accurate utterances produced by CRTTs.

The following section will report the results of CRTTs' ability to interact in ways that are strategically useful for interaction. The descriptive statistics of test-raters' evaluation showed that CRTTs used more interaction strategies while they were engaged in the interaction with CGTTHs. They used least interaction strategies with CGTTSs.

Table 4.16 Descriptive Statistics of Interaction Strategies

	Mean	Std. Deviation	Median	N
Rating_ wCGTTLS	1.4	.63	1.00	15
Rating_ wCGTTSs	1.07	.26	1.00	15
Rating_ wCGTTHs	1.87	1.13	2.00	15

Then the normality of the data was examined.

Table 4.17 Tests of Normality^b

		Kolmogorov-Smirnov ^a			Shapiro-Wilk		
CGTTLS1_CGTTHs3		Statistic	df	Sig.	Statistic	df	Sig.
Ave_Strategies	1	.416	15	.000	.705	15	.000
	3	.370	15	.000	.617	15	.000

a. Lilliefors Significance Correction

b. Ave_Strategies is constant when NETTL1_NETTS2_NETTH3 = 2. It has been omitted.

It showed that this data is not normally distributed, which only allowed non-parametrical analysis. Friedman’s test as well as Kendall’s test were conducted (refer to Table 4.22).

Table 4.18 Friedman test

	Mean Rank		15
Rating_wCGTTLs	2.03	Kendall's W ^a	.231
Rating_wCGTTSs	1.63	Chi-Square	6.938
Rating_wCGTTHs	2.33	Df	2
		Asymp. Sig.	.031

a. Kendall's Coefficient of Concordance

As shown in Table 4.18, the test result was not significant, $\chi^2(2, N=15) = 6.94, p < .05$, and the Kendall coefficient of concordance of .23 indicated that there was statistical significance in terms of differences among the three mean percentage of strategically useful utterances produced by CRTTs. As the result was statistically significant, post-hoc analysis was performed. However, as SPSS does not offer a way to run the post-hoc analysis of Friedman’s test, Wilcoxon’s test (i.e., two independent sampled test) was performed for each pair (i.e., rating with CGTTL vs. rating with CGTTS, rating with CGTTLs vs. rating with CGTTHs, and rating with CGTTSs vs. rating with CGTTHs).

Table 4.19 reports the rank analysis between CRTTs’ performance with CGTTSs and that with CGTTLs.

Table 4.19 Wilcoxon test

		N	Mean Rank	Sum of Ranks		Rating_wCGTTSs - Rating_wCGTTLs
Rating_wCGTTSs	Negative Ranks	4 ^a	2.50	10.00	Z	-1.890 ^a
Rating_wCGTTLs	Positive Ranks	0 ^b	.00	.00	Asymp. Sig. (2-tailed)	.059
	Ties	11 ^c				
	Total	15				

a. Rating_wCGTTSs < Rating_wCGTTLs

b. Rating_wCGTTSs > Rating_wCGTTLs

c. Rating_wCGTTSs = Rating_wCGTTLs

A Wilcoxon Signed-ranks test indicated that the rating which CRTTs got while they were engaged in interaction with CGTTLs ($Mdn = 1.00$) was higher than with CGTTSs ($Mdn = 1.00$), $Z = 1.89$, $p > .05$, $r = .49$

The following tables show the result of the test with CGTTHs and CGTTSs data.

Table 4.20 Wilcoxon test

		N	Mean Rank	Sum of Ranks
Rating_wCGTTHs	Negative Ranks	1 ^a	3.50	3.50
Rating_wCGTTSs	Positive Ranks	8 ^b	5.19	41.50
Ties		6 ^c		
Total		15		

- a. Rating_wCGTTHs < Rating_wCGTTSs
- b. Rating_wCGTTHs > Rating_wCGTTSs
- c. Rating_wCGTTHs = Rating_wCGTTSs

	Rating_wCGTTHs - Rating_wCGTTSs
Z	-2.326 ^a
Asymp. Sig. (2-tailed)	.020

- a. Based on negative ranks.
- b. Wilcoxon Signed Ranks Test

A Wilcoxon Signed-ranks test indicated that the rating which CRTTs got while they were engaged in interaction with CGTTHs ($Mdn = 2.00$) was higher than with CGTTSs ($Mdn = 1.00$), $Z = 2.36$, $p < .05$, $r = .6$

Table 4.21 Wilcoxon test

		N	Mean Rank	Sum of Ranks
Rating_wCGTTHs	Negative Ranks	3 ^a	4.00	12.00
Rating_wCGTTLs	Positive Ranks	6 ^b	5.50	33.00
Ties		6 ^c		
Total		15		

- a. Rating_wCGTTHs < Rating_wCGTTLs
- b. Rating_wCGTTHs > Rating_wCGTTLs
- c. Rating_wCGTTHs = Rating_wCGTTLs

	Rating_wCGTTHs - Rating_wCGTTLs
Z	-1.310 ^a
Asymp. Sig. (2-tailed)	.190

- a. Based on negative ranks.
- b. Wilcoxon Signed Ranks Test

A Wilcoxon Signed-ranks test indicated that the rating which CRTTs got while they were engaged in interaction with CGTTHs ($Mdn = 1.00$) was higher than with CGTTLs ($Mdn = 1.00$), $Z = 1.31$, $p > .05$, $r = .34$. The results showed that the differences in test-raters' evaluation of CRTTs' performance are statistically significant comparing that with CGTTHs and CGTTSs. That is, the rating differences between the interaction with CGTTLs and that with CGTTSs or that with CGTTHs and CGTTLs are not statistically significant.

Exit Survey

The following section reports the results of the exit survey that CRTTs and CGTTs completed in after each PA. This section will only report CRTTs responses. The exit survey questions asked about CRTTs' evaluation of their own performance and their CGTTs' performance. Although the information from this exit survey was not systematically incorporated in the research question of this study, the results are still reported to provide CRTTs' perception after they were done with each PA. The questions used the 1 through 5 Likert scale.

The first question asked their evaluation of the difficulty of each PA. CRTTs expressed that they felt the PA with CGTTLs was slightly more difficult than the ones with CGTTSs or CGTTHs. However, the differences were minimal.

Table 4.22 Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
TestDffwCGTTLs	15	1.00	4.00	2.6000	1.05560
TestDiffwCGTTSs	15	1.00	4.00	2.5333	.91548
TestDffwCGTTHs	15	1.00	4.00	2.4000	1.05560
Valid N (listwise)	15				

The next question is about their evaluation of their performance.

Table 4.23 Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
OwnLevelwCGTTLS	15	1.00	4.00	2.9333	.79881
OwnLevelwCGTTSs	14	2.00	4.00	3.0000	.55470
OwnLevelwCGTTHs	15	2.00	5.00	3.0000	.84515
Valid N (listwise)	14				

CRTTs evaluated their level as middle level ($M=3.0$). Their evaluations to the minimum and maximum level of their performance were slightly lower when they were engaged in CGTTLS (i.e., 1.00 & 4.00) than CGTTSs (i.e., 2.00 & 4.00) or CGTTHs (i.e., 2.00 & 5.00).

The following question was about CRTTs' evaluation of their CGTTs' performance during each PA.

Table 4.24 Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
CGTTLperformance	15	1.00	4.00	2.93	.80
CGTTSperformance	15	2.00	5.00	3.27	.80
CGTTHperformance	15	2.00	5.00	3.40	.99
Valid N (listwise)	15				

While CRTTs evaluated CGTTHs' level highest ($M = 3.4$), they did CGTTLS' lowest ($M= 2.93$). As the mean differences were larger than those from other questions, a one-way ANOVA was run. The normality test of this data did not confirm all the data was normally distributed, both a one-way ANOVA and a non-parametric test were run.

Table 4.25 Tests of Normality

	CGTT	Kolmogorov-Smirnov ^a			Shapiro-Wilk		
		Statistic	df	Sig.	Statistic	df	Sig.
CGTTLevel	1	.333	15	.000	.819	15	.006
	2	.297	15	.001	.865	15	.028
	3	.195	15	.128	.896	15	.082

a. Lilliefors Significance Correction

Table 4.26 show the result from a one-way ANOVA.

Table 4.26 ANOVA

CGTTLevel

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	1.733	2	.867	1.157	.324
Within Groups	31.467	42	.749		
Total	33.200	44			

The result of a one-way ANOVA did not indicate any statistical significance ($p > .5$) and

$F(2, 42) = 1.16$. A Friedman test was also run to examine the result.

Table 4.27 Ranks

	Mean Rank
CGTTLevelCGTTL	1.80
CGTTLevelCGTTS	2.10
CGTTLevelCGTTH	2.10

Table 4.28 Test Statistics

N	15
Kendall's W ^a	.051
Chi-Square	1.543
Df	2
Asymp. Sig.	.462

a. Kendall's Coefficient of Concordance

The test result was not significant, $\chi^2(2, N=15) = 1.54, p < .5$, and the Kendall coefficient of concordance of .05 indicated that there was no statistical significance.

The results of this study are summarized in Table 4.29.

