INVESTIGATING CHINESE SECOND LANGUAGE PRAGMATIC COMPETENCE IN INTERACTION USING PAIRED SPEAKING TESTS

A DISSERTATION SUBMITTED TO THE GRADUATE DIVISION OF THE UNIVERSITY OF HAWAI‘I AT MĀNOA IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

IN

EAST ASIAN LANGUAGES AND LITERATURES (CHINESE)

By

Xue Xia

August 2018

Dissertation committee:

Haidan Wang, Co-Chair
Song Jiang, Co-Chair
James Dean Brown
Thom Hudson
Li Jiang
Seongah Im

ACKNOWLEDGEMENTS

I am very happy to write this page and say goodbye to my Ph.D. life in Hawaii, where I had many wonderful memories.

First of all, I would like to thank all of my committee members: Dr. Haidan Wang, Dr. Song Jiang, Dr. James Dean Brown, Dr. Thom Hudson, Dr. Seongah Im, and Dr. Li Jiang. Thank them all for their teamwork and guidance throughout my dissertation journey. My Co-Chair-Professor Wang has given me unlimited help throughout my entire Ph.D. study life. She treats students like the way moms love their children, providing guidance and being full of tolerance and understanding. She is approachable and we can always talk like respected colleagues. She has had the most influence on me in my Ph.D. life. Co-Chair, Professor Song Jiang, also has played a significant role in my Ph.D. studies. When facing difficulties, he is always there to help and encourage me. His wisdom and generosity gave me great strength. Dr. Brown presented a lot of practical comments and suggestions for my dissertation. He is very responsive and helpful all of the time. Dr. Hudson made a deep lasting impression on me. I really appreciated that for an entire semester one year, he gave us a free R workshop and printed all the materials for us every week. He gave a lot of very important suggestions at the beginning of my dissertation design, and helped me set up a good foundation. Dr. Im is also a professor I respect and like very much. I have taken many of her statistics classes, and was often amazed by her solid professional knowledge. She is very supportive and often makes me feel very warm. Dr. Li Jiang came to our department in the late stages of my Ph.D., so I didn't have the chance to take her classes. But she was still willing to serve on my doctoral committee, and I am very grateful to her.

I want to thank those who offered me suggestions and help in my dissertation writing process. Dr. Alfred Rue Burch helped inform the theoretical framework. Dr. Yi Zhang, Dr. Mingming Zhang, Tianzi, Dr. Lixin Liu, Dr. Li Zhang, Jing Wang generously helped me find participants. Dr. Liulin Zhang and Mengying Zhai gave me suggestions for my data. I also thank all the lovely Chinese language learners who participated in my speaking test. I am grateful to our graduate advisor-Cherry Rojo Lacsina, who sent me a lot of useful reminders. I also thank Dean Laura Lyons and Graduate Advisor Mee-Jeong Park for their encouragement and support.

I would also like to thank my colleagues at the University of Wisconsin-Milwaukee. My colleague Andrew Olson has been very understanding and supportive throughout my dissertation writing process. Our department Chair Dr. Andrew Porter and Associate Dean Dr. Jasmine Alinder have strong supported me as well.

I am always thankful for the research funds I received to help me complete the dissertation, including: the NFMLTA/NCOLCTL Graduate Students Research Support Award (by National Council of Less Commonly Taught Languages), Cheng & Tsui Professional Development Award (Cheng & Tsui Company), The University of Hawai'i at Mānoa -Peking University Exchange Program Award (by Center for Chinese Studies at UHM) and EALL research fund.

Finally, I want to express my gratitude to my parents, sister, and other family members for their encouragement and support. Thanks to the church sisters' prayers and friends' encouragement. At the same time, I would also like to thank Nate for supporting and helping me during my dissertation writing. His efficiency, productivity, self-discipline, and diligence I learn from all the time.

ABSTRACT

Second language pragmatic competence, the ability of language learners to understand and perform the pragmatic functions of target languages in social interactions (Taguchi & Roever, 2017), develops over time and is an important research area. Youn (2015) defines L2 pragmatic competence in interaction as the ability of interactive participants to use different pragmatic and interactive resources to achieve pragmatic meaning and conduct actions in organized sequences. In the current study, the peer-to-peer paired speaking test, considered as a way of classroom assessment, was employed to investigate Chinese learners' second language (L2) pragmatic competence in interaction in the personal language use domain. An analytical rubric was developed based on related conversational organizations and interaction relevant studies for raters to award scores. Mixed method design was employed to analyze the data – test takers' in-test discourse.

The results indicate that open role-play and situational topic discussion (extended discourse) tasks were effective in eliciting interactions for assessing the construct. The test content was based on the needs analysis of the most commonly used situations, topics, and language functions in this domain. When using the analytical rubric to assess test takers' in-test discourse, inter-rater reliability did not meet established thresholds. The detailed results of DA for 12 excerpts of in-test discourse not only identified additional components (language use and situation response), but also distinguished new interactional features within the three major interactional rating categories (turn-taking organization, sequence organization and topic management). DA revealed that all the rating categories were distinguishable across three different competence levels, a finding that was confirmed via quantitative analyses (descriptive

statistics and repeated measures ANOVA). Based on the general interactional features summarized from in-test discourse, the developmental trajectory of Chinese learners' L2 pragmatic competence in interaction was summarized by five elements: frequency, proactivity, complexity, content and coherence. Specifically, as L2 pragmatic competence in interaction develops, learners are more active, and their cognitive abilities are more capable of dealing with faster turn-takings, more complex structures, and the more coherent delivery of deeper content. The findings from the mixed method approach were strengthened and could help to revise the analytical rating rubric and improve future rater training. In summary, the findings offer the potential to contribute to the future assessment of Chinese learners' L2 pragmatic competence in interaction.

TABLE OF CONTENTS

Acknowledgement…… III
Abstract……. V
List of Tables…… X
List of Figures……... XI
List of Excerpts……. XII
List of Abbreviations…… XIII

Chapter 1. Introduction……... 1
Previous Studies……. 5
Research Gaps……..... 9
Study Purposes and Research Questions……. 10
Summary……..... 11

Chapter 2. Review of the Literature……... 13
Interactional Competence…… 13
Conversational Organization……... 20
Employing the Discursive Approach…….. 26
The Operation of L2 Pragmatic Competence in Interaction……... 29
Paired Speaking Tests……. 32
Discourse Analysis…….. 34
Data Exploration Strategy and Steps……... 35
Mixed Methods Design……... 37
Summary……. 39

Chapter 3. Method…… 41
Participants……... 41
Instruments……... 43
Rating Criteria…….. 49
Procedure……. 50
Data Analysis……... 54
Summary…….. 56

Chapter 4. Results…… 57
Quantitative Analysis……... 57
Descriptive statistics…… 57
Repeated measures ANOVA……... 60
Reliability estimation…….. 64
Correlation analyses…… 67

Qualitative Analysis……. 69
Transcribing and coding…….. 69
Discourse analysis……... 77
High-competence group…….. 78

Middle-competence group..... 95
Low-competence group..... 105
A Synopsis of the DA results.....… 112
Summary..... 117

Chapter 5. Discussion..... 120
Target Domain and Task Design..... 120
Rating Reliability..... 124
The Measured Construct..... 127
The Reliability of Mixed Methods Approach..... 136
Summary..... 142

Chapter 6. Conclusion..... 144
Implications..... 145
Limitations..... 147
Suggestion for Future Research..... 148
Summary..... 149

Appendix A:	Background Questionnaire……	151
Appendix B:	Survey about International Students' Needs of Chinese Use in the Personal Domain from Teachers' Perspective..	154
Appendix C:	The Speaking Test……	163
Appendix D:	Rating Criteria for Speaking Test……	185
Appendix E:	DA Transcription Conventions: Adapt from Atkinson & Heritage (1984)..	188
References	……...	189

LIST OF TABLES

1.	Interactional Features Selected and Defined by Wang.…….....	19
2.	Test-takers' Background Information.…….	42
3.	Background Information, Participants in the Needs Analysis Questionnaire.…	44
4.	Three Themes of the Personal Language-use Domain……	45
5.	The Most Commonly Mentioned Situations, Topics and Speech Acts……...	46
6.	Structure of the Three Independent Speaking-proficiency Tasks……	47
7.	Structure of the Three Interactive Tasks Assigned to the Paired Test-takers…..	48
8.	Supporting Analyses for Each Research Question……..	55
9.	Descriptive Statistics for the Rating Categories……..	58
10.	Task Means across Levels……...	59
11.	The Means of Rating Categories across Levels……...	60
12.	ANOVA Source Table for Scores by Competence Level and Task……	61
13.	ANOVA Source Table for Scores by Competence Level and Rating Category.	62
14.	Inter-rater Reliability for the Five Rating Categories (Tasks Considered……... Separately)	65
15.	Inter-rater Reliability for the Five Rating Categories (All Tasks Combined)…. and for the Entire Test	66
16.	Internal Consistency Reliability for Each of the Five Rating Categories; the…. Three Rating Categories with Interactional Features; and Overall	67
17.	Summary of Coding Results, Raters' Notes for "Language Use"……...	72
18.	Summary of Coding Results, Raters' Notes for "Situation Response"……...	73
19.	Summary of Coding Results, Raters' Notes for "Turn-taking Organization"….	73
20.	Summary of Coding Results, Raters' Notes for "Sequence Organization"…….	74
21.	Summary of Coding Results, Raters' Notes for "Topic Management"……...	75
22.	Summary of Coding Results, Raters' Notes for "Other Elements"…….	76
23.	Summary of Distinguishing Features of Language Use, by Competence Group	112
24.	Summary of Distinguishing Features of Situation Response, by Candidates'… Competence Levels	114
25.	Summary of Distinguishing Features of Turn-taking Organization, by…….. Candidates' Competence Levels	114
26.	Summary of Distinguishing Features of Sequence Organization, by…….. Candidates' Competence Levels	115
27	Summary of Distinguishing Features of Topic Management, by Candidates'… Competence Levels	117

LIST OF FIGURES

1. The scoring by competence level and rating category…….. 63
2. Relationship between the paired and solo speaking tasks……. 68
3. Five elements of the developmental trajectory of Chinese learners' L2…….. 134 pragmatic competence in interaction