Table 4.29 Summary of research questions and results

Research Question	Statistical Test	Results
RQ1: Does PA test-takers' use of grammatically accurate L2 utterances vary in relation to the developmentally-equal and unequal status of their pairing?	Friedman test	$\chi^2 (2, N=15) = 1.67, p < .5$ No. CRTTs' use of grammatically accurate L2 utterances did not vary in relation to their CGTTs. CRTTs consistently produced grammatically accurate or inaccurate utterances regardless of their CGTTs' L2 developmental stages.
RQ2: Does PA test-takers' use of sociolinguistically appropriate and interactionally strategic L2 utterances vary in relation to the developmentally-equal and unequal status of their pairing?	Friedman test	$\chi^2 (2, N=15) = 1.66, p > .5$, Kendall coefficient of concordance = .06 $\chi^2 (2, N=15) = 1.04, p > .5$, Kendall's $W = .04$ No. CRTTs' use of sociolinguistically appropriate and interactionally strategic utterances was consistent regardless of their CGTTs' L2 developmental stages.
RQ3: Does PA test-raters' rating of linguistic accuracy vary in relation to the developmentally-equal and unequal status of their pairing?	Friedman test	$\chi^2 (2, N=15) = .13, p > .5$, Kendall coefficient of concordance = .004 No. Test-raters' evaluation of CRTTs' ability to produce linguistically accurate utterances was consistent regardless of CGTTs' L2 developmental stages.
RQ4: Does PA test-raters' rating of sociolinguistic appropriateness and interaction strategies vary in relation to the developmentally-equal and unequal status of their pairing?	Friedman test	$F (2, 28) = .36$ Wilks' $\lambda = .96, p > .5$, partial $\eta^2 = .042$. $\chi^2 (2, N=15) = .05, p > .5$ Kendall's $W = .002$ $\chi^2 (2, N=15) = 6.94, p < .05$, Kendall's $W = .23$ No. Test-raters' evaluation of CRTTs' ability to interact in ways that were sociolinguistically appropriate did not vary. Yes. Test-raters' evaluation of CRTTs' ability to interact in ways that were strategically useful for SLA varied in relation to CGTTs' L2 developmental stages. In particular, the score differences between CRTTs' interaction with CGTTHs and CGTTSs were statistically significant.

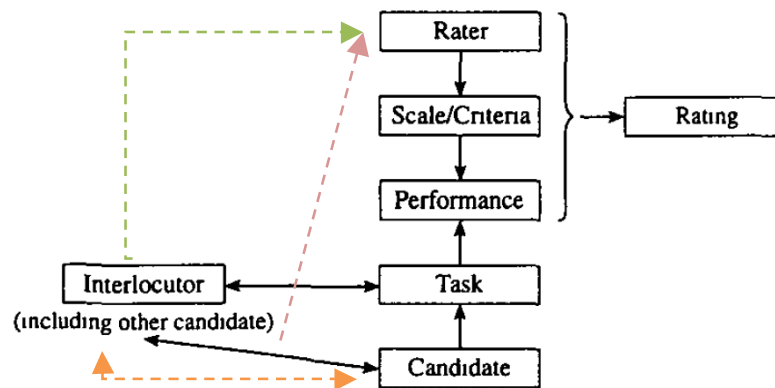
The following chapter will discuss the results and conclude this study.

CHAPTER FIVE: Discussion and Conclusion

Introduction

The goal of the study was to better understand paired assessment (PA) as an approach to evaluating L2 learners by analyzing the interaction between two non-native speaking test-takers. In particular, the impetus of the study was the paucity of research studies on the perspectives on two paired test takers' influence on each others' performance. Previously, the focus of language assessment had been largely on individual test-takers' performance and their cognitive processes they revealed during an exam than on a test-taker-in-interaction (Chalhoub-Deville, 2003; McNamara, 1997). The study was developed in order to add to the lack of research on another test-taker's influence on test-takers during PA. Criticizing the overemphasis on individuals' cognitive processes during language assessment, McNamara (1997) argued that interaction between test-takers should be a target of rating as well. In contrast to Bachman who considered the social variation as an undesirable and noisy factor in language assessment, McNamara regarded it as an ignored construct in assessment. As shown below, the framework advanced by McNamara called for greater attention to a test-taker-in-interaction with another person in a testing setting.

Figure 5.1 Language assessment framework (Candidate: test-taker)



(Adopted from McNamara 1996:86)

As discussed earlier, this framework provided a meaningful way to draw test-developers' and test-users' attention to the existence of interlocutors⁵ in testing settings and their influence on other test-takers. This is especially the case with respect to their importance in terms of surfacing test-takers' cognitive processes and interaction ability. It is believed that the existence of another test-taker as an interlocutor, in which case there are two test-takers, can help teachers to evaluate students efficiently in a large classroom setting.

Nonetheless, there are claims that the existence of interlocutors can weaken the reliability and fairness of a language assessment. In particular, the invisible influence and interaction among test-takers, interlocutors (i.e., another test-taker or a tester), and test-raters in the processes of language assessment and evaluation as indicated in the dotted line in Figure 5.1.needed to be examined. The study addressed the concerns regarding the influence that each test-taker will give and receive and the possible variation that this might reveal in their performance.

⁵ McNamara uses this term if there is any other person who interacts with a test-taker in a test. Interlocutor may include either a tester or another test-taker.

The study examined CRTTs-in-interaction by focusing on their linguistic and interaction ability variation in relation to their CGTTs' L2 developmental stages. In the study, thirty second language (L2) learners⁶ interacted as status-equal test-takers to shed light on their ability to produce linguistically accurate utterances and interact in ways that were sociolinguistically appropriate and strategically useful for SLA. Their performance was examined twofold: their utterances were transcribed and analyzed for the target features, and two test-raters were hired to evaluate their performance. The following four research questions were developed to examine test-takers' performance systematically:

1. Does PA test-takers' use of grammatically accurate L2 utterances vary in relation to the developmentally-equal and unequal status of their pairing?
2. Does PA test-takers' use of sociolinguistically appropriate and interactionally strategic L2 utterances vary in relation to the developmentally-equal and unequal status of their pairing?
3. Does PA test-raters' rating of linguistic accuracy vary in relation to the developmentally-equal and unequal status of their pairing?
4. Does PA test-raters' rating of sociolinguistic appropriateness and interaction strategies vary in relation to the developmentally-equal and unequal status of their pairing?

The study examined the other test-takers' (i.e., CGTTs) influence, with a special focus on CGTTs' L2 developmental stage differences, on test-takers' (i.e., CRTT) performance.

While controlling other compounding variables of CGTTs in its research method, the study ruled out the potential influence of other variables and only measured the influence of the L2 developmental stage differences on CRTTs' performance. The findings of the study revealed that CRTTs' performance did not vary when they were paired with

⁶ Fifteen were evaluated test-takers (CRTTs) and the other fifteen were non-evaluated test-takers (CGTTs) in the study.

CGTTs whose L2 developmental stages differed. However, test-raters' evaluation of CRTTs' performance was found to vary in important ways. A summary of the findings is listed below:

1. Test-takers' ability to produce linguistically accurate utterances did not vary in relation to other test-takers' L2 developmental stages during PAs.
2. Test-takers' ability to produce utterances in ways that were sociolinguistically appropriate and strategically useful did not vary in relation to other test-takers' L2 developmental stages during PAs.
3. Test-raters' evaluation of test-takers' ability to produce linguistically accurate utterances did not vary in relation to other test-takers' L2 developmental stages during PAs.
4. Test-raters' evaluation of test-takers' ability to produce utterances in ways that were sociolinguistically appropriate did not vary in relation to other test-takers' L2 developmental stages during PAs. However, variation was shown in the test-raters' evaluation of test-takers' ability to interact in ways that are strategically useful for SLA. The variation was noted despite test-takers' consistent performances as revealed through the transcription analysis.

This chapter will discuss these findings based on the target dimensions. The first section will discuss the results of research questions one and three which were about CRTTs' linguistic accuracy and its rating. The second section will discuss the findings of research questions two and four. As such, these focused on CRTTs' ability to interact in ways that were sociolinguistically appropriate and strategically useful for SLA. Finally, it will conclude with suggestions and implications for pedagogy and future research.

Findings regarding test-takers' linguistic accuracy

Research questions one and three examined whether CRTTs' ability to produce linguistically accurate utterances would vary in relation to their CGTTs' L2 developmental stages. Research question one examined the transcribed utterances

produced by CRTTs to check whether their linguistic accuracy varied in relation to their CGTTs' L2 developmental stages. Research question three examined test-raters' evaluation of CRTTs' performance with a focus on their linguistic accuracy. Analysis of utterances and rating revealed that CRTTs' ability to produce linguistically accurate utterances was demonstrated and evaluated consistently regardless of their CGTTs' L2 developmental stages. Descriptive analysis showed CRTTs produced linguistically more accurate utterances while they interacted with CGTTSs than CGTTLs or CGTTHs. However, inferential statistics did not confirm that these differences were statistically significant. In the following sections, each research question will be discussed in detail.

Discussion of Results for Research Question 1

The results found in analysis of CRTTs' utterances were drawn from both descriptive and inferential statistics. Firstly, the results from the descriptive statistics revealed that the degree of CRTTs' linguistically accurate utterances varied. The raw mean percentages of error-free utterances were different in relation to CGTTs' L2 developmental stages. CRTTs performed better with CGTTSs ($M = 86.34\%$) than with CGTTLs ($M = 80.83\%$) or CGTTHs ($M = 84.88\%$). However, secondly inferential statistics yielded from repeated measures ANOVA did not support the claim that these differences were statistically significant ($p > .5$ $F(3, 13) = .84$). That is, although there existed differences in descriptive statistics, the differences were not distinctive enough to claim that CRTTs performed differently.

These results were partially consistent with those in Iwashita (1999); however, the interpretation was different. Her descriptive analysis of the data showed differences in

CRTTs' performance in terms of grammar and expression uses. This data analysis allowed her to argue that her test-takers produced more grammatically accurate language while they were paired with same-level test-takers than they were with high-level test-takers. If they were low-level test-takers, they produced linguistically more accurate utterances when they were paired with higher-level test-takers. However, as she did not run any inferential statistics, it is not clear whether the differences her data revealed were statistically significant.

These findings are also consistent with those revealed in other SLA studies. First, as studies of L2 acquisition showed, the degree of linguistic accuracy did not vary by CGTTs' production, which confirms that learners' uptake of the other learners' erroneous production does not usually happen (e.g, Gass & Selinker, 2003). Second, CRTTs' consistent performance of linguistic accuracy may have something to do with the grammatical features targeted in the linguistic dimension. Although global accuracy was examined, the following morphological and syntactical features were chosen in order to ensure high inter-rater reliability and inter-coder reliability (Iwashita, et al., 2008). The target features in the study included 1) morphological features such as verb tense, third person singular, and plural markers and 2) syntactical features such as prepositions, article use, and word order. These individual features were examined in CRTTs' transcriptions. More research should be conducted; however, results of this analysis suggested that these features were development-bound rather than interaction-bound. That is, performance of these morphological and syntactical features was more influenced by CRTTs' L2 development. Their performance would vary when their L2 acquisition of

these forms were not completed. In other words, performance of these features was intrapersonal and thus did not vary in relation to other test-takers. Instead, their performance of these forms was cognitively constructed rather than constructed through interaction. That is, these features were intrapersonal and cognitive constructs rather than interpersonal and interaction ones. Hence, CRTTs' ability to produce linguistically accurate utterances of these features would not vary despite the changes of external factors, in this case CGTTs' L2 developmental stages.