LIST OF EXCERPTS

1. High-competence pairs (Ban & Zeng)……. 78
2. High-competence pairs (Ti & Ai)..……... 81
3. High-competence pairs (Huang &Wang)…… 81
4. High-competence pairs (Ya & Men)…… 84
5. Middle-competence pairs (A & Ban)……... 96
6. Middle-competence pairs (Tian & Bai)…… 98
7. Middle-competence pairs (Fu & Xie)…….. 100
8. Middle-competence pairs (Wu & Da)…….. 102
9. Low-competence pairs (Ge & A)……. 106
10. Low-competence pairs (Ye & Li)…… 107
11. Low-competence pairs (Fu & Da)…… 108
12. Low-competence pairs (Zhou & Cui)…….. 110

LIST OF ABBREVIATIONS

1.	ANOVA	Analysis of variance
2.	APs	Adjacency pairs
3.	CA	Conversation analysis
4.	CEFR	The Common European Framework of Reference for Languages
5.	CTT	Classical test theory
6.	DA	Discourse analysis
7.	EAP	English for academic purposes
8.	FCE	First Certificate of English
9.	FPP	First pair-part
10.	L2	Second language
11.	MMR	Mixed methods research
12.	OPIs	Oral proficiency interviews
13.	SLA	Second language acquisition
14.	SPP	Second pair-part

CHAPTER 1

INTRODUCTION

A seminal term in second language acquisition (SLA), *interlanguage* was first proposed by Selinker (1972), and refers to the development system of a learner's target language. An important body of research has subsequently focused on interlanguage pragmatic competence – the ability of language learners to understand and perform the pragmatic functions of their target languages in social interactions (Taguchi & Roever, 2017) – and how it develops over time. The term "second language" (L2) refers to any languages other than their native languages that people learn, whether in natural contexts or through education (Krashen, 1981). As such, the terms *interlanguage pragmatics* and *L2 pragmatics* are interchangeable, and the latter will be employed in the current study.

In the past three decades, L2 pragmatics has become one of the core areas of SLA research. The communicative competence model (Bachman, 1990; Bachman & Palmer, 1996, 2010; Canale & Swain, 1980; Celce-Murcia, 2007; Celce-Murcia, Dörnyei, & Thurrell, 1995) positions pragmatic competence as an important component of L2 ability. This model has now developed into various versions, notably one based on interaction (Celce-Murcia, 2007; Celce-Murcia, el., 1995), interpreted as speakers' communicative competence in an interactive environment. However, most of these models emphasize that, in addition to strategic abilities and a grasp of grammar and discourse conventions, it is critical that learners have an understanding of social conventions, also known as communication rules if they are to avoid communication errors. In other words, L2 speakers must have both pragmatic knowledge (of language tools for communicating in the target language) and sociolinguistic knowledge (of

cultural rules and norms, expectations of different social roles and appropriate behaviors, etc.). Moreover, these two types of knowledge must be mapped to each other so that learners can choose the language forms appropriate to achieving their communicative goals in specific contexts.

Thus, pragmatic competence can be said to be multi-dimensional and to have multiple layered. In addition to the two main aspects of pragmatic knowledge described above, the application of non-linguistic knowledge can reflect how learners want to present themselves in social interactions. The understanding and evaluation of context, as part of social pragmatics, are essentially dynamic; and the term social pragmatic knowledge (Taguchi & Roever, 2017) has been coined to refer to a person's ability to unravel a complex background involving a range of elements (e.g., background, relationships, influences, attitudes, and positions) while at the same time detecting subtle changes in these elements, and adapting to such changes when interacting with other people.

Youn (2015), in defining the construct of pragmatic competence in interaction, highlighted its two distinct theoretical foundations: (1) communicative competence as revised by Celce-Murcia et al. (1995) and Celce-Murcia (2007); (2) discursive pragmatics (Kasper, 2006). As briefly noted above, Celce-Murcia's (2007) model incorporates the concept of interactional competence (Kramsch, 1986), which emphasizes the role of individual conversational knowledge in accomplishing diverse pragmatic actions, and can be subdivided into actional competence, conversational competence and paralinguistic competence. However, Celce-Murcia's model has been criticized for neglecting learners' knowledge of sequence organization, that is, the effective and meaningful organization of

interactions into conversations via series of turns (Schegloff, 2007). The theory of discursive pragmatics (Kasper, 2006; Kasper & Ross, 2013) compensates for this deficiency, by emphasizing sequence structure within conversation analysis (CA), and more generally, how interlocutors achieve pragmatic meanings and conduct actions in organized sequences. Nevertheless, the common ground between the Celce-Murcia variant of the communicative competence model and discursive pragmatics should not be overlooked; both aim to achieve a better understanding of interactions and of how conversationalists generate and understand the meaning of conversations. L2 pragmatic competence in interaction was defined as the ability of the participants in an interaction to use various pragmatic and interactive resources to achieve pragmatic meaning and to conduct actions in organized sequences (Youn, 2015).

The study of L2 pragmatics has been dominated by speech act theory (Searle, 1969) and politeness theory (Brown & Levinson, 1987). Recent research also describes the development of interactional competence, while studies of routine formulae (Coulmas, 1981) and implicature (Levinson, 2000) are no longer as popular as they once were. Longitudinal studies have also predominated, as being best suited to revealing the developmental trajectory of L2 pragmatic competence in interaction. However, one cross-sectional approach – speaking assessments, in which learners are grouped and the groups compared – has also been found effective, as a means of predicting the development of L2 pragmatic competence (Taguchi & Roever, 2017).

Amid the wide array of existing language assessments, proficiency tests are among the most important, and are mainly used to assess L2 learners' accuracy, fluency, and ability to use a variety of discourse strategies. Due to the prevalence of

communicative competence models and of the communicative language teaching approach, oral proficiency interviews (OPIs) have become one of the most influential types of oral assessments. However, their validity has been questioned, since OPIs and natural conversation are distinct interactions with very different interactional patterns and features (e.g., Brown, 2005; Lazaraton, 2002; van Lier, 1989; Young & He, 1998). Moreover, as mentioned above, some language-testing specialists have criticized most models of communicative competence for concentrating on individuals' static language performance from a cognitive perspective, while neglecting the dynamic nature of interactions and their social dimension (e.g., Chalhoub-Deville, 2003; McNamara, 1996). Communicative competence is built on a psycholinguistic foundation, and its core theories on a rational model and a cognitivist paradigm, whereby thoughts indicate actions, and conversely, actions express speakers' intentions and reveal their mental states (Edwards, 1997). It also presumes that communicative competence can be inferred from individual test-takers' cognitive abilities. However, this is again to ignore the inseparability from social context of all interactions, which are co-constructed by all of their participants. Kramsch (1986, p. 386) first used "interactional competence" as an alternative to the notion of "language proficiency" as the target of L2 learning, and defined it as a "dynamic process of communication built through the collaborative effort of the interactional partners". Since then, interactional competence has attracted considerable attention from scholars of both L2 pragmatics and language assessment.

This context helps explain the recent growth in the popularity of paired and group language assessments, which are considered to be best suited to eliciting interactional features when assessing L2 pragmatic competence in interaction. These

techniques have been applied extensively, not only in small-scale classroom assessments (e.g., Brooks, 2009; Ducasse & Brown, 2009) but also in large-scale language tests, such as the University of Cambridge ESOL examinations (e.g., Galaczi, 2014; Taylor, 2001). In paired speaking tests, two test-takers are placed together, and administrators guide and observe them during the whole process, without any direct conversational involvement – in contrast to OPI, where in most cases are testers ask questions and the test-takers reply to them. Thus, power and status between the interlocutors are more balanced in paired speaking tests, which can therefore produce more everyday conversation-like interactions (Kley, 2015). Moreover, since it can elicit a greater variety of interactional features (Brooks, 2009), pairing provides more evidence from which to infer the test-takers' levels of L2 pragmatic competence in interaction (Fulcher, 2003; Fulcher & Davidson, 2007). In addition, since these tests resemble the pair-work activities that are frequently used in classroom language instruction, they may have a positive washback effect for L2 learners (Ducasse & Brown, 2009; Taylor, 2011).

Previous Studies

Apart from the work of Youn (2013, 2015), little research on how to directly test L2 pragmatic competence in interaction has been conducted, with most studies focusing instead on how to assess interactional competence. More empirical research, including research on languages other than English in L2 pragmatic competence in interaction, is therefore urgently needed.

Based on Youn's (2013, 2015) research and other studies (Galacizi, 2014) on how to assess L2 interactional competence, researchers can continue to explore how to better assess L2 pragmatic competence in interaction. Research findings have already

contributed greatly to our understanding of the constructs of interactional competence and L2 pragmatic competence in interaction, the design of appropriate task types, and the creation of valid rating scales. The relevant prior work will be discussed in the following four categories: (1) language and language-use domain; (2) task type; (3) test content; and (4) research methods.

Language and language-use domain. To date, the use of peer-to-peer paired speaking tests to investigate L2 interactional competence has been mainly in L2 English contexts (e.g., Brooks, 2009; Galaczi, 2004, 2008, 2014; He & Dai, 2006; May, 2009, 2011; Wang, 2015; Youn, 2013, 2015), and the language-use domain of such research has normally been English for academic purposes (EAP) (e.g., Brooks, 2009; May, 2009, 2011; Wang, 2015; Youn, 2013). The small number of studies that have been conducted in languages other than English have included Ducasse and Brown's (2009) research on the interactional competence of college students learning elementary Spanish in Australia, and Kley's (2015) paired speaking tests of intermediate L2 German learners in the U.S.

Task type. Recently, the key findings about the constructs in paired speaking test discourse have been based primarily on peer-to-peer formal discussions (e.g., Brooks, 2009; Ducasse & Brown, 2009; Galaczi, 2004, 2008, 2014; He & Dai, 2006; May, 2011). However, Youn (2013, 2015), Wang (2015) and Kley (2015) have all analyzed whether other task types are suitable for eliciting various interactional features. Youn (2013, 2015), for example, used an open role-play task, finding that it allowed test-takers to naturally negotiate and interact with each other, and that it was a valid task type for investigating L2 pragmatic competence in interaction within test discourse. Wang (2015) explored the differential effects of task types including spot-the-difference, story

completion, decision-making and free discussion on interactional patterns, interactional features and competence scores. Wang found no clear correlation between task types and interactional features, and also pointed out that free-discussion tasks, which belongs to the category of extended conversation (see Taguchi & Roever, 2017), could be used in classroom-based achievement tests, since they elicit more natural conversations as well as more types of interactional features. Lastly, Kley (2015) compared the differences in repair systems across a jigsaw task and a discussion task, and found that they exhibited more similarities than differences.

Test content. Some language testers have investigated the test-content constructs from a macro point of view – for example, Galaczi (2004, 2008), who identified three major interactional patterns – while others (e.g., Brook, 2009; Ducasse & Brown, 2009; Galaczi, 2014; May, 2011; Wang, 2015) have taken a micro approach, analyzing the interactional features highlighted in actual test discourse.

Galaczi (2004, 2008, 2014) adopted CA to analyze data obtained from the peer-to-peer formal discussion section of the oral exam of the University of Cambridge ESOL First Certificate of English (FCE) examination, and identified three broad patterns of interaction – collaborative, parallel, and asymmetric – as well as a blended form, comprising any two of the three. Galaczi also analyzed the relationship between interactional patterns and paired speaking test scores, and found that those students who performed a collaborative interactional pattern tended to receive the highest scores, while those who exhibited parallel interaction always obtained the lowest scores.

Many researchers have also recognized that interactional features within test discourse can aid our understanding of the constructs of L2 pragmatic competence in

interaction and interactional competence, and have designed rating scales appropriate to assessing students' relative performance in these areas. Ducasse and Brown's (2009) pioneering research on which interactional features (e.g., "turn length", "turn domination", and "turn speed") tended to make interactions more successful, from raters' perspectives, divided those features into three categories: non-verbal interpersonal communication (gaze and body language), interactive listening (supportive listening and comprehension), and interaction management (horizontal and vertical management). Wang (2015), on the basis of Ducasse and Brown's model, further refined the second and third categories, and – according to the frequency of the relevant interactional features' appearance in previous empirical studies – selected 17 principal features and defined them operationally. Some other researchers have also summarized interactional features: for example, Brooks's (2009) "expressing incomprehension", "paraphrasing", "topic closure", "asking for help", and so on.

Most research has merely noted what interactional features were salient in language testing discourse. However, Galaczi (2014) conducted a deeper exploration of what interactional features could distinguish among L2 learners in terms of their language proficiency, and found that while "topic development", "listener support", and "turn-taking management" appeared at all levels of test discourse, the type and frequency of occurrence of each feature were not the same, and thus could be considered distinguishing features.

Research methods. To date, most research on L2 pragmatic competence in interaction and on interactional competence that has relied on peer-to-peer paired speaking tests has also used mixed methods to some extent. Both of the most widely

accepted mixed-methods approaches, "sequence" and "dominance" (Dörnyei, 2007, p. 169), have been represented.

In sequence approaches, the qualitative-method aspect has consisted largely of transcribing within-test discourse (e.g., Brooks, 2009; Galaczi, 2014; He & Dai, 2006; Youn, 2013), often with interactional features coded according to various coding schemas (e.g., Brooks, 2009; Galaczi, 2014; He & Dai, 2006). Some studies further analyzed the test discourse in depth, employing discourse analysis (DA) (e.g., Brooks, 2009) or CA (e.g., Galaczi, 2004, 2014; Youn, 2013, 2015). As for their quantitative aspects, most sequence studies have focused on the frequency of particular interactional features' occurrence in test discourse (e.g., Brooks, 2009; Galaczi, 2014; He & Dai, 2006).

In the dominance approach, some studies that were chiefly qualitative (e.g., Kley, 2015) transcribed all the documents related to raters' ratings, such as their verbal reports, and followed this with CA (e.g., Ducasse & Brown 2009) or DA (e.g., May, 2011). Some other studies were conducted primarily under the framework of CA, with a minor quantitative element (e.g., Kley, 2015). On the other hand, some studies such as Wang's (2015) were dominated by quantitative analysis, with a large amount of inferential statistical analysis performed based on transcriptions of the test discourse and coding.

Research Gaps

As the above summary suggests, research using peer-to-peer paired speaking tests to investigate L2 pragmatic competence in interaction and/or interactional competence is still in its infancy. As such, unsurprisingly, various research gaps exist. Not only have there been very few such studies in the field of SLA, but none at all were mentioned in the most recent literature reviews of L2 pragmatics (Yang, 2018) and

speaking assessment (Liao, 2018) in the Chinese L2 field. With regard to task type, formal discussion tasks were employed in most research, presumably because it was conducted in formal settings (e.g., Brooks, 2009; Ducasse & Brown, 2009; Galaczi, 2004, 2008, 2014; He & Dai, 2006; May, 2011), with only a handful of studies using other task types, albeit still for academic purposes, as discussed above (Kley, 2015; Wang, 2015; Youn, 2013). In terms of distinguishing interactional features, the majority of researchers have devoted their efforts simply to establishing which such features actually were present in their data, and only a few studies have focused on these features' potential as a means of differentiating between different L2 proficiency levels (e.g., Galaczi, 2014; Youn, 2013).

Therefore, more experimental studies using peer-to-peer paired speaking tests to investigate both L2 pragmatic competence in interaction and L2 interactional competence are urgently needed, especially in languages other than English. Likewise, the language test domain should be extended beyond academic purposes; more diversified task types ought to be employed, especially ones suitable for use in non-formal settings; and more research attention should be paid to the potential practical value of distinguishing among interactional features.

Study Purposes and Research Questions

To help fill some of the above-mentioned research gaps, the three primary purposes of the current study are: (1) to investigate the developmental trajectory of learners' L2 pragmatic competence in interaction, via the distinguishing interactional features in their test discourse, to further deepen our understanding of the construct of L2 pragmatic competence in interaction in a Mandarin Chinese context; (2) to design open

role-play and situational topic discussion tasks that are appropriate to the peer-to-peer paired speaking test format, and ensure that they are adequate to eliciting diverse interactional features; and (3) to design rating rubrics for the assessment of Chinese learners' L2 pragmatic competence in interaction. It will be guided by the following four research questions:

1. How effectively do the three paired speaking tasks developed in this study reflect Chinese learners' L2 pragmatic competence in interaction? To what extent do these tasks strike a balance between standardization and authenticity?

2. When using an analytical rubric with interactional features, to what extent can raters ensure the reliability and consistency of their rating?

3. What features useful for distinguishing between varied levels and tasks are identifiable in test-takers' paired test discourse? How much can those distinguishing interactional features deepen our understanding of the developmental trajectory of Chinese learners' L2 pragmatic competence in interaction?

4. To what extent are the findings from mixed methods design reliable and how can they enhance the validity of the future assessment of Chinese learners' L2 pragmatic competence in interaction?

Summary

The purpose of this chapter has been to introduce the background and context of the current research and to emphasize that little or no research has previously been conducted on the assessment of L2 pragmatic competence in interaction in the Chinese language. Following a summary of the relevant major previous studies, the gaps in this research field were noted, and the purposes of the current study and the corresponding

research questions were outlined. In the next chapter, the literature related to the assessment of L2 pragmatic competence in interaction will be reviewed.

CHAPTER 2

REVIEW OF THE LITERATURE

This chapter reviews the literature relevant to the current research. By way of describing the construct of L2 pragmatic competence in interaction, it introduces the construct of interactional competence; conversational organization; how discursive approaches have been employed in L2 pragmatics and test-performance discourse; and how the construct of L2 pragmatic competence in interaction has been used in assessment studies. Then, it reviews the literature on the methodology used in the current study, including paired speaking tests, DA, various data-exploration strategies and steps, and mixed method research (MMR).

Interactional Competence

A broad consensus holds that L2 pragmatic competence in interaction is profoundly affected by the construct of interactional competence. A thorough understanding of this construct requires a knowledge of specific language assessment practices as well as theoretical findings from SLA.

SLA theory. After Kramsch first defined and used interactional competence in 1986, Young (2000, 2008, 2011, 2012) developed his own theoretical framework of interactional competence that differs from communicative competence in several aspects, and has motivated a considerable body of research on interactional competence in language assessment.

Features of interactional competence. For Kramsch (1986), interactional competence has the following four characteristics: it features co-construction, is tied to specific discursive practices, is distinguished by intersubjectivity, and is constructed by

general interactional resources.

Among these characteristics, many scholars propose or assume that co-construction is the most critically important. Jacoby and Ochs (1995, p. 171), for instance, defined interactional competence as "the joint creation of a form, interpretation, stance, action, activity, identity, institution, skill, ideology, emotion, or other culturally meaningful reality." Similarly, Young and He (1998, p. 7) emphasized that interactional competence is "not an attribute of an individual participant" but rather "something that is jointly constructed by all participants". In other words, interactional competence does not refer to a single person's static ability to engage in a discursive practice, but to all interlocutors' dynamic ability to take joint action in a particular context.

The second characteristic refers to how interactional competence pertains to language used in a particular discursive practice, rather than being a language ability independent of such context (Young, 2000). McNamara and Roever (2006) highlighted the differences between OPI and natural conversations, and many studies of interaction in oral assessment have reported similar results (e.g., Okada & Greer, 2013). In OPI, the power relationship between testers and test-takers is unequal; and in terms of turn-taking, the testers always initiate turns and the test-takers only respond to them. This is vastly different from everyday conversation, in which all participants have much more freedom to initiate turns and to respond to others' turns, or elect not to. This has raised important questions about the validity of using OPI test scores as proxies for candidates' oral ability in daily life (Young & He, 1998).

The third characteristic closely connected with interactional competence, intersubjectivity, was defined by Wells (1981, p. 46) as:

> Any act of linguistic communication involves the establishment of a triangular relationship between the sender, the receiver, and the context of situation: The sender intends that, as a result of his communication, the receiver should come to attend to the same situation as himself and construe it in the same way.

If conversations are to go smoothly, all participants need to have a shared knowledge of a sequence organization, so that listeners can understand speakers' meanings and then give appropriate responses (Heritage, 1984; Young, 2008; Youn, 2013).