Moreover, the findings may suggest that as these features can be selected for linguistic accuracy evaluation in PA, the concern about the influence of the other test-takers to a test-taker should be reconsidered. The foremost concern raised regarding PA is that test-takers, who share the same testing status and are evaluated with another test-taker in PA, can cause variability in their own performance by influencing each other (e.g., Foot, 1998). They argued that even testers who are trained to interact with test-takers can elicit inconsistent performance of test-takers as their interaction behavior can fluctuate, and having another test-taker in a testing setting would only cause unreliable testing results. However, the results of research question one revealed that CRTTs performed consistently regardless of their CGTTs' L2 developmental stages.

Discussion of Results for Research Question 3

The test-raters' mean rating of CRTTs' ability to produce linguistically accurate utterances revealed large differences neither in descriptive nor inferential statistics. Although the mean rating that CRTTs received while interacting with respect to their interaction with CGTTLs was lowest ($M = 4.27$) and that with CGTTHs was highest (M

= 4.37), these differences were negligible. The Friedman test confirmed that there were no differences among the three mean percentage scores of grammatically accurate utterances produced by CRTTs ($\chi^2(2, N=15) = .13, p > .5$, Kendall's $W = .004$).

The results found in Iwashita (1999) did not support this finding. Her findings showed differences in the raw mean rating scores; her test-takers received higher scores when high-level test-takers interacted with the same-level test-takers and low-level test-takers interacted with higher-level test-takers. However, again as she did not conduct an inferential statistical analysis, it is not straightforward to conclude that the differences test-raters' evaluation was statistically meaningful. On the other hand, studies conducted by Davis (2009) and Csepes (2002) supported the findings of the study. When conducting Rasch analysis and *Chi-square* analysis, they did not find any statistically significant differences in test-raters' evaluation.

These findings provided a supportive rationale for implementing PA as part of a testing battery. Concerns regarding test-raters' inconsistent evaluation in relation to test-takers' pairing are a leading factor to create hesitancy of employing PA in high-stakes testing as well as classroom assessment. However, the test-raters' consistent evaluation of CRTTs' performance, which was revealed in the statistical analysis and the comparison with CRTTs' performance analyzed in transcriptions, provided encouragement to include linguistic accuracy with a focus on CRTTs' morphological and syntactical features as a testing construct in PA.

Findings regarding test-takers' interaction ability

Research questions two and four examined whether CRTTs' ability to interact in ways that were sociolinguistically appropriate and strategically useful for SLA would vary in relation to their CGTTs' L2 developmental stages. Research question two examined the transcribed utterances produced by CRTTs to check whether their interaction ability varied in relation to their CGTTs' L2 developmental stages. Research question four examined test-raters' evaluation of CRTTs' performance with a focus on their interaction. Analysis of utterances revealed that CRTTs' ability to interact in ways that were sociolinguistically appropriate and strategically useful for SLA was demonstrated consistently regardless of their CGTTs' L2 developmental stages. Test-raters' evaluation of CRTTs' ability to interact in ways that were sociolinguistically appropriate was consistent regardless of CGTTs' L2 developmental stages; however, the evaluation of their ability to interact in ways that were strategically useful for SLA varied.

In the following sections, each research question will be discussed in detail.

Discussion of Results for Research Question 2

Research question two examined CRTTs' interaction ability in ways that were sociolinguistically appropriate and strategically useful for SLA through transcription analysis. The analyses revealed that CRTTs' performance did not vary regardless of their CGTTs' differences. The mean differences of producing sociolinguistically appropriate utterances, in CRTTs' utterances with CGTTLs (10.75%), CGTTSs (7.76%), and CGTTHs (10.33%) were negligible. The inferential statistics confirmed that these differences were not significant (Kendall's $W = .05$). Moreover, the degree of using

interaction strategies such as confirmation checks, clarification requests, and comprehension checks were not found to vary as well ($wCGTTLs = 1.89\%$, $wCGTTSs = 1.13\%$, and $wCGTTHs = 1.04\%$). Inferential statistics also confirmed that the mean differences were not statistically significant (Kendall's $W = 0.04$).

As these two domains of interaction ability were not specifically examined in the previous studies, the comparison between the results of the study and those in the previous studies was not straightforward. However, attempts were made to compare the interaction ability examined in the study with that of other studies. First, CRTTs' ability to interact in ways that were sociolinguistically appropriate was compared to the findings of Nakatsuhara (2006). Her findings regarding the goal-orientation and interaction contingency revealed in their utterances confirmed that there were negligible differences in relation to other test-takers' L2 levels.

Second, an attempt was made to situate the findings regarding CRTTs' ability to interact in ways that were strategically useful for SLA in the language assessment research studies. Referring to a good number of research studies in SLA, it is noted that the L2 level differences in dyads including the L2 developmental stage differences, created an environment where they negotiate meaning as they employ confirmation checks, clarification requests, and comprehension checks (Iwashita, 2001; Porter, 1986; Yule & Macdonald, 1990). Nonetheless, no differences were found in CRTTs' use of those interaction strategies in the study. Moreover, the instances of these strategies were quite limited.

Although more empirical evidence is needed, these results may indicate that CRTTs' understanding of the goal of the assessment was an important factor for their performance. That is, as they were aware that they were in testing situations and limited in time to make decisions, it seemed that they did not attempt to challenge their CGTTs or change the direction of interaction. In addition, as they were asked to complete as much of a task as they could in the limited time, they might pursue a more efficient way to reach a conclusion. It is also possible that they only developed cohesive utterances with explicit expressions regardless of their CGTTs' L2 developmental stages. Furthermore, it is assumed that CRTTs did not try to use the aforementioned interaction strategies, which they might consider prolonged the interaction or challenged their CGTTs. Hence, it seems that the time limits and CRTTs' psychological tension from tests prevented them from attempting various interaction strategies. Moreover, instead of employing interaction strategies to make their CGTTs clarify and modify what they have said, CRTTs might have guessed what their CGTTs said and continued interaction. These results provided strong supportive evidence against the concerns related to test-takers' inconsistent performance due to the influence from another test-taker.

Discussion of Results for Research Question 4

Test-raters' rating of CRTTs' interaction ability revealed that their rating of CRTTs' ability to interact in ways that were sociolinguistically appropriate was consistent regardless of their CGTTs' L2 developmental stages. However, their rating of CRTTs' ability to interact in ways that were strategically useful for SLA varied. To be specific, the rating CRTTs received, when they engaged in interaction with CGTTHs and

CGTTSs showed statistically significant differences. CRTTs' mean scores of interacting in ways that were sociolinguistically appropriate were 2.47 out of five when interacting with CGTTLs, 2.57 with CGTTSs, and 2.27 with CGTTHs. CRTTs' mean scores of producing interactionally strategic utterances were 1.4, 1.07, and 1.87, respectively. Inferential statistics demonstrated that the differences in CRTTs' mean scores of interacting in ways that were sociolinguistically appropriate were not statistically meaningful ($F(2, 28) = .36$, *Wilks'* $\lambda = .96$, $p > .5$, partial $\eta^2 = .042$, $\chi^2(2, N=15) = .05$, $p > .5$, Kendall's $W = .002$). In contrast, the differences in CRTTs' mean scores of interacting in ways that were strategically useful for SLA were statistically significant ($\chi^2(2, N=15) = 6.94$, $p < .05$, Kendall's $W = .23$). In particular, the differences between the mean scores with CGTTHs and CGTTSs were statistically significant ($Z = 2.36$, $p < .05$, $r = .6$).

Compared to other studies, it was concluded that the findings from the study were not consistent with them. Other studies revealed that test-raters' evaluation varied as CRTTs interacted with different CGTTs. For instance, Iwashita (1999) examined her test-takers' communicative ability in which descriptive analysis showed differences in the mean scores. Low-level test-takers received higher scores when interacting with higher-level test-takers, and high-level test-takers received higher scores when interacting with the same-level test-takers. The discrepancy in the results may be related to the operationalization of interaction ability in the study. That is, as this dissertation study operationalized interaction ability differently than other studies by narrowing down interaction ability to producing cohesive utterances, the discrepancies may come from

this. Nonetheless, as test-raters' evaluation was consistent with CRTTs' performance as revealed through transcription analysis, the findings should be considered reliable and not haphazard.

In contrast, despite the fact that CRTTs' ability to interact in ways that were strategically useful for SLA did not vary across their CGTTs' L2 developmental stages, test-raters' evaluation varied. Test-raters' evaluation of CRTTs' performance indicated that CRTTs' ability to interact in ways that were strategically useful for SLA was best when they interacted with CGTTHs. Transcription analysis revealed that the frequency and needs of those strategies were least in interaction with CGTTHs. This finding indicated that test-raters could be influenced by CGTTs' L2 developmental stages when evaluating CRTTs' interaction strategies. That is, it is possible that test-raters were influenced by 1) the interaction between CRTTs and CGTTHs and 2) dearth of frequency and needs of interaction strategies, which resulted in their evaluating CRTTs' ability to use interaction strategies higher. More research should be conducted to better understand test-raters' interpretation of the situations that require interaction strategies and influence of CGTTs' L2 level in assessment situations.