The theory of interactional competence encompasses a set of general resources required to construct interactions, which Young (2011) enumerated as seven, organized into three categories: (1) identity resources (participation framework); (2) linguistic resources (register; modes of meaning); and (3) interactional resources (speech acts, turn-taking, repair and boundaries). A *participation framework* refers to all the participants' identities in the interaction. *Register* means the features of pronunciation, vocabulary, and grammar, etc., that form a specific practice. *Modes of meaning* are how participants construct interpersonal, experiential and textual meanings in interaction. *Speech acts* are the selection of actions and those actions' sequential organization in a practice. *Turn-taking* refers to how participants select the next speaker, and when they may end a turn and start the next turn. *Repair* refers to the means interlocutors use to respond to interactional difficulties. And *boundaries*, the opening and closing of a practice, can be used to distinguish one practice from another. Only when in possession of all of these resources can participants co-construct a discursive practice (Young, 2000).

Differences between interactional competence and communicative competence. Interactional competence, though strongly rooted in the theory of communicative

competence (Young, 2011), emphasizes joint construction with others rather than individuals' knowledge, and employs different methodologies for defining constructs (Chapelle, 1998). The three main factors that distinguish interactional competence from communicative competence are summarized below.

First, interactional competence elaborates communicative competence's model considerably. Specifically, Celce-Murcia (2007) added three subcomponents – actional, conversational, and paralinguistic competence – to the original four-subcomponent model, wherein a learner's communicative competence generally includes linguistic, pragmatic, discourse, and strategic competence (Bachman & Palmer, 2010). Among Celce-Murcia's additions, actional competence refers to one's knowledge of how to perform various actions in different interactions; conversational competence mainly relates to the turn-taking system; and paralinguistic competence is related to the employment of body language, physical touching, silence/pauses, and so forth. These three competences are also covered by Young's (2008) seven resources for forming interactions – specifically, the linguistic and interactional resource categories – although the terminologies differ. In short, the traditional theoretical model of communicative competence does not emphasize people's competence to interact in conversations, and interactional competence fills this gap.

More specifically, interactional competence is what a person does with others to communicate accurately, appropriately, and effectively, rather than what he/she *knows* about doing so (Young, 2011): a further reminder of the centrality of co-construction to interactional competence as a concept. Ross and Kasper (2013) criticized pragmatic competence, one of the four subcomponents of the communicative competence model, on

the grounds that it implies a one-sided psycholinguistic perspective, at the expense of the social dimension. Potentially, this denial of co-construction, albeit partial, calls into question the entire theoretical underpinnings of the communicative competence model.

As briefly noted above, interactional competence also employs a different methodology for defining constructs than communicative competence does. As Chapelle (1998) explained, a construct can be defined as a trait, as a behavior, or as the combination of both. If it is defined as a trait, a person's consistent performance relates to his/her fundamental knowledge and speech-production processes; but if defined as behavior, such performance is connected with the observational context. For Ross and Kasper (2013), communicative competence's definition belongs to the first of these two categories. However, as Young (2011) pointed out, neither is entirely appropriate to a definition of interactional competence. The third or combined method, also known as interactionalist definition (Weir, Vidakovic, & Galaczi, 2013), has thus been adopted to define interactional competence. Under this definition, a person's language-assessment performance not only can be used to interpret that person's underlying traits, but can also reflect the influence of the specific context in which the interaction occurs. In this way, interactional competence can broaden the speaking construct, which is constituted by both communicative terms and interactional perspectives.

Specific language assessment practices. Understanding of the construct of interactional competence has been deepened by various researchers' investigations of specific assessment practices from both macro (Galaczi, 2004, 2008) and micro perspectives (Brooks, 2009; Ducasse & Brown, 2009; Galaczi, 2014; May, 2011; Wang, 2015).

The macro perspective. In peer-to-peer paired speaking tests, according to Galaczi (2004, 2008), there are three major types of interactional pattern – collaborative, parallel, and asymmetric – as well as a blended form that combines any two of the three core patterns. Based on mutuality and equality, the *collaborative pattern* refers to interactions with high mutuality and high equality. In tests, this means that all the participants co-construct the interaction, and sharing and expanding each other's ideas. The *parallel pattern* indicates interactions that are high-equality, yet exhibit a low level of mutuality. In test discourse, this would mean that all the speakers show a high degree of topic initiations, but fall short when it comes to expanding one another's topics. The *asymmetric pattern* pertains to interactions with low equality, and in which not all participants exhibit high mutuality. In test discourse, this is often manifested as an imbalance in the number of topic initiations, and/or in elaborations that are only contributed to by one or a few participants, while others occupy subordinate positions. Lastly, any two of the three patterns above may be combined at different points in an interaction: for instance, it could begin as parallel but end as asymmetric.

The micro perspective. A larger number of scholars has investigated interactional competence from a micro point of view, that is, through analyzing the interactional features discernible in test discourse. As briefly discussed above, on the theoretical foundation laid by Ducasse and Brown (2009), Wang (2015) built two categories in addition to the original three: interactive listening (supportive listening and comprehension), and interaction management (horizontal and vertical management). Based on the frequency of these interactional features' occurrence in previous empirical studies, Wang identified 17 principal interactional features and gave them the operational

Table 1

Interactional Features Selected and Defined by Wang (2015, p. 18)

Category	Subcategory	Interactional Feature	Definition
Interactive Listening	Signaling comprehension	Filling a silence	The action of suggesting or providing missing word(s) the other partner is searching for
		Making comments	Relevant statements indicating comprehension
		Agreeing/disagreeing	Agreeing or disagreeing with a partner
		Correcting a mistake	Correcting a partner's mistake or helping a partner out
	Signaling support	Back-channeling	Signaling attention while the other interlocutor maintains the floor
		Prompting	Actions to elicit or encourage a partner to elaborate
Interactional Management	Topic management	Initiation	Signaling the start of a new topic in a conversation
		Development	The actions of interlocutors in expanding a topic to develop a conversation or interaction
		Connection	Moves in which one interlocutor refers to the other's idea or topic to facilitate an interaction
	Turn-taking management	Turn length	Measured through mean utterance length
		Turn speed	How fast the two partners respond to each other
		Turn domination	How interlocutors compete for the floor
	Using questions	Agreement	Asking for agreement
		Confirmation	Asking for confirmation or checking comprehension
		Opinion	Asking for opinions
		Information	Asking for information
		Floor-offer	Offering the floor

definitions shown in Table 1. Other interactional features have also been summarized by various researchers, such as “expressing incomprehension”, “paraphrasing”, “topic closure”, and “asking for help” (Brooks, 2009).

To sum up, it would seem that, far from conflicting, theoretical models and the findings of studies on specific practices complement and enrich each other, enabling us to understand the construct of interactional competence in considerable depth.

Conversational Organization

In addition to a working understanding of the construct of interactional competence, anyone seeking to study L2 pragmatics in interaction requires some knowledge of conversational organization. Scholars’ understanding of conversational organization is influenced by Celce-Murcia’s (2007) communicative competence model and by CA, which studies the sequential organization of conversations as a means of accessing participants’ understandings of and collaboration in social interaction (Hutchby & Wooffitt, 1998). CA, in turn, was based on Garfinkel’s ethnomethodology and Goffman’s interaction analysis (Schiffrin, 1994), as well as – in the field of linguistics – Sacks, Schegloff and Jefferson’s (1974) work on the social organization of everyday interaction.

As briefly noted above, Celce-Murcia’s (2007) model of communicative competence added a subcomponent called interactional competence, which she further subdivided into actional, conversational, and paralinguistic competence. *Actional competence* refers to knowledge of “how to perform common speech acts and speech act sets in the target language involving interactions”: for instance, “information exchanges, interpersonal exchanges, expression of opinions and feelings, problems (complaining,

blaming, regretting, apologizing, etc.), future scenarios (hopes, goals, promises, predictions, etc.)" (p. 48). *Conversational competence* includes "how to open and close conversations", "how to establish and change topics, how to get, hold, and relinquish the floor", "how to interrupt", and "how to collaborate and backchannel, etc." (p. 48). Lastly, *paralinguistic competence* covers the use of "non-linguistic utterances" such as "silence and pauses" (p. 49).

Conversational dominance. Within the study of interactions and conversational organization, the topic of conversational dominance has received considerable attention. Linell, Gustavsson and Juvonen (1988) explored the distinguishing features of conversational dominance and divided it into three subtypes: *quantitative dominance*, referring to how much a person talks; *topical dominance*, which is related to the words and topics used when introducing new content; and *interactional dominance*, which relates to the quality of initiations and responses. Itakura (2001, p. 1861) proposed the quantification of asymmetries, the most systematic approach to date for investigating conversational dominance. Asymmetries are imbalances in participation power and in the display of interactional features. Itakura's application of this idea suggested that conversational dominance had three dimensions: sequential, participatory, and quantitative. *Sequential dominance* refers to a speaker controlling the direction of interactions via questions and the initiation of new topics. *Participatory dominance* consists of limiting others' speaking power by using interruptions and overlaps; and *quantitative dominance* relates to interlocutors' relative contributions to an interaction in terms of the numbers of words used.

Turn-taking organization. As an organized activity, turn-taking is the core concept of CA research (Lazaraton, 2002). As Sacks et al. (1974) observed, the basic fact about conversation is that only one person speaks at a time. Although speaker-change can be characterized by tiny overlaps as well as tiny gaps, the key attribute of a conversation is that its participants take turns, without any obvious gaps or overlaps. Sacks et al. also noted several ways to achieve speaker change: the next speaker can be selected by the previous speaker; the next speaker can choose him or herself; or the current speaker can continue to talk.

Sequence organization. The second core idea of CA is that talks in interaction are organized in sequences (Sacks, 1992), with each sequence creating a context for the next utterance. Sequences are seen as resources that can be employed to implement and respond to social actions such as invitations, praises, complaints, and agreements or disagreements. According to Wong and Waring (2010), studies of sequence have focused on mainly on simple sequences, although some more complex sequences have also been investigated. Adjacency pairs (APs) and response tokens belong to the former category, while preference organization and topic management belong to the latter.