Implications for Pedagogy and Future Research

The findings concerning test-takers' consistent performance regardless of the other test-takers' L2 developmental stages suggested that PA can be used as a reliable assessment tool to elicit learners' L2. Moreover, test-takers' consistent performance provided empirical evidence to decline the claim that another test-taker would only cause unreliable performance of test-takers, which is one of the foremost concerns that caused

hesitancy of employing PA. The findings that test-raters' consistent evaluation of test-takers' ability to produce linguistically accurate utterances and interact in ways that were sociolinguistically appropriate also suggested that rating could be consistent, and test-raters could be free from the influence of the other test-takers' L2 developmental stages in PA. It also provided empirical evidence to reject the claim that test-raters' inconsistent rating is predictable as their rating can be vulnerable to the other test-takers' performance.

However, concerns regarding test-raters' evaluation of test-takers' ability to interact in ways that were strategically useful for SLA remained. The findings revealed that test-takers received higher scores in interacting with higher-level test-takers than the same or lower-level test-takers. The findings implied that test-raters were possibly influenced by interaction between test-takers and higher-level test-takers. It is possible that test-raters interpreted the lower degree of interaction strategies used in test-takers' utterances as evidence of test-takers' higher ability to interact in ways that are strategically useful for SLA than in other situations. It implied that test-raters' evaluation could be influenced by the other test-takers. It finally suggested more systematic research on examining the procedure of test-raters' evaluation.

The overall findings and the procedures of data collection and analysis of the study strongly supported the pedagogical use of PA; however, at the same time, there remain many research issues regarding the framework and processes of test-development. Issues to be investigated include 1) target constructs, 2) data analysis approaches, 3) testing tasks, 4) pairing methods, and 5) rating mechanism. The findings of the study implied that the linguistic features examined in the study were robust with respect to

external factors such as the other test-takers' performance, which also supported more research to investigate linguistic features that can be consistently presented and reliably evaluated. The findings regarding the interaction ability suggested that the operationalization of interaction ability can be more inclusive and broadened to obtain more in-depth information regarding test-takers' ability to interact in ways that are sociolinguistically appropriate. The complicated nature of demonstrating and evaluating the ability to interact in ways that are strategically useful for SLA indicates that examining test-takers' use of interaction strategies can develop a bridge between SLA and language assessment. As Bachman and Cohen (1998) indicated, it has been understood while language assessment research mainly focus on the "results of acquisition", the focus of SLA research has been principally placed on the "factors and processes" of L2 acquisition (Bachman & Cohen, 1998, pp. 1 – 5). L2 learners' ability to interact in ways that are strategically useful for SLA in order to negotiate meaning with another learner can connect the interests in processes and outcome levels, which is one of the leading purposes of employing PA as a testing tool.

The ways to analyze the spoken data in the study also provided insights for future research on data quantification. The unit of analysis employed in the study was utterances which led data analysis based on the pitch contour and pauses. It is possible that the more detailed analysis method adopted in the study might have yielded different results from previous studies, which revealed no differences in terms of the amount of test-takers' talk in relation to the other test-takers' L2 developmental stages. Pica (1983) provided strong encouragement to reexamine the spoken data with different units of

analysis to determine data analysis method that can provide more in-depth outlook of learner language.

More research studies are also needed for designing and implementing PA. The study used only decision-making tasks in order to control the consistency in testing tasks. That said, more information is desired in order to know the relationship between the types of testing-tasks and nature of learner language elicited through the tasks. A quest to develop testing-tasks which can elicit more information about test-takers' processes and outcome levels in relation to L2 acquisition is on-going (Purpura, 2004). In particular, as indicated in Pica et al. (1993), depending on interactant roles, interaction requirement, goal orientation, and outcome options, L2 learners' language and interaction ability may be demonstrated differently. These differences will make meaningful contribution to examining wider ranges of learner language elicited in PA.

More research studies on different pairing methods are needed as well. The line of research includes investigating not only the influence of a range of test-takers' characteristics but also the possible influence of non-face-to-face interaction on test-takers' performance. Many high-stakes testing organizations such as ETS and CAL promote the testing setting where a test-taker interacts with a computer which is run by a pre-installed program. As they admit, however, this testing method only produces limited samples of learner language. One suggestion to supplement the limitations of the testing setting is to pair test-takers over the online system and observe and evaluate those test-takers' performance. It will require more empirical research findings regarding

reliability and validity of the assessment; however, it would be worthwhile to investigate the use of PA in the setting.

Taken together, results of this dissertation study and their implications provided strong support for employing PA as a part of testing battery and for recognizing the need for more research on PA as a possible bridge between the fields of SLA and language assessment.

Appendix A. Participant information

ID #	L1	Gender	Nationality	How long (yrs)	Eng Country (mons)	Test 1 + 2	Test 3
CRTT#1	Mandarin	F	China	10 - 15	7 - 12	34	4.1
CRTT#2	Mandarin	F	China	10 - 15	7 - 12	39	4.5
CRTT#3	Mandarin	F	China	10 - 15	18 - 24	36	4.1
CRTT#4	Mandarin	F	China	15 - 20	24 - 36	39	4.1
CRTT#5	Mandarin	F	China	10 - 15	1 - 6	39	4.1
CRTT#6	Mandarin	F	China	15 - 20	1 - 6	34	4.2
CRTT#7	Mandarin	F	China	10 - 15	1 - 6	34	4.2
CRTT#8	Mandarin	F	China	15 - 20	3 - 5	39	4.1
CRTT#9	Mandarin	F	China	15 - 20	3 - 5	35	4.2
CRTT#10	Mandarin	F	China	15 - 20	7 - 12	34	4.1
CRTT#11	Mandarin	F	China	10 - 15	1 - 6	39	3.9
CRTT#12	Mandarin	F	China	10 - 15	18 - 24	39	3.9
CRTT#13	Mandarin	F	China	10 - 15	1 - 6	39	3.6
CRTT#14	Mandarin	F	China	10 - 15	1 - 6	39	3.8
CRTT#15	Mandarin	F	China	10 - 15	1 - 6	39	4.2
CGTTL #1	Mandarin	F	China	10 - 15	1 - 6	31	4.2
CGTTL #2	Mandarin	F	China	15 - 20	7 - 12	29	4.1
CGTTL #3	Mandarin	F	China	10 - 15	18 - 24	29	4.3
CGTTL #4	Mandarin	F	China	10 - 15	1 - 6	25	4.2
CGTTL #5	Mandarin	F	China	10 - 15	1 - 6	23	4
CGTTS #1	Mandarin	F	China	10 - 15	1 - 6	34	4.3
CGTTS #2	Mandarin	F	China	10 - 15	1 - 6	37	4.2
CGTTS #3	Mandarin	F	China	15 - 20	1 - 6	33	4.1
CGTTS #4	Mandarin	F	China	10 - 15	1 - 6	39	4.2
CGTTS #5	Mandarin	F	China	15 - 20	1 - 6	34	4.2
CGTTH #1	Mandarin	F	China	15 - 20	1 - 6	40	4.7
CGTTH #2	Mandarin	F	China	10 - 15	7 - 12	40	5.1
CGTTH #3	Mandarin	F	China	10 - 15	1 - 6	40	4.6
CGTTH #4	Mandarin	F	China	15 - 20	1 - 6	40	4.8
CGTTH #5	Mandarin	F	China	10 - 15	1 - 6	40	4.7

Research Participants Needed

A doctoral student at the Educational Linguistics, the University of Pennsylvania recruits research participants who meet the following criteria.

1. Social status in the States: English language learners
2. Social status in your own country: Undergraduate/graduate student
3. Nationality: Mandarin
4. Gender: Male
5. Age: 20s – 30s

Purpose of the study

The proposed study looks at issues in speaking assessment.

Procedure of the study

1. If you meet all of the above criteria, please contact Jiyoong Lee at jiyoong@dolphin.upenn.edu.
2. You will receive a confirmation e-mail shortly with time and place of your tests.
3. You will take 3 speaking tests with 3 different people respectively.
4. Each test lasts about 5 minutes but including waiting and preparation time, you may want to secure 1 hour for this study.
5. You don't have to prepare anything but should be present on time.
6. You will be paid \$10 for participating in this study.

If you have any further questions, please contact Jiyoong Lee (jiyoong@dolphin.upenn.edu).

Appendix C Screening test

Written assessment

1. Information

Please do NOT refer to a dictionary or a grammar book. Do NOT consult with your friends, teachers, or supervisors for this test. Once you go to the next page, you cannot turn back. It will take about 30 minutes to complete this test.

Please write your name.

First Name _____
 Last Name _____
 Nationality _____
 First language _____
 Age _____
 Major _____
 Email Address _____
 Phone Number _____

What is your terminal degree in your own country or elsewhere?

1. Bachelor's
 2. Master's
 3. Ph.D.

When did you start learning English?

How long have you studied English?

1 - 3 years
 4 - 6 year
 7 - 9 years
 10 - 15 years
 15 - 20 years

Page 1

Written assessment

How long have you lived in an English speaking country?(including current residence in America.)

1 - 6 months
 7 - 12 months
 18 month - 2 years
 3 - 5 years
 6 - 8 years
 more than 10 years

Page 2

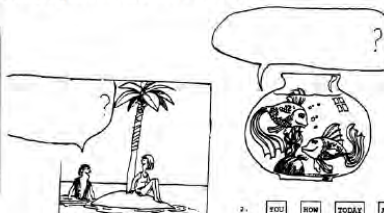
Written assessment

2. Picture completion task

In this activity you will write some questions looking at cartoons. The word for questions are in boxes, but they are not in correct order. Put the words in the correct order and write them.

Example questions and answers are below.

Example questions and answers are below.



1. YOUR ISLAND THIS IS


Please place words in squares in order.

Ex 1. _____
 Ex 2. _____

The answers are:
 Ex 1. Is this your island?
 Ex 2. How are you today?

Page 1

Written assessment




1. IS SATELLITE DISH
 THE NEW


2. ANSWER WHAT THE
 IS TO NUMBER 7

Please place words in squares in order.

1. _____
 2. _____



3. WIN THE WORD QUEST
 GAME CAN THE



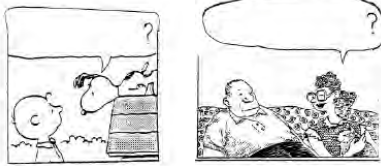
4. ARE ANGRY YOU WHY

Please place words in squares in order.