Adjacency pairs. The basic building block of the sequence is an AP: two turns, taken by two speakers, ordered as a first pair-part (FPP) and a second pair-part (SPP), with particular types of FPP requiring specific types of SPP correspondence (Schegloff, 1968). Schegloff (2007) also noted three types of expansion in a sequence: *pre-expansion* (before FPP), which is designed to ensure that the speaker's actions will run smoothly; *insert expansion* (after the FPP and before the SPP), aimed at clarifying the FPP or obtaining preliminary information before producing the SPP; and *post-expansion*, which

can be an acknowledgment or assessment, and is intended to terminate the sequence. Careful attention to APs can reveal how speakers' mutual understandings are completed and manifested through interaction.

Response tokens. Another important set of sequences consists of response tokens, with which a listener gradually increases his/her participation in an interaction (Wong & Waring, 2010). They mainly include the following types: acknowledging previous information (*acknowledgments*); repeating or simply reorganizing the words of the previous speaker (*recycling*); offering evaluations of what has just been said (*assessments*); and giving signals before taking the floor (*listener speakership*).

Listeners' most basic activity in an interaction is generally the use of an acknowledgment or recycling token to display the listening-comprehension relationship: for example, *mm*, *hm*, or *okay* (Jefferson, 2002). However, a more convincing way to respond is to offer one's own assessment (Goodwin, 1986). Assessments can be either brief or extended. A typical assessment is to use turns to express agreement or disagreement, though minimal assessments, such as *great*, are frequently used as well. Right before listeners take the floor, some like to signal this via tokens such as *yeah* (Gardner, 2006; Jefferson, 1993).

Preference structure. In a third major type of sequence organization, known as preferences (Pomerantz,1984; Sacks, 1987), the speaker establishes the conversation in a way that suits the other party, and designs a turn that minimizes the threat of losing face by either of them (Sacks & Schegloff, 1979). There are several alternative, non-equivalent ways of designing first-pair and second-pair parts, some being preferred and others dispreferred (Pomerantz & Heritage, 2012), and turns can be packaged or shaped

to indicate that they are preferred or dispreferred. However, the preferred option appears natural, normal or as expected, and is selected whenever possible (Wong & Waring, 2010). It should be borne in mind that preferences are not personal preferences, but based on a sequence structure, and alternative choices of specific actions are usually preferred or dispreferred because of structural rules (Sacks, 1987; Schegloff, Jefferson, & Sacks, 1977; Lerner, 1996). For example, in first-pair parts, an offer is better than a request, since the former is good for others while the latter will cause them trouble. If a request is to be implemented, within the interlocutors' specific circumstances, it should not be made directly, and the sequence should be used to maximize the likelihood that the person receiving the request will accept it. According to Wong and Waring (2010), the key feature of preference structure in a specific context is usually an unmarked turn shape, such as no delay, mitigative devices, or accounts.

Preference organization does not constrain all adjacency pairs. For instance, when responding to many *wh*- questions, there is no need to concern oneself with preference structure (Wong & Waring, 2010). However, when second-pair parts respond to first-pair parts, preference structure becomes involved if there are alternative options such as agree/disagree and accept/reject (Schegloff & Lerner, 2009). In addition to requests and offers (Davidson, 1984), actions that involve preference structures include agreement/disagreement (Pomerantz, 1984), invitations (Davidson, 1984; ten Have, 2007), and compliments (Pomerantz, 1978).

Topic management. Conversational topics can be developed in a stepwise progression, or shifted abruptly. Topic management also belongs to the realm of sequence (Wong & Waring, 2010), and forms larger sequences, as briefly discussed above. Noting

the difficulties inherent in analyzing it, van Lier (1989, p. 147) remarked that topic management "has survived many years of non-definition". Similarly, Atkinson and Heritage (1984) wrote that "'topic' may well prove to be among the most complex conversational phenomena to be investigated and, correspondingly, the most recalcitrant to systematic analysis", and also pointed out that "topical maintenance and shift are extremely complex and subtle matters" (p. 165). Brown and Yule (1983) noted that it is difficult to draw a line between sentences, on the one hand, and on the other, the information between sentences.

Galaczi (2004, 2014) and Wong and Waring (2010) divide topic management into five categories: topic initiation; topic development; topic shift; topic termination; and topic incomplete. *Topic initiation* refers to the speaker introducing a new topic. There are different ways of doing this, including asking the other person a question, either generic or specific (Button & Casey, 1985); announcing new information about oneself, or news one is in possession of (Button & Casey, 1985); pre-topical sequences, used to recognize each other (Maynard & Zimmerman, 1984); and setting talk, related to the situation in which the conversation occurs (Maynard & Zimmerman, 1984).

Topic development (Galaczi, 2004; Wong & Waring, 2010) is related to the speaker's actions in developing the newly initiated topic, irrespective of who initiated it. Developing one's own topic can take two forms: pursuit and building. The former refers to situations in which, after the speaker initiated a new topic, he/she did not obtain the expected response, and thus continued to reiterate the topic's initially mentioned aspects. The latter refers to the speaker continuing to contribute new information to his or her own topics. Developing others' topics is also of two general types: minimal acknowledgment

and extension. The first refers to short replies such as *yes*, and the second to the current speaker contributing to topics that were previously initiated by the other speaker.

Topic shift (Wong & Waring, 2010) comprises talking about new aspects of the current topic or gradually shifting to new topics. It can be performed in two directions: using disjunctive markers such as *anyway* or *by the way*; or shifting in a stepwise fashion (Schegloff & Sacks, 1973). This stepwise approach can take three forms: (1) a pivot connecting the new aspect of the topic to the previous aspect; (2) a semantic relationship that is built between the current talk and previous talk; and (3) a summary of previous topics before moving on to new ones.

Topic termination (Wong & Waring, 2010) refers to the speaker having an intention to terminate a topic. This may be signaled by the use of pre-closing markers such as *well* or *okay* (Schegloff & Sacks, 1973), or the use of assessment tokens such as *great* or *very good* at topical boundaries (Antaki, 2002; Heritage, 1984; Waring, 2008; Wong & Waring, 2010). Lastly, *topic incomplete* (Galaczi, 2004) means that the speaker did not complete the topic, either spontaneously or because he/she was interrupted.

Employing the Discursive Approach

The current study's use of a combination of the discursive approach from L2 pragmatics studies with investigation of test-takers' in-test discourse is unprecedented, and is intended to provide a novel research perspective. As indicated by Kasper (2006, 2009), L2 pragmatics based on traditional theory neglects interaction; and for this reason, she proposed a new way of studying L2 pragmatics, namely discursive pragmatics, from the perspective of CA. In addition, McNamara, Hill and May (2002) have proposed that in-test discourse be studied qualitatively, for instance, using CA or DA.

In L2 pragmatics. Over the past three decades, various theoretical frameworks for L2 pragmatics have been created (Kasper, 2009). In this context, it should first be noted that L2 pragmatics has long been influenced by the theories and concepts of rational pragmatics: in particular, speech-act theory (Searle, 1969, 1975) and politeness theory (Brown & Levinson, 1987). The theoretical premise of these approaches is that speakers are individual rational actors, who choose their own means for meeting previous speakers' expectations after decoding/encoding information.

Operating under the traditional paradigm of speech-act research, early work on L2 pragmatics was based on cross-cultural pragmatics (Searle, 1969), and was thus essentially a comparison of different cultures rather than the study of pragmatics acquisition *per se* (Bardovi-Harlig, 1999; Kasper & Schmidt, 1996). And theoretically, it was dominated by the perspective of individual cognition. However, many researchers have since raised objections to the rational speech-act model's general concept as well as its data-collection and data-analysis methods, and a great deal of disagreement continues to swirl around issues of pragmatics in interaction (e.g., Youn, 2013).

Therefore, a variety of alternative theoretical approaches have been proposed to explain the interactions that are observed empirically. These alternatives originate from a variety of knowledge-bases and disciplinary foundations (D'Hondt, 2009), including ethnomethodology (Garfinkel, 1967), interactional order (Goffman, 1983), interactional sociolinguistics (Gumperz, 1982), and CA (Sacks et al., 1974). Among these, CA is widely considered an efficient and productive alternative to rational pragmatics. It is noteworthy that Kasper (2006), who originated the concept of "discursive pragmatics", advocated studying speech acts from the perspective of CA. Doing so would obviously

differ from the pragmatics derived under the principles of speech acts and politeness (Heritage, 1990; Schegloff, 1993; Searle, 1992). Speech-act and politeness studies focus on speakers' meanings and their strategies for achieving goals in interactions, whereas discursive pragmatics attends to how people complete the actions of daily life through interactions, and more specifically, to what speakers do through conversation rather than what they might have said. More and more research is employing the discursive approach in the study of L2 pragmatics, due to its detailed micro-analysis methods and its focus on the sequence of interactions (e.g., Galaczi, 2014, Youn, 2015).

In-test discourse. Coincidentally, McNamara et al. (2002) also pointed out that the most promising methods of speaking-assessment research to have been developed in the previous 15 years were qualitative ones, such as CA and discourse analysis. And indeed, qualitative methods have since been found effective in analyzing the validity of speaking assessments, since they can be used to explain how participants are able to construct pragmatic meanings and complete actions together in social interactions.

Research applying the discursive approach to analysis of L2 learners' interactional competence through speaking assessments has reported many interesting results, and demonstrated its value in the ongoing development of speaking-assessment techniques (e.g., Brown, 2006; Grabowski, 2009; Young & He, 1998; Ikeda, 1998; Ross, 1992; Swain, 2001; Young, 1995; Young & Milankov, 1992). The same approach has also contributed greatly to the conceptualization of interactional competence and how to operationalize it in assessment (e.g., Galaczi, 2004, 2008, 2014; Gan, 2010; Lazaraton, 2002; Young, 2008; Young & He, 1998).

The Operation of L2 Pragmatic Competence in Interaction

Over the past three decades, the body of research on L2 pragmatics assessment has gradually grown (e.g., Grabowski, 2009; Hudson, Detmer & Brown, 1992, 1995; Roever, 2006; Ross & Kasper, 2013; Walters, 2007, 2009). As noted earlier, previous pragmatics were mainly based on theories of individuals' speech acts (Searle, 1969, 1975) and politeness (Brown & Levinson, 1987), and ignored the role of interaction (Kasper & Ross, 2013; Youn, 2015). As such, previous research on L2 pragmatics assessment naturally ignored how to assess pragmatics when it involved interaction. Nevertheless, even as increasing attention is paid to speaking assessment of L2 pragmatic competence in interaction, only a few studies have done (e.g., Grabowski, 2009; Youn, 2013). This places pragmatics, as a test construct, at risk of validity challenges (Roever, 2011); and finding more effective means of assessing L2 pragmatics that involve interaction has emerged as an urgent new research direction.