3. _____
 4. _____

Page 1

Written assessment



5. COOKIES MY
WHERE ARE
6. WE TO A MOVIE
DO CAN

Please place words in squares in order.

5. _____
6. _____



7. BET WHAT I
FOR \$2,000 CAN
8. ARE THE NEW
YOU COACH

Please place words in squares in order.

7. _____
8. _____

Page 1

Written assessment



9. DOING ARE
WE WHEN TO EAT
10. YOUR LEAVE PLANE
DOES AT 1:00

Please place words in squares in order.

9. _____
10. _____



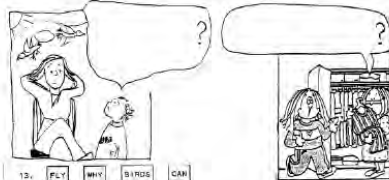
11. MY IS FINISHED
HOMEWORK
12. DO GET ME
CHANNEL 22

Please place words in squares in order.

11. _____
12. _____

Page 2

Written assessment



13. FLY WHY BIRDS CAN
14. I WEAR CAN
THIS DRESS

Please place words in squares in order.

13. _____
14. _____



15. IS FOR BREAKFAST
IT TIME
16. MY IS MOTHER
WHEN COMING HOME

Please place words in squares in order.

15. _____
16. _____

Page 3

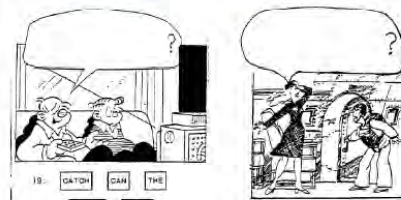
Written assessment



17. TRAIN DOES THE
WHEN LEAVE
18. BUY I WHERE
ICE DREAM CAN

Please place words in squares in order.

17. _____
18. _____



19. CATCH CAN THE
BATMAN JOKER
20. WANT WHY DO
TO JUMP YOU

Please place words in squares in order.

19. _____
20. _____

Page 4

Written assessment

Please place words in squares in order.

- 19. _____
- 20. _____

Written assessment

3. Preference task

Here are some sentences. In some of the sentences, some of the words are placed incorrectly. Look at each pair of sentences and choose the answer that you think is best.

Notice that many of the sentences are QUESTIONS. Check to see if there is a question mark at the end of the sentence before you make your decision.

- A. What is your brother doing?**
- B. What your brother is doing?**

- only A is correct
- only B is correct
- A and B are correct
- A and B are incorrect
- don't know

- A. Tomorrow Bill is going on vacation.**
- B. Bill is going on vacation tomorrow.**

- only A is correct
- only B is correct
- A and B are correct
- A and B are incorrect
- don't know

- A. Where I can buy a bicycle?**
- B. Where can I buy a bicycle?**

- only A is correct
- only B is correct
- A and B are correct
- A and B are incorrect
- don't know

Written assessment

- A. Sometimes Alexandra cleans her room.**
- B. Alexandra sometimes cleans her room.**

- only A is correct
- only B is correct
- A and B are correct
- A and B are incorrect
- don't know

- A. Why children like McDonald's?**
- B. Why do children like McDonald's?**

- only A is correct
- only B is correct
- A and B are correct
- A and B are incorrect
- don't know

- A. Do you can play the guitar?**
- B. Can you play the guitar?**

- only A is correct
- only B is correct
- A and B are correct
- A and B are incorrect
- don't know

- A. When can you visit your uncle?**
- B. When you can visit your uncle?**

- only A is correct
- only B is correct
- A and B are correct
- A and B are incorrect
- don't know

Written assessment

- A. Are the students watching television?**
- B. The students are watching television?**

- only A is correct
- only B is correct
- A and B are correct
- A and B are incorrect
- don't know

- A. Lisa has a large very car.**
- B. Lisa has a large car very.**

- only A is correct
- only B is correct
- A and B are correct
- A and B are incorrect
- don't know

- A. What can we watch on TV tonight?**
- B. What we can watch on TV tonight?**

- only A is correct
- only B is correct
- A and B are correct
- A and B are incorrect
- don't know

Written assessment

A. The teachers like to cook?
B. Do the teachers like to cook?

- only A is correct
- only B is correct
- A and B are correct
- A and B are incorrect
- don't know

A. Harry runs to his house quickly.
B. Harry runs quickly to his house.

- only A is correct
- only B is correct
- A and B are correct
- A and B are incorrect
- don't know

A. Can I take the dog outside?
B. Do I can take the dog outside?

- only A is correct
- only B is correct
- A and B are correct
- A and B are incorrect
- don't know

Written assessment

A. What time is it?
B. What is the time?

- only A is correct
- only B is correct
- A and B are correct
- A and B are incorrect
- don't know

A. Tom to work drives a motorcycle.
B. Drives Tom a motorcycle to work.

- only A is correct
- only B is correct
- A and B are correct
- A and B are incorrect
- don't know

A. Where your parents are working?
B. Where are your parents working?

- only A is correct
- only B is correct
- A and B are correct
- A and B are incorrect
- don't know

Written assessment

A. You take the bus to school?
B. Do you take the bus to school?

- only A is correct
- only B is correct
- A and B are correct
- A and B are incorrect
- don't know

A. When are you going to eat breakfast?
B. When you are going to eat breakfast?

- only A is correct
- only B is correct
- A and B are correct
- A and B are incorrect
- don't know

A. Are you a good student?
B. A good student you are?

- only A is correct
- only B is correct
- A and B are correct
- A and B are incorrect
- don't know

Written assessment

A. Why he's at home today?
B. Why is he at home today?

- only A is correct
- only B is correct
- A and B are correct
- A and B are incorrect
- don't know

A. Can the children speak Spanish?
B. The children can speak Spanish?

- only A is correct
- only B is correct
- A and B are correct
- A and B are incorrect
- don't know

A. To visit New York John wants.
B. John wants to visit New York.

- only A is correct
- only B is correct
- A and B are correct
- A and B are incorrect
- don't know

Written assessment

A. Does your teacher is sick today?
B. Is your teacher sick today?

only A is correct
 only B is correct
 A and B are correct
 A and B are incorrect
 don't know

A. The boys like not the girls.
B. The boys like the girls not.

only A is correct
 only B is correct
 A and B are correct
 A and B are incorrect
 don't know

A. Where the teacher is going?
B. Where is the teacher going?

only A is correct
 only B is correct
 A and B are correct
 A and B are incorrect
 don't know

Written assessment

A. Why can fish live in water?
B. Why fish can live in water?

only A is correct
 only B is correct
 A and B are correct
 A and B are incorrect
 don't know

A. Does your sister is talking on the phone?
B. Is talking your sister on the phone?

only A is correct
 only B is correct
 A and B are correct
 A and B are incorrect
 don't know

A. Carol hates the smell of cigarettes.
B. Carol the smell of cigarettes hates.

only A is correct
 only B is correct
 A and B are correct
 A and B are incorrect
 don't know

Written assessment

A. What is your favourite film?
B. What film is your favourite?

only A is correct
 only B is correct
 A and B are correct
 A and B are incorrect
 don't know

A. Do they like pepperoni pizza?
B. They like pepperoni pizza?

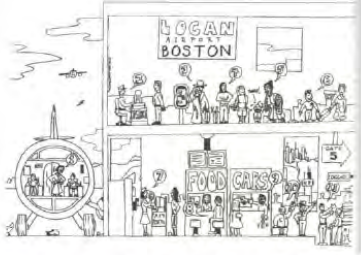
only A is correct
 only B is correct
 A and B are correct
 A and B are incorrect
 don't know

Written assessment

4. Production task

Asking questions at the airport

An airport is a very busy place. People ask for directions. They ask for help with their baggage. Some people need information about renting cars or taking taxis. Sometimes children get lost. In the picture below, people are asking questions. For example, Number 4 seems to be asking, 'What time is it?', 'Do you have the time?', 'It's three o'clock, isn't it?', or 'What time do you think our plane arrives here?'



On the lines below, write the questions that you think each person is asking.

Person #1 _____
 Person #2 _____
 Person #3 _____
 Person #4 _____
 Person #5 _____
 Person #6 _____
 Person #7 _____
 Person #8 _____
 Person #9 _____
 Person #10 _____
 Person #11 _____

Written assessment

5. Availability

Please choose time that you are NOT available.

	2/28/10	3/1/10	3/2/10	3/3/10
11:00 am - 12:00 pm	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
12:00 pm - 1:00 pm	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
4:00 - 5:00 pm	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
5:00 pm - 6:00 pm	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Appendix D. Tasks

Name: _____ Date: _____ Test-time: _____

Speaking assessments

You will take 3 speaking tests with three different partners. You will find 5 pictures in each test. You will listen to the instruction for each test. After listening to the instruction, you can take notes of your ideas and expressions you will use in the test. You will have 2 minutes to prepare for a test and talk about the topic with your partner for 3 minutes. When you are done with the test, turn to the next page. You will have a short survey and then according to your test-administrator's direction, move your seats. Your performance will be video-taped and audio-taped; however, your privacy will be kept confidentially as noted in the consent form you signed earlier. Please speak up during the tests. Your performance will be graded based on the following criteria; grammatical accuracy, your interaction with your partner, and your effort to continue the conversation. If you have any questions, please ask your questions now.

Before each test, use the following expression to introduce yourself. Say only the following information. Do NOT mention your real name, age, major, or job.

- Nice meet you!
- I'm _____ **YOUR NAME** _____.
- I'm learning English here.

I'm glad to take this speaking test with you.