Both discursive pragmatics (Kasper, 2006) and the knowledge of conversation organization derived from CA and from Celce-Murcia's (2007) communicative competence model have contributed to the definition and conceptualization of L2 pragmatic competence in interaction in test discourse. Turn-taking organization is the most basic component of conversations (Sacks et al., 1974), and thus logically should be the main topic of investigation in the assessment of L2 pragmatic competence in interaction. Test-takers should be examined as to whether they understand that the basic principle of conversation is that only one person speaks at a time, and that optimal turn-taking between speakers should feature no gaps and no overlap. Another perspective that

should be considered is which method the interlocutors use to select the next speaker (i.e., other-selection, self-selection, or the current speaker continuing to talk).

Sequence organization is another basic building-block of conversations, and mainly consists of adjacency pairs, response tokens, preference organization, and topic management. As mentioned earlier, the adjacency pair (paired turns of different speakers) is the basic unit that embodies intersubjectivity, and its turns should be relevant: e.g., greeting-greeting, ask-answer, offer-accept/reject or request-grant/reject. In any adjacency pair, the first-pair part plays the role of creating normative expectations for the action of the second-pair part, and serves as a basis for interpretation (Sacks et al., 1974). Therefore, when the second-pair part is missing, it is necessary to use L2 pragmatic competence in interaction to explain the reason behind.

Response tokens (Wong & Waring, 2010) can be used to evaluate whether test-takers are able to acknowledge that information provided by previous speakers has been received, and to repeat words used by those other speakers to indicate they have been listening. At a higher level of interaction, whether speakers comment on the previous speakers' statements and signal that they want to take the floor to be the next speaker.

Preference organization is also important for understanding the completion of actions in interactions. Some such actions are "positive" or "preferred", such as accepting invitations or expressing agreement, while others are "negative" or "dispreferred", such as rejecting invitations or disagreeing; and these differences are associated with clear differences in turn-taking structure (Pomerantz, 1984). Preferred actions typically cause overlaps, or occur without any delay between turns, whereas dispreferred actions lead to proper pauses and the use of hesitant markers, such as *well* and *uh*. The lack of these

normative features in interactions can jeopardize communication, so they are also components of L2 pragmatic competence in interaction.

In terms of topic management, test-takers can be examined on whether they can perform topic initiation, topic development, topic shifts, and topic termination (Galaczi, 2004, 2014; Wong & Waring, 2010), and what linguistic forms they will employ to accomplish these four functions.

Non-native speakers' use of sequence organizations varies along with their L2 skill levels. Al-Gahtani and Roever (2012) reported that, during a role-play task, most low-level learners omitted optional pre-requests (e.g., saying "Can you help me?") before the first pair containing a request (Schegloff, 2007), whereas advanced learners in the same context tended to use them. Non-native speakers' understanding of the pragmatic meanings of sequence organizations is also often limited. For instance, Walters (2009) found that, on a CA-based listening-comprehension test with various sequence organizations, non-native English speakers performed poorly compared to native ones, further indicating that learners' competence in sequence organization strictly constitutes L2 pragmatic competence in interaction.

However, when it comes to analyzing in-test discourse, the construct of L2 pragmatic competence in interaction needs to be clearly distinguishable from general concepts in the non-test environment. While in-test discourse, it is necessary to understand in great detail both *what* learners produce in interactions and *how* they produce it, including what strategies they use to initiate conversations, develop topics, provide audience feedback, and so forth.

Paired Speaking Tests

The model of communicative competence has drawn increasing scholarly attention since the 1980s, and has had a major influence on the definition of constructs in speaking assessments, prompting the emergence of paired speaking assessments. Such tests are considered capable of assessing interaction-relevant construct studies (Youn, 2015). Many scholars have pointed out that, as compared with OPI, paired speaking tests can elicit more symmetrical interactional patterns (e.g., Galaczi, 2008; Iwashita, 1998; Kormos, 1999; Lazaraton, 2002; Taylor, 2001), more diverse interactional features (e.g., Ducasse & Brown, 2009; Wang, 2015), and a wider range of language functions and roles (e.g., Galaczi, ffrench, Hubbard, & Green, 2011; Skehan, 2001). They also provide their participants with more opportunities to showcase their conversational skills (e.g., Brooks, 2009; O'Sullivan, Weir, & Saville, 2002), and provide better oral language sampling than other test types, such as OPI (Skehan, 2001).

From the findings of previous empirical research, it can be inferred that the speaking construct in language assessments is broader in paired speaking assessments than in OPI (Weir et al., 2013). Specifically, due to its models of lexico-grammatical accuracy and appropriateness, cohesion, organization, and fluency, it emphasizes interactive management features such as turn-taking management, topic initiation, and interactive listening (e.g., Ducasse & Brown, 2009; Galaczi, 2010; May, 2009).

Within the paired speaking test format, open role-play allows candidates to negotiate the interactive process without being instructed to achieve any specific interactional outcomes, and it has been shown capable of eliciting L2 pragmatic performance that is close to naturally occurring conversations (Youn, 2015). Because

"extended discourse" is considered a marker of interactional competence (Taguchi & Roever, 2017, p. 128), Galaczi (2014) chose to analyze a two-way collaborative discussion task belonging to *Extended discourse – part 3* of the University of Cambridge FCE. The two candidates did not have pre-assigned roles, and had to conduct two-way discussions and fully control the interaction. Galaczi argued that this task produced natural language output in which learners were likely to exhibit topic-management skills. Douglas and Selinker (1985) pointed out that such tasks allow candidates to participate more interactively, display more talk, and have more control over the language they use than they would in OPI. Riggenbach (1998) also asserted that tasks featuring greater flexibility in theme selection and interaction control might more truly reflect test-takers' interactive skills than less flexible tasks would.

Using paired speaking tests may encourage collaboration in classroom settings (Saville & Hargreaves, 1999; Taylor, 2000), but in formal-assessment contexts, it may cause both measurement and fairness problems, insofar as the lower-proficiency candidate can depress the test scores of both parties. To explore the influence of interlocutors' relative proficiency on paired speaking tests, Davis (2009) divided students into two groups – one with relatively high and the other with low English proficiency – and tested each person twice: once with a partner with similar proficiency, and once with a partner with higher or lower proficiency. The results indicated that the interlocutor's proficiency had no significant effect on the test's measurement ability; and most of the paired groups produced collaborative interactions (see also Galaczi, 2008). Overall, this suggests that concerns about differences in candidates' L2 skill levels should not be taken to outweigh the advantages of using the paired speaking test format.

The paired speaking test has also faced other challenges based on interlocutor effects, however (O'Sullivan, 2002). Factors mentioned in the literature as having a potential impact on test scores and/or in-test discourse include the participants' familiarity with each other (O'Sullivan, 2002), their gender (O'Sullivan, 2002), and their personality types (Ockey, 2009). Nevertheless, these factors are complex and often occur in mixed combinations, and no study to date has demonstrated the existence of a linear relationship between test-takers' scores and any of them.

Discourse Analysis

Definition. Though acknowledged as of the most important approaches to the study of discourse, DA is a blanket term for a variety of disparate methods developed for studying texts, based on various theories, and therefore can be hard to define (Gill, 2000; Silverman, 2006). Gill (2000) argued that scholars tend to hold one of four principal views of DA: (1) that it only concerns discourse itself; (2) that it considers language to be constructive and constructed; (3) that it emphasizes discourse as a form of function or action; and (4) that it is fundamentally rooted in the rhetorical organization of discourse.

Due to its potential usefulness in interpreting interactions in test discourse, the third of the above points of view is worthy of special attention here, due to its implication that discourse can be analyzed from the perspective of sociolinguistics or function, depending on whether one focuses on its "function orientation" or its "action orientation" (Gill, 2000; Gumperz, 1996). Discourse analysts view all discourse as social practice, and talk participated in by two or more people as being inherently interactive. A primary objective of DA is to explain the function or action of these cooperative talks.

Transcripts. In discourse studies, recording is a valuable tool to help

researchers capture useful but fleeting information, such as pauses and overlaps. By examining records of interactions, certain interesting aspects of discourse can be further investigated. However, recording by itself is far from sufficient for the systematic study of interactions. Thus, transcription of discourse is necessary, as it is helpful for retaining information that disappears quickly, and for organizing the disordered aspects of discourse.

For the most part, transcripts are not intended to be either exhaustive or objective, instead being both selective and interpretive in character. Decisions as to what to select while transcribing depend, to a large extent, on researchers' interests and theoretical background. Guided by their research questions, researchers make decisions about what information should be retained, what features should be analyzed, what approaches can be employed to search for such features, and what kinds of layout should be used to display information (Edwards, 2001; Gumperz & Berenz, 1993).

To some degree, transcripts in DA are less detailed than those in CA. However, this does not imply that the former are inferior. In reality, there is no perfect transcript (Edwards, 2001; Silverman, 2006). Noaks and Wincup (2004) noted that the degree of detail in a transcript is governed by many factors, including research questions and methods, and time- and resource limitations. The most important principle is to establish and abide by a rationale for choosing a particular style of transcription.

Data Exploration Strategy and Steps

The data-exploration strategy proposed by ten Have (2007) is a circulatory system for the processing of transcriptions of conversational data, focusing on basic concepts of conversational organization such as turn-taking, sequence, and repair.

Pomerantz and Fehr (1997), meanwhile, proposed steps for making data exploration systematic and comprehensive, as follows: (1) selection of a sequence; (2) recognition of the type of action implemented in the sequence (e.g., topic development methods, types of response tokens); (3) consideration of the form of the speaker's action (e.g., using statements or questions), and when and how the turn-taking is processed (e.g., after pauses, overlaps/latches, or interruptions); and (4) based on previous analysis, thinking about the speaker's role in interaction (e.g., collaborative, non-cooperative, dominant, or passive).

One of the crucial premises of DA is that it should not be driven by pre-existing theories or hypotheses, but instead should describe conversational organization via an "emic" perspective (e.g., insider's view) (Hutchby & Wooffitt 1998). Therefore, while points of analytical interest can be known in advance, a DA study's conclusions should arise entirely from the data.