Test 1



Partner Name: _____

Your choices: 1. _____ 2. _____

Reason for choice 1:

Reason for choice 2:

Expressions or words you want to use during the test:

Go to the next page

1. How difficulty was this test?

1----- 2 ----- 3 ----- 4 ----- 5

Easiest

Most difficult

2. What do you think your level was in this test?

1----- 2 ----- 3 ----- 4 ----- 5

Lowest

Highest

3. What do you think your partner's level was in this test?

1----- 2 ----- 3 ----- 4 ----- 5

Lowest

Highest

4. How familiar are you with your partner in this test?

1----- 2 ----- 3 ----- 4 ----- 5

Least

Most

5. What movies did you choose? _____, _____

6. What were the two movies you and your partner decided?

_____, _____

7. How well do you think you performed in this test?

1----- 2 ----- 3 ----- 4 ----- 5

I did the worst job.

I did the best job.

Test 2



Partner Name: _____

Your choices: 1. _____ 2. _____

Reason for choice 1:

Reason for choice 2:

Expressions or words you want to use during the test:

Go to the next page

1. How difficulty was this test?

1----- 2 ----- 3 ----- 4 ----- 5

Easiest

Most difficult

2. What do you think your level was in this test?

1----- 2 ----- 3 ----- 4 ----- 5

Lowest

Highest

3. What do you think your partner's level was in this test?

1----- 2 ----- 3 ----- 4 ----- 5

Lowest

Highest

4. How much do you know your partner in this test?

1----- 2 ----- 3 ----- 4 ----- 5

Least

Most

5. What suggestions did you choose? _____, _____

6. Were you and your partner able to make a decision? Yes / No

7. What were the two suggestions you and your partner decided?

_____, _____

8. How well do you think you performed in this test?

1----- 2 ----- 3 ----- 4 ----- 5

I did the worst job.

I did the best job.

Test 3



Partner Name: _____

Your choices: 1. _____ 2. _____

Reason for choice 1:

Reason for choice 2:

Expressions or words you want to use during the test:

Go to the next page

1. How difficulty was this test?

1----- 2 ----- 3 ----- 4 ----- 5

Easiest

Most difficult

2. What do you think your level was in this test?

1----- 2 ----- 3 ----- 4 ----- 5

Lowest

Highest

3. What do you think your partner's level was in this test?

1----- 2 ----- 3 ----- 4 ----- 5

Lowest

Highest

4. How much do you know your partner in this test?

1----- 2 ----- 3 ----- 4 ----- 5

Least

Most

5. Were you able to make a decision in test 1? Yes/ No

6. What movies did you choose? _____, _____

7. What were the two movies you and your partner decided?

_____, _____

8.

1----- 2 ----- 3 ----- 4 ----- 5

Least

Most

9.

1----- 2 ----- 3 ----- 4 ----- 5

Least

Most

10. How well do you think you performed in this test?

1----- 2 ----- 3 ----- 4 ----- 5

I did the worst job.

I did the best job.

**Please return this note to your test-administrator.
Thank you for your participation!**

Appendix E Directions for the tests

Instruction to test-administrators:

1. Greet the test-participants with smile.
2. Please check the list and distribute the name tags to them.

(Test-participants Lab ID: **Chocolate, Cinnamon, Caramel**, Sumatra, Hazlet, Limon)

3. Make sure that the core participants sit **right hand side from your perspective**.
4. Before each test, distribute the testing material (task 1, 2, and 3) and collect the material once each exam is done.
5. When the tests are over, distribute the **exit survey**.
6. Now, please play the CD you've been given.

CD: You will take 3 speaking tests with three different partners. You will find 5 pictures in each test. You will listen to the instruction for each test. After listening to the instruction, you can take notes of your ideas and expressions you will use in the test. You will have 2 minutes to prepare for this test and talk about the topic with your partner for 3 minutes. When you are done with each test, turn to the next page. You will have a short survey and then according to your test-administrator's direction, move your seats. Your performance will be video-taped and audio-taped; however, your privacy will be kept confidentially as noted in the consent form you signed earlier. Please speak up during the tests. Your performance will be graded based on the following criteria; grammatical accuracy, your interaction with your partner, and your effort to continue the conversation. If you have any questions, please ask your questions now.

7. Stop the CD.
8. Answer test-participants' questions.
9. Make sure which version (version 1, 2, or 3) you have for the current session.
10. Distribute tasks accordingly. Make sure everybody has the handout.
11. Resume the CD.

CD: Instructions. (Please refer to the handout attached. 30 minutes.)

CD: Now all the tests are over. When you are done with survey, please remain seated until other participants finish their survey. You will receive the compensation when you leave the room.

12. Stop the CD.
13. Distribute the exit survey.
14. Distribute the compensation when everybody is done with survey.

Scripts

Version 1

Task 1

The film club at your college has asked you to choose two films which would be interesting for the students to watch and then discuss. Here are the films they are considering. First, talk to each other about how interesting these different types of film would be. Then decide which two would be the best for students to discuss. Feel free to take notes of your thoughts and expressions you want to use during the speaking assessment. You will have 2 minutes to organize your thoughts and expressions. When you hear a beep, introduce yourself to your partner, exchange your idea, and make a decision. You will have 3 minutes to talk to your partner. When you hear another beep, please stop your conversation. And according to you test-administrator's direction, move your seats.

2 minutes

Beep

3 minutes

Beep

3 minutes

Task 2

I'd like you to imagine that a local café wants to attract more people. Here are some of the suggestions they are considering. First, talk to each other about how successful these suggestions might be. Then decide which two would attract most people. Feel free to take notes of your thoughts and expressions you want to use during the speaking assessment. You will have 2 minutes to organize your thoughts and expressions. When you hear a beep, introduce yourself to your partner, exchange your idea, and make a decision. You will have 3 minutes to talk to your partner. When you hear another beep, please stop your conversation. And according to you test-administrator's direction, move your seats.

2 minutes

Beep

3 minutes

Beep

3 minutes

Task 3

Here are some pictures of things that can make living in a city enjoyable. First talk to each other about how these things can help people to enjoy life in a city. Then decide which two things are the most important. Feel free to take notes of your thoughts and expressions you want to use during the speaking assessment. Feel free to take notes of your thoughts and expressions you want to use during the speaking assessment. You will have 2 minutes to organize your thoughts and expressions. When you hear a beep, introduce yourself to your partner, exchange your idea, and make a decision. You will have 3 minutes to talk to your partner. When you hear another beep, please stop your conversation.

2 minutes

Beep

3 minutes

Beep

Version 2

Task 1

Here are some pictures of things that can make living in a city enjoyable. First talk to each other about how these things can help people to enjoy life in a city. Then decide which two things are the most important. Feel free to take notes of your thoughts and expressions you want to use during the speaking assessment. Feel free to take notes of your thoughts and expressions you want to use during the speaking assessment. You will have 5 minutes to organize your thoughts and expressions. When you hear a beep, introduce yourself to your partner, exchange your idea, and make a decision. You will have 5 minutes to talk to your partner. When you hear another beep, please stop your conversation. And according to you test-administrator's direction, move your seats.

5 minutes

Beep

5 minutes

Beep

3 minutes

Task 2

The film club at your college has asked you to choose two films which would be interesting for the students to watch and then discuss. Here are the films they are considering. First, talk to each other about how interesting these different types of film would be. Then decide which two would be the best for students to discuss. Feel free to take notes of your thoughts and expressions you want to use during the speaking assessment. You will have 5 minutes to organize your thoughts and expressions. When you hear a beep, introduce yourself to your partner, exchange your idea, and make a decision. You will have 5 minutes to talk to your partner. When you hear another beep, please stop your conversation. And according to you test-administrator's direction, move your seats.

5 minutes

Beep

5 minutes

Beep

3 minutes

Task 3

I'd like you to imagine that a local café wants to attract more people. Here are some of the suggestions they are considering. First, talk to each other about how successful these suggestions might be. Then decide which two would attract most people. Feel free to take notes of your thoughts and expressions you want to use during the speaking assessment. You will have 5 minutes to organize your thoughts and expressions. When you hear a beep, introduce yourself to your partner, exchange your idea, and make a decision. You will have 5 minutes to talk to your partner. When you hear another beep, please stop your conversation.

5 minutes

Beep

5 minutes

Beep

Version 3

Task 1

I'd like you to imagine that a local café wants to attract more people. Here are some of the suggestions they are considering. First, talk to each other about how successful these suggestions might be. Then decide which two would attract most people. Feel free to take notes of your thoughts and expressions you want to use during the speaking assessment. You will have 5 minutes to organize your thoughts and expressions. When you hear a beep, introduce yourself to your partner, exchange your idea, and make a decision. You will have 5 minutes to talk to your partner. When you hear another beep, please stop your conversation. And according to you test-administrator's direction, move your seats.

5 minutes

Beep

5 minutes

Beep

3 minutes

Task 2

Here are some pictures of things that can make living in a city enjoyable. First talk to each other about how these things can help people to enjoy life in a city. Then decide which two things are the most important. Feel free to take notes of your thoughts and expressions you want to use during the speaking assessment. Feel free to take notes of your thoughts and expressions you want to use during the speaking assessment. You will have 5 minutes to organize your thoughts

and expressions. When you hear a beep, introduce yourself to your partner, exchange your idea, and make a decision. You will have 5 minutes to talk to your partner. When you hear another beep, please stop your conversation. And according to you test-administrator's direction, move your seats.

5 minutes

Beep

5 minutes

Beep

3 minutes

Task 3

The film club at your college has asked you to choose two films which would be interesting for the students to watch and then discuss. Here are the films they are considering. First, talk to each other about how interesting these different types of film would be. Then decide which two would be the best for students to discuss. Feel free to take notes of your thoughts and expressions you want to use during the speaking assessment. You will have 5 minutes to organize your thoughts and expressions. When you hear a beep, introduce yourself to your partner, exchange your idea, and make a decision. You will have 5 minutes to talk to your partner. When you hear another beep, please stop your conversation.