Galaczi's (2004, 2008, 2014) proposals for topic analysis deemed "topic sequence" the most appropriate analysis unit, on the grounds that topics and topic management are very complicated and difficult to analyze. Brown and Yule (1983) had previously argued that attempts to identify topics are doomed to failure, due to their abstract nature and the difficulty of determining the boundaries between one set of information and another. Indeed, conversations can gradually progress from one topic to another in a manner that the interlocutors are not conscious of (Button, 1991; Button & Casey, 1984; Jefferson, 1984), making it difficult to determine not only what the topic of a given conversation is, but also how and if that topic can be separated from other ones (Button & Casey, 1984). Galaczi (2004) used the prompts of a topic-discussion task in a

paired speaking test as the basis for determining the topic, as follows. A sequence was deemed to begin with the discussion related to a prompt, and all developments associated with that prompt were classified as part of that topic sequence. Each transcription was divided into discontinuous topic sequences, indicating that topic shifts were based on prompts. This approach allowed Galaczi to conduct systematic and consistent topic analysis across various research projects and purposes.

Based on ten Have's (2007) strategy and Pomerantz and Fehr's (1997) steps, in addition to identifying a conversation's topic sequence and the steps within that sequence, Galaczi (2004) analyzed the forms of speakers' action, because there are multiple ways to complete certain actions, and choosing one over another is always meaningful. Specifically, the foci of analysis include the manner of turn-taking (self-selection or other selection) and its timing (after gaps or with overlaps/latches), and the termination method (voluntary or mandatory). Finally, based on all of the above analysis, the speaker's role in interaction should also be analyzed, and classified as collaborative, non-cooperative, dominant, or passive according to that person's contributions to the interaction.

Mixed Methods Design

There has been a long-term dispute between the supporters of quantitative and qualitative methodologies, manifested not least in the so-called Paradigm War of the 1970s and 1980s (Gage, 1989). Although there are a large number of differences between the two methodologies, they share some similarities as well. As Johnson and Onwuegbuzie (2004) pointed out, researchers formulate research questions based on observations, and make every effort to reduce bias and other potential sources of invalidity, regardless of what methodologies they employ. Based on these commonalities,

the two methodologies were combined into a new independent research methodology, namely MMR. Johnson, Onwuegbuzie, and Turner (2007) defined this approach as

> the type of research in which a researcher or team of researchers combines elements of qualitative and quantitative research approaches (e.g., use of qualitative and quantitative viewpoints, data collection, analysis, inference techniques) for the broad purposes of breadth and depth of understanding and corroboration[.] (p. 123).

MMR has been controversial since it came into being. One central point that its opponents have made is that the two research paradigms that comprise it have distinct overall consistencies, which are different both from worldviews and from inference methods (Guba & Lincoln, 1989). Faced with such challenges, some of the scholars using MMR (Johnson et al., 2007; Teddlie & Tashakkori, 2003) adopted the pragmatic stance that gathering evidence for answers to one's research questions as efficiently as possible makes more sense than simply focusing on the supposed incompatibility between two paradigms.

However, as Brown (2014) pointed out, the fact that a piece of research is neither purely qualitative nor purely quantitative does not mean that it must be MMR. Rather, to qualify as MMR, qualitative and quantitative methods must be used systematically and complementarily, that is, to balance out each other's weaknesses. If the two methods are simply used simultaneously or sequentially, without any such interaction between them, it would be better to refer to the study in question as multi-method. Brown (2014, p. 134) also proposed techniques for improving the legitimacy of MMR, including techniques of "convergence", "divergence", "elaboration",

"clarification", "exemplification" and "interaction". Among these, convergence and divergence are the most frequently used. The former refers to disparate sources of data coming together to support similar conclusions, while the latter refers to conflicts between data sources that can lead to more in-depth exploration. In the language-assessment field, many successful empirical studies have utilized MMR and demonstrated its feasibility and applicability (e.g., Grabowski, 2009; Jang, 2005; Lee & Greene, 2007; Norris, 2008; Walter, 2007; Wang, 2015; Youn, 2013).

Summary

This chapter has presented the literature relevant to the current research. Specifically, its first part covered the construct of L2 pragmatic competence in interaction. First, from the perspective of theoretical studies of SLA, it introduced the construct of interactional competence, including its key features and how it differs from communicative competence, and summarized the interactional patterns (a macro point of view) and features (a micro point of view) identified by prior research on specific language-assessment practices. Second, it summarized the current state of knowledge of conversational organization, including conversational dominance, turn-taking organization, and sequence organization (adjacency pairs, response tokens, preference structure, and topic management). Third, it introduced a new approach for investigating L2 pragmatics and in-test discourse: the discursive approach. Finally, it covered how L2 pragmatic competence in interaction has been conceptualized in assessment studies.

The second part of this chapter discussed the methodology employed in the current research. First, it clarified why the paired speaking test format is best suited to eliciting the in-test discourse most suitable to this study's aims. Then, it illustrated how

the DA approach has been used in analyzing such discourse, including the definitions of DA and how data is transcribed for further analysis. After that, the present study's data exploration strategy and steps were set forth; and last but not least, the rationale for and benefits of using mixed methods throughout the study were explained.

CHAPTER 3

METHOD

The critically important foundation of this study's demonstration of the development trajectory of Chinese L2 pragmatic competence through in-test discourse is the design of a speaking test capable of generating rich data. This chapter introduces the specifics of the mixed-methods approach used for eliciting and analyzing such data. Prior research (Johnson et al., 2007; Teddlie & Tashakkori, 2003) has indicated that mixed methods would be ideal for the collection of evidence suitable to answering the present study's research questions.

Participants

The participants in this study comprised 90 test-takers and two raters.

Examinees. A total of 90 adult Chinese learners studying in five Chinese universities took part in the study voluntarily. As shown in Table 2, 54.4% were female and 45.6% were male. Their ages ranged from 18 to 38, and they had 22 different native-language backgrounds: Korean (21.1%), Russian (11.1%), Arabic (10%), Indonesian (10%), Vietnamese (9%), Thai (8.9%), English (3.3%), Mongolian (3.3%), Persian (3.3%), Armenian (2.2%), French (2.2%), Japanese (2.2%), Kazak (2.2%), Portuguese (2.2%), Turkish (2.2%), Bengali (1.1%), German (1.1%), Lao (1.1%), Latvian (1.1%), Nepali (1.1%), Tajik (1.1%) and Uzbek (1.1%). The time they had spent living in China ranged from 1 month to 12 years, with a mean of 24 months and a median of 17 months. Their years of learning Chinese ranged from 4 months to 14 years, with a mean of 41 months and a median of 30 months. Not counting exchange students (who made up 17.8% of the sample), the majority of the test-takers, 55.6%, were either undergraduates (35.6%)

Table 2

Test-takers' Background Information

Number	90		
Gender	Female 54.4% Male 45.6%		
Age	18-38		
L1	>7%	2%-7%	<2%
	Korean 21.1% Russian 11.1% Arabic 10% Indonesian 10% Vietnamese 9% Thai 8.9%	English 3.3% Mongolian 3.3% Persian 3.3% Armenian 2.2% French 2.2% Japanese 2.2% Kazakh 2.2% Portuguese 2.2% Turkish 2.2%	Bengali 1.1% German 1.1% Lao 1.1% Latvian 1.1% Nepali 1.1% Tajik 1.1% Uzbek 1.1%
Time living in China	1 month to 12 years		
Time learning Chinese	4 months to 14 years		
Program in China	Undergraduate study: 35.6% Master's study: 21.1% Preparatory program for undergraduate study: 20% Exchange program: 17.8% Ph.D. study: 5.6%		

or taking preparatory classes for undergraduate study (20%). The remainder were in master's (21.1%) or Ph.D. programs (5.6%).

Most of the participants had not previously participated in a standardized test of their Chinese-language abilities. Therefore, their Chinese proficiency levels were mainly assessed by their Chinese instructors. Since the interactive task in this study's paired speaking test required both of its participants to have similar language proficiency levels, the students were sorted into three groups – low, middle, and high proficiency – according to information provided by the instructors. This process resulted in 28 students being placed in in the low-proficiency group, 34 in the middle-proficiency group, and 28

in the high-proficiency group. To further evaluate candidates' Chinese-language proficiency levels, this study devised three independent speaking tasks, all based on the individual test format of the Common European Framework of Reference for Languages (CEFR), and administered them to all participants prior to their completion of the interactive tasks.

Raters. The raters for this study's paired tasks and its independent oral-proficiency tasks were two female native speakers of Chinese. One of them, who had 6.5 years of Chinese-language teaching experience, has held a Ph.D. degree in Chinese linguistics and language from a university in the United States, and now is teaching at another American university. The other rater, a doctoral student in Chinese linguistics and language studies, had been teaching the Chinese language for 4 years.

Instruments

The main instruments used in the current study included the background questionnaire, the test instruments, and the rating criteria.

Background questionnaire. Before being tested, all the participants were asked to fill out a background questionnaire (see Appendix A), aimed at capturing their Chinese learning and testing experience as well as their demographic details.

Test instruments: The test in this study is divided into two parts: solo tasks and interactive tasks. The design of the two parts of the test content was based on a needs analysis.

Needs analysis*:* As noted previously, according to the CEFR (2001), language learners' foreign-language use can be divided into four domains: personal, public, educational, and occupational, of which the first two are more difficult to delineate than

the latter two. A great many daily communications fall into the personal language-use domain, which is crucial to L2 Chinese students, especially those who are studying and living in China, but its boundary remains indistinct. Thus, an open-ended questionnaire (see Appendix B) about international students' language-use needs in the personal domain was administered to both Chinese-language teachers and international students who belonged to the target population. Background information on the participants in the needs analysis is shown in Table 3.

Table 3
Background Information, Participants in the Needs Analysis Questionnaire

Teachers		International students	
Number	14	Number	12
Gender	Female 64% Male 36%	Gender	Female 75% Male 25%
Age	23-40	Age	20-28
Teaching Experience	0.5 to 12 years	L1	Kazakh (41.7%) Mongolian (16.7%) French (8.3%) Spanish (8.3%) Thai (8.3%) Lao (8.3%) Montenegrin (8.3%)
		Current courses taken	Intermediate to advanced level
		HSK[1]	Level 4 to 6

The three major themes that could be discerned from their responses were personal relationships, frequently used language functions, and locations, as illustrated in detail in Table 4.