5 minutes

Beep

5 minutes

Beep

References

- American Council on the Teaching of Foreign Languages. (1999). *Speaking guidelines revised*. Retrieved August 2, 2008, from <http://www.actfl.org/files/public/Guidelinespeak.pdf>
- Allen, J., & Yen, M. (2001). *Introduction to Measurement Theory*. Prospect Heights, IL: Waveland Press, Inc.
- Alderson, J., & Banerjee, J. (2001). State-of-the-Art Review. Language testing and assessment (Part I). *Language Teaching*, 34 (3). 213-36.
- 2002: State-of-the-Art Review. Language testing and assessment (Part II). *Language Teaching*, 35 (2). 79-113.
- Bachman, L. (1990). *Fundamental Considerations in Language Testing*. Oxford: Oxford University Press.
- Bachman, L., & Palmer, S. (1981). The construct validation of the FSI oral interview. *Language Learning*, 31(1), 67-86.
- Bachman, L.F., & Palmer, A.S. (1996). *Language Testing in Practice: Designing and Developing Useful Language Tests*. Oxford etc.: OUP.
- Bachman, L. F. & Savignon, S. J. (1986). The evaluation of communicative language proficiency: A critique of the ACTFL oral interview. *Modern Language Journal*, 70 (4), 380-389.
- Bailey, N., Madden, C.G., & Krashen, S.D. (1974). Is there a 'natural sequence' in adult Second Language learning? *Language Learning*, 24, 235-243.
- Banerjee, J., & Luoma, S. (1997). Qualitative approaches to test validation. In C. Clapham & D. Corson (Eds.), *Encyclopedia of language and education Volume 7: Language testing and assessment*. 275-287. Dordrecht, The Netherlands: Kluwer Academic Publishers.
- Barnwell, D. (1996). *A history of foreign language testing in the United States from its beginnings to the present*. Tempe: Bilingual Review Press.
- Beebe, L. (1977). The influence of the listener on code-switching. *Language Learning*, 27, 331-339.
- Beebe, L., & Zuengler, J. (1983). Accommodation theory: An explanation for style shifting in second language dialects. In N. Wolfson & E. Judd (eds.), *Sociolinguistics and Second Language Acquisition*, Newbury House, Rowley, MA.
- Berkoff, N. A. (1985). Testing oral proficiency: A new approach. In Y. P. Lee, A. Fok, R. Lord, & G. Low (Eds.), *New directions in language testing*. 93-99. London: Pergamon Press.
- Berry, V. (2007). *Personality differences and oral test performance*. Frankfurt: Peter Lang.
- Bonk, W. J., & Ockey, G. J. (2003). A many-facet Rasch analysis of the second language group oral discussion task. *Language Testing*, 20(1).89-110.
- Bonk, W.J., & Van Moere, A. (2004). *L2 group oral testing: The influence of shyness/outgoingness, match of interlocutors' proficiency level, and gender on individual scores*. Paper presented at the Language Testing Research Colloquium, Temecula, California.
- Brown, A. (2003). Interviewer variation and the co-construction of speaking proficiency. *Language Testing*, 20 (1), 1-25.
- Brown, A. (2005). *Interviewer variability in oral proficiency interviews*. Frankfurt am Main: Peter Lang.
- Brooks, L. (2009). Interacting in pairs in a test of oral proficiency: Co-constructing a better Performance. *Language Testing*, 26 (3). 341-366.
- Butler, Y.G., & Lee, J. (2004). On-task versus off-task self-assessments among elementary school students. Proceeding of *Second Language Research Forum*.
- Canale, M. (1983). From communicative competence to communicative language pedagogy. In

- Richards, J. C., & Schmidt, R. W. (Eds.), *Language and Communication*, 2-27. London: Longman.
- Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics*, 1(1), 1-47.
- Carlsen, C. (2003). *Guarding the Guardians rating scale and rater training effects on reliability and validity of scores of an oral test of Norwegian as a second language*. Unpublished MA Thesis, Nordisk institutt Universitetet i Bergen.
- Chalhoub-Deville, M. (2003). Second language interaction: Current perspectives and future trends. *Language Testing*, 20, 369 - 383.
- Chapelle, C.A., & Douglas, D. (2006). *Assessing language through computer technology*. Cambridge: Cambridge University Press.
- Chen, J. (2007). On how to solve the problem of the avoidance of phrasal verbs in the Chinese context. *International Education Journal*, 8 (2), 348-353.
- Csepes, I. (2002). Is testing speaking in pairs disadvantageous for students? A quantitative study of partner effects on oral test scores, Nov. *ELTy*, 9(1).
- Crookes, G. (1990). The utterance, and other basic units for second language discourse analysis. *Applied Linguistics*, 11, 189-199.
- Davis, L. (2009). The influence of interlocutor proficiency in a paired oral assessment. *Language Testing*, 26 (3), 367-396.
- Doughty, C., & Pica, T. (1986). 'Information-gap' tasks: Do they facilitate second language acquisition? *TESOL Quarterly* 20(2), 305-326.
- Douglas, D., & Selinker, L. (1992). Analyzing Oral Proficiency Test Performance in General and Specific Purpose Contexts. *System* 20, 317 - 328.
- Ducasse, A. M., & Brown, A. (2009). Assessing paired orals: Raters' orientation to interaction. *Language Testing*, 26 (3), 423-444.
- Duff, P. (1986). Another look at interlanguage talk: taking task to 'task'. In R.R. Day (Ed.). *Talking to Learn: Conversation in Second Language Acquisition* Rowley, MA: Newbury house. 237-326.
- Dulay, H. & Burt, M. (1974). Natural sequences in child second language acquisition. *Language Learning*, 24, 37- 53.
- Falsgraf, C. (2009). The ecology of assessment. *Language Teaching*, 42(4), 491-503.
- French, A. (1999). Study of qualitative differences between CPE individuals and paired test formats. Internal UCLES EFL Report.
- Folland, D., & Robertson, D. (1976). Towards objectivity in group oral testing. *English Language Teaching Journal* 30, 156-167.
- Foot, M C. (1999). Relaxing in Pairs, *ELT Journal*, 53(1), 36-41.
- Foster, P., & Skehan, P. (1996). The influence of planning and task type on second language performance. *Studies in Second Language Acquisition*, 18, 299-323.
- Fulcher, G. (1996). Testing tasks: issues in task design and group oral. *Language Testing*, 13 (1), 23- 49.
- Fulcher, G. (1997). "The Testing of Speaking in a Second Language." In Clapham, C. M. and Corson, D. (eds.) *Language Testing and Assessment. Encyclopedia of Language and Education, Vol. 7*, Dordrecht: Kluwer Academic Publishers, 75 - 85.
- Galaczi, D. (2008). Peer-peer interaction in a speaking test: The case of the first certificate in English examination. *Language Assessment Quarterly*, 5(2), 89-119.
- Gass, S. (1997). *Input, interaction, and the second language learner*. Hillsdale, NJ: Erlbaum.
- Gass, S., & Varonis, E. (1994). Task variation and non-native/non-native negotiation of meaning In Gass S., and Madden, C. (Eds). *Input in Second Language Acquisition*. 149 -

161. Rowley, MA: Newbury House.
- Gass, S., & Varonis, E. (1989). Incorporated repairs in nonnative discourse. In M.R. Eisenstein (Ed.), *The Dynamic Interlanguage: Empirical Studies in Second Language Variation*. 71-86. New York: Plenum Press.
- Gass, S., & Varonis, E. (1994). Input, interaction, and second language production. *Studies in Second Language Acquisition*, 16, 283-302.
- Gass, S. M., & Mackey A. (2007). *Data elicitation for second and foreign language research*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Goldschneider, J., & DeKeyser, R. (2001). Explaining the “Natural order of L2 morpheme acquisition” in English: A meta-analysis of multiple determinants. *Language Learning*, 51, 1–50.
- Green, A. (1998). *Verbal Protocol analysis in language testing research: A handbook* (Vol. 5). Cambridge: Cambridge University Press.
- Green, S., & Salkind, N. (2005). *Using SPSS for Windows and Macintosh* (4th ed.). Upper Saddle River, NJ: Pearson Prentice Hall.
- Hilsdon, J. (1995). The group oral exam: advantages and limitations. In Alderson, J. & North, B., (eds), *Language testing in the 1990s: the communicative legacy*. Hertfordshire: Prentice Hall International, 189–97.
- Hughes, A. (2003). *Testing for language teachers* (2nd ed.). Cambridge ; New York: Cambridge University Press.
- Hymes, D. H. (1972). On Communicative Competence. In Pride, J. B., & Holmes, J.(Eds.), *Sociolinguistics*, 269-293. Baltimore, USA: Penguin Education, Penguin Books Ltd.
- Ingram, E. (1977). Basic concepts in testing. In J. P. B. Allen & A. Davies (Eds.), *Testing and experimental methods: Volume 4. Edinburgh course in applied linguistics* . 11–37. London: Oxford University Press.
- Iwashita, N. (1999). The Validity of the Peer-peer Interview in Oral Performance Assessment, *Melbourne Papers in Language Testing*, 5(2), 51–65.
- Iwashita, N. (2001). The effect of learner proficiency on corrective feedback and modified output in nonnative-nonnative interaction. *System* 29, 2.267-287.
- Iwashita, N., Brown, A., McNamara, T., & O’Hagan, S. (2008). Assessed levels of Second Language Speaking Proficiency: How distinct? *Applied Linguistics*. 29(1), 24-49.
- 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.
- Jones, E., & Gerard, H. B. (1967). *Foundations of Social Psychology*, Wiley, New York.
- Kitajima, R. (2009). Negotiation of meaning as a tool for evaluating conversational skills in the OPI. *Linguistics and Education*, 20. 145–171.
- Kowal, M., & Swain, M. (1994). Using collaborative language production tasks to promote students’ language awareness, *Language Awareness*, 3. 73-93.
- Kowal, M., & Swain, M. (1997). From semantic to syntactic processing: How can we promote it in the immersion classroom? In R.K. Johnson & M. Swain (Eds.), *Immersion education: International perspectives*. 284-309. Cambridge: Cambridge University Press.
- Lazaraton, A. (1996). Interlocutor Support in Oral Proficiency Interviews: The Case of CASE, *Language Testing*, 13, 151–172.
- Lazaraton, A. (2002). Quantitative and qualitative approaches to discourse analysis. *Annual Review of Applied Linguistics*, 22, 32-51.
- Lazaraton, A. (2008). Utilizing qualitative methods for assessment. In E. Shohamy & N.H. Hornberger (Eds). *Encyclopedia of language and education* (2nd ed.): 7 *Language testing and assessment* 197-209. New York: Springer.
- Lazaraton, A., & Davis, L. (2008). A microanalytic perspective on discourse, proficiency,

- and identity in paired oral assessment. *Language Assessment Quarterly*, 5, 313-335.
- Lantolf, J.P., (2000). Second language learning as a mediated process. *Language Teaching* 33, 79–96.
- Lantolf, J. P., & M. Ahmed. (1989). Psycholinguistic perspectives on interlanguage variation: AVygotskian analysis. In S. M. Gass, L. Selinker & D. Preston (eds.), *Variation in second language acquisition: Psycholinguistic issues* (93-108). Clevedon, England: Multilingual Matters.
- Leeser, M. J. (2004). Learner proficiency and focus on form during collaborative dialogue. *Language Teaching Research*, 8(1), 55-81.
- Liao, Y., & Fukuya, Y. J. (2004). Avoidance of Phrasal Verbs: The Case of Chinese Learners of English. *Language Learning*, 54(2), 193-226.
- Long, M. (1980). Inside the "black box": methodological issues in classroom research on language learning. *Language Learning* 30(1), 1-42.
- Long, M. (1996). The role of the linguistic environment in second language acquisition. In W. C. Ritchie & T. K. Bhatia (Eds.), *Handbook of language acquisition: Vol. 2. Second language acquisition*. 413–468. New York: Academic Press.
- Long, M., & Porter, P. (1985). Group work, Interlanguage talk, and Second Language Acquisition. *TESOL Quarterly*, 19(2), 207-228.
- Luoma, S. (2004). *Assessing speaking*. Cambridge: Cambridge University Press.
- Lumley, T. & A. Brown. (2005) 'Research methods in language testing.' In E. Hinkel (Ed.) *Handbook of Research in Second Language Teaching and Learning*. 833-855. Mahwah, NJ: Lawrence Erlbaum Associates Inc.
- Lyster, R. (2002). Negotiation in immersion teacher-student interaction. *International Journal of Educational Research*, 37, 237-253.
- Lyster, R. (2007). *Learning and teaching languages through content: A counterbalanced approach*. Amsterdam/Philadelphia: John Benjamins.
- Mackey, A. (1999). Input, interaction, and second language development. *Studies in Second Language Acquisition*, 21, 557-587.
- Mackey, A. (2002). Beyond production: Learners' perceptions about interactional processes. *International Journal of Educational Research*, 37, 379–94.
- Mackey, A., & Philp, J. (1998). Conversational interaction and second language development: Recasts, responses, and red herrings? *Modern Language Journal* 82. 338–56.
- Mackey, A., Oliver, R., & Leeman, J. (2003). Interactional input and the incorporation of feedback: An exploration of NS-NNS and NNS-NNS adult and child dyads. *Language Learning*, 53, 35–66.
- McNamara, T. (1996). *Measuring second language performance*. London: Longman.
- McNamara, T. (1997). Interaction in second language performance assessment: whose performance?. *Applied Linguistics*, 18(4). 446-466.
- McNamara, T., Hill, K., & May, L. (2002) Discourse and assessment. *Annual Review of Applied Linguistics* 22. 221-242.
- McNamara, T., & Roever, C. (2006). *Language testing: The social dimension*. Malden, MA: Blackwell.
- Nakatsuhara, F. (2006). The impact of proficiency level on conversational styles in paired speaking tests. *Cambridge ESOL Research Notes*, 25.
- Nevo, D., & Shohamy, E. (1986). Evaluation standards for the assessment of alternative testing methods: An Application. *Studies in Educational Evaluation*. 12(2), 149-158.
- O'Sullivan, B. (2002). Learner acquaintanceship and oral proficiency test pair-task performance. *Language testing* 19(3). 277 – 295.
- Perrett, G. (1990). The language testing interview: A reappraisal. In: J.H. de Jong & D.K.

- Stevenson, Editors, *Individualizing the assessment of language abilities*, 225–238. Multilingual Matters, Philadelphia.
- Pica, T. (1983). Methods of Morpheme Quantification: Their Effect on the Interpretation of Second Language Data. *Studies in Second Language Acquisition* 6. 69-78.
- Pica, T. (1994). Research on negotiation: What does it reveal about second language learning conditions, processes, and outcomes? *Language Learning*, 44 (4), 493–527.
- Pica, T. & Doughty, C. (1985). The role of group work in classroom second language acquisition. *Studies in Second Language Acquisition*, 7, 233-248.
- Pica, T., Young, R., & Doughty, C. (1987). The impact of interaction on comprehension. *TESOL Quarterly*, 21(4), 737–758.
- Pica, T., Holliday, L., Lewis, N., & Morgenthaler, L. (1989). Comprehensible output as an outcome of linguistic demands on the learner. *Studies in Second Language Acquisition*, 11(1), 63–90.
- Pica, T., Kanagy, R., & Falodun, J. (1993). Choosing and using communication tasks for second language instruction and research. In G.G.Crooks. (Ed.). *Tasks and Language Learning*. 9-34. Bristol: Multilingual Matters, Ltd.
- Pica, T., Lincoln-Porter, F., Paninos, D. & Linnell, J. (1996). Language Learners' Interaction: How Does It Address the Input, Output, and Feedback Needs of L2 Learners? *TESOL Quarterly*, 30,59-85.
- Pica, T., & Lee, J. (2009). A Comparison Study of Three Approaches to Drawing Attention to Article Form and Function on an Information-Gap Task. Proceeding of *Task Based Language Teaching*.
- Pienemann, M., & Johnston, M. (1987). Factors influencing the development of language proficiency. In D. Nunan (Ed.), *Applying second language acquisition research*. 45-141. Adelaide, Australia: National Curriculum Research Centre, AMEP.
- Pienemann, M., Johnston, M., & Brindley, G. (1988). Constructing an acquisition-based procedure for second language assessment. *Studies in Second Language Acquisition* 10(2). 217-243.
- Porter, P. (1986). How learners talk to each other: input and interaction in task centered discussions. In R. Day (ed), *Talking to learn*. 220 – 222. Cambridge, MA: Newbury House.
- Purpura, J.E. (1998). Investigating the effects of strategy use and second language test performance with high- and low-ability test takers: a structural equation modeling approach. *Language Testing* 15, 333-79.
- Robinson, P. (1995). Task complexity and second language narrative discourse. *Language Learning*, 45, 99–140.
- Salaberry, R. (2000). Revising the revised format of the ACTFL Oral Proficiency Interview. *Language Testing*, 17(3), 289-310.
- Sato, C. J. (1985). *The syntax of conversation in Interlanguage development*. Unpublished dissertation, University of California-Los Angeles.
- Savignon, S.J. (1997). *Communicative Competence: Theory and Classroom Practice*. New York: McGraw-Hill. 2nd edition.
- Shohamy, E., Reves, T., & Bejerano, T. (1986). Introducing a new comprehensive test of oral proficiency. *English Language Teaching Journal* 40. 212–220.
- Skehan, P. & Foster, P. (1999). The influence of task structure and processing conditions on narrative retellings. *Language Learning*, (49), 93-120.
- Sollenberger, H.E. (1978). Development and current use of the FSI Oral Interview test. In J.L.D. Clark, *Direct testing of speaking proficiency: Theory and application* .1-12. Princeton, NJ: Educational Testing Service.

- Spada, N. & Lightbown, P.M. (1993). Instruction and the development of questions in L2 classrooms. *Studies in Second Language Acquisition*, 15(2), 205-224.
- Spada, N. & Lightbown, P. M. (1999). Instruction, L1 influence and developmental "readiness" in second language acquisition. *Modern Language Journal*, 83(1), 1-22
- Storch, N. (2002). Patterns of interaction in ESL pair work. *Language Learning*, 52, 119-158.
- Swain, M. (2001). Examining dialogue: another approach to content specification and to validating inferences drawn from test scores. *Language Testing*, 18(3), 275 – 302.
- Swender, E. (Ed.). (1999). *Oral proficiency interview tester training manual*. New York: ACTFL.
- Tarone, E. (1985). Variability in Interlanguage use: A study of style-shifting in morphology and syntax. *Language Learning*, 35, 373–404.
- Tarone, E. (1988). *Variation in Interlanguage*. London: Edward Arnold Publishers
- Tarone, E. (1999). Expanding our vision of English language learner education in Minnesota: Implications of state population projections, *Minne/WITESOL Journal*. 16,1-13.
- Van Lier, L. (1989). Reeling, writhing, drawling, stretching, and fainting in coils: Oral Proficiency Interviews as conversation. *TESOL Quarterly*, 23(3), 489-508.
- Van Moere, A. (2006). Validity evidence in a university group oral test. *Language Testing* 23 (4) 411–440.
- Watanabe, Y. (2008). Peer–peer interaction between 12 learners of different proficiency levels: their interactions and reflections. *Canadian Modern Language Review*, 64 (4). 605-635
- Watanabe, Y., & Swain, M. (2007). Effects of proficiency differences and patterns of pair interaction on second language learning: Collaborative dialogue between adult ESL learners. *Language Teaching Research*, 11(2), 121-142.
- White, L., Spada, N., Lightbown, P., & Ranta, L. (1991). Input enhancement and L2 question formation. *Applied Linguistics*, 12, 416-432.
- Young, R. (2002). Discourse approaches to oral language assessment. *Annual Review of Applied Linguistics*, 22, 243–262.
- Young, R., & Milanovic, M. (1992). Discourse Variation in Oral Proficiency Interviews, *Studies in Second Language Acquisition*, 14, 403–424.
- Young, R., & He, A. (Eds.). (1998). *Talking and testing: Discourse approaches to the assessment of oral proficiency*. Amsterdam/Philadelphia: Benjamins.
- Yule, G., & Macdonald, D. (1990). Resolving referential conflict in L2 interaction: the effect of proficiency and interactive role. *Language Learning* 40. 539–56.