[1] *Hanyu Shuiping Kaoshi* (Chinese Proficiency Test): a standard instrument for measuring the Chinese-language proficiency of non-native speakers in China. Levels 4 to 6 of the HSK are equivalent to levels A2 to C1 of the CEFR (Lu, 2017).

Table 4

Three Themes of the Personal Language-use Domain

Theme	Detail
Relationships	• Friends • Strangers • Family members • Other social relations (e.g., teacher-student; co-workers)
Frequently Used Language /Functions	• Exchanging ideas or engaging in discussions (e.g., casual chatting, topic discussion, expression of emotions) • Solving problems (e.g., asking for help, handling conflicts) • Practicing specific speech acts (e.g., making plans to go out together; inviting someone to a party)
Locations	• At a social event • In a place of entertainment/recreation (e.g., a shopping mall) • On a trip • On campus • Online • Other social sites (e.g., teacher's office)

A list of the personal language-use situations, topics and speech acts most commonly mentioned in the needs analysis is presented in Table 5. It will be noted from this table that conversational topics varied considerably, depending on whether the respondent's interlocutor was a friend or a stranger.

According to the results of the needs analysis, personal language use domain was delimited from the following three themes: relationships, frequently used language/functions, and locations. Based on this scope, the commonly used situations, topics (with friends and strangers) and the speech acts were also summed up. The two parts of the speaking test (see Appendix C) were designed based on the three themes and the three common used aspects.

Table 5

The Most Commonly Mentioned Situations, Topics and Speech Acts

Category	Detail
Situations	• Basic daily casual chatting • Informal discussions • Inviting friends to go out • Organizing or participating in activities • Asking for help • Chatting online
Topics with Friends	• Coping with life in China (e.g., national and cultural differences, eating or shopping habits, means of transportation, environmental issues) • Interests and hobbies (e.g., traveling; eating out) • Work and study (e.g., educational differences; learning Chinese) • Personal feelings (e.g., impressions of China; other friends and acquaintances) • Philosophy of life (e.g., dreams; goals) • Love life (e.g., love stories; dating problem) • News (e.g., politics; entertainment; gossip)
Topics with Strangers	• Basic personal information (e.g., name; nationality) • Interests and hobbies • Work and study • Personal feelings • Sharing of experiences (e.g., eating; shopping) • Asking for help (e.g., asking for directions; borrowing something)
Speech Acts	• Invitation • Request • Inquiry/answer • Apology • Greeting • Agreement/disagreement

Solo tasks. These three 1-minute solo tasks were used to explore the relationships between the candidates' L2 pragmatic competence in interaction and their L2 language proficiency levels. The language-function foci, situations, and topics for these tasks are summarized in Table 6, below.

Table 6

Structure of the Three Independent Speaking-proficiency Tasks

Instructional language and approach	Chinese characters Pinyin Romanization Sound recording Picture prompts
Language-function foci	Task 1: Providing descriptions and expressing opinions Task 2: Comparing and contrasting Task 3: Speculating and imagining
Topic	Task 1: A place one has traveled to Task 2: Comparison of eating habits in different two countries/places Task 3: Imagining you are a teacher
Timing	1 minute per task

Paired Interactive tasks. This part of the test also emerged from the scope of the personal language-use domain. The performance observed in the assessment tasks needed to reflect use of the target language in real life if it was to generate meaningful scores. Thus, to connect candidates' performance on speaking tasks to this target domain, the tasks had to reflect the competence required to cope with representative real-life situations.

To be able to elicit the discourse closer to natural occurring conversations, two task types were chosen: open role-play and situational topic discussion tasks. There are two open role-play tasks and one situational topic discussion task in this part of the test. Commonly occurring situations were used to develop the content of the three tasks.

In the open role-play tasks, candidates at all three proficiency levels were given the same tasks. Unlike in closed role-play tasks, the test-takers had no fixed interactive objectives in the open role-play tasks; the purpose of this aspect of the design was to elicit more natural interactions. However, if students are to be assessed using a uniform standard, it is reasonable to expect that test tasks will be standardized in terms of their

Table 7

Structure of the Three Interactive Tasks Assigned to the Paired Test-takers

	Open Role-play	Situational Topic Discussion
Instructional language and approach	Chinese characters Pinyin Romanization Sound recording Picture prompts	
Number	2	1
Same task to all proficiency levels?	Yes	No
Language-function focus	General: • Maintaining communication • Achieving situational communicative goal • Using speech acts • Evaluating Specific: Task 4: Agreement/disagreement and offering suggestions Task 5: Invitation, request and apology	General: • Maintaining communication • Exchanging opinions • Explaining and justifying reasons • Agreeing or disagreeing • Reaching an agreement through negotiation Specific: Task 6.1 Expressing opinions Task 6.2 Comparing and contrasting Task 6.3 Constructing hypotheses
Situation	Task 4: Getting to know each other at an on-campus Chinese event Task 5: Inviting a friend to a party and borrowing a coffee machine from him/her after Chinese class	Tasks 6.1 to 6.3: The partners became friends after meeting at the event. They are talking while waiting at the bus stop on the way to go grocery shopping together.
Topic	Task 4: • Personal information • Impressions of China Task 5: • Inviting friend to a party • Asking for help	Task 6.1: Interests and hobbies Task 6.2: National differences Task 6.3: Urban livability
Timing	No time limit	

language-function foci, situations and topics; and the design allowed for this, as indicated in Table 7. In the two open-role play tasks, the two candidates will represent role A and role B, and they will be given different situations and not informed of what the other person's situation is.

In the situational topic discussion task, in contrast, the tasks assigned to the three proficiency levels of candidates differed. The low-level group discussed their interests and hobbies (e.g., reading books, watching movies, playing videogames, exercising, traveling); the mid-level group compared differences and similarities between countries (e.g., shopping habits, means of transportation, recreation/entertainment, environmental issues, educational modes); and the high-level group discussed issues around urban livability (e.g., pollution, social security, economic development, friendliness of residents). It will be noted from this that the topics discussed were increasingly difficult and abstract as one moved up the proficiency ladder. The purpose of this was to allow students of all ability levels to talk about topics that elicited their natural language to the greatest degree.

Rating Criteria

Two rating rubrics were used: one to score the three solo tasks and the other to assess the three interactive tasks.

For the solo tasks. The present study's rubric for assessing the solo tasks retained four of the CEFR's rating categories, that is, *range*, *accuracy*, *fluency*, and *coherence* (see Appendix D) according to the needs of current study. Since these tasks were not the research focus of this study, the original CEFR scoring system was also

streamlined, with raters only providing an overall score for each task on a three-point scale, rather than breaking their scores down into categories.

For the paired interactive tasks. An analytical rating rubric (Brown, 2012) (see Appendix D) was used to evaluate paired interactive tasks. Based on the theories of interactional competence and conversational organization, coupled with the findings of prior studies on the assessment of L2 interactions (Galaczi, 2014; Youn, 2013), the researcher designed an analytical rating rubric to measure L2 pragmatic competence in interaction. This rubric includes five categories: (1) *language use*, (2) *situation response*, (3) *turn-taking organization*, (4) *sequence organization* and (5) *topic management*. These categories are further grouped into three levels, according to their competence levels. Raters were required to rate each category, since L2 pragmatic competence in interaction – being a new assessment construct – should be assessed in light of the most detailed possible information on the test-taker's performance.

Procedure

After the speaking test was designed, to test its validity, a small-scale pilot study was conducted. Based on the findings of the pilot study, the original test was modified. After that, the real test of this study was started. After the test was completed, the researchers found two raters and trained them for the next step – rating. Then they began to rate the test independently and participated in the online interview after the rating was completed.

Pilot study. Three pairs of students – one from each of the low, middle, and high language-proficiency ranges – participated in a pilot study. The two students in the low-level pair were from a third-year Chinese-language class. The middle-level students

were from a fourth-year Chinese-language class, and had experience of short-term study abroad in China. One of the two students in the advanced pair was taking the same fourth-year Chinese-language course mentioned above, and had previously done missionary work in China, while the other was a graduate student in Chinese and had studied in China for many years.

Based on these six test-takers' performance, the pilot study indicated that the main instrument's goals had only been partially achieved. After finishing the paired speaking test, all six participants reflected that the situations, topics and language functions of the tasks were moderately difficult at their respective proficiency levels, and very frequently encountered in real life, based on their language-learning experience. As such, they felt the tasks could elicit a high number interactions, and was good speaking practice. However, transcription and analysis of the in-test discourse of the three pairs in the first role-play and topic-discussion tasks revealed that the low-proficiency participants had difficulty in comprehending the speaking tasks due to their relatively low level of knowledge of Chinese characters. Thus, the *pinyin* Romanization system for Chinese was added to the instructional aids (see Appendix C). It was also found that the turns elicited by the original topic-discussion and decision-making tasks were extremely long, and the frequency of turn-taking was low: salient features of formal discussions, as distinct from everyday casual conversation, which should feature short turns and frequent turn-taking. Thus, a new situation was added to the original discussion task: the two speakers, who had become friends at a Chinese-themed weekly campus event called "Chinese corner" and frequently spend time with each other, are talking naturally while waiting at a bus stop on the way to go grocery shopping together. This familiar situation

was provided so that the participants could discuss some day-to-day topics naturally and informally, as befits the personal language-use domain being tested.

Test administration. Before the administration of any of the tests, the participants were divided into the three proficiency levels discussed above, based on their instructors' assessments, and two students were paired within each level. Both members of each pair arrived at the testing site at the same time, and each of them completed their solo tasks before proceeding to their shared interactive tasks.

The researcher explained the tasks in detail to ensure that all the participants understood the test process and the meaning and requirements of each task. Including the provision of these instructions, and the solo tasks, the total test time for each pair was approximately one hour. The whole process was audio recorded.

Rater training. Before conducting the rater training, the researcher made it clear to the raters that the test-takers' recordings were to be treated as confidential. Each rater was also asked to provide her background information, including educational attainment and teaching experience.

The rater training was performed in two sessions: the first before rating the solo tasks, and the second before rating the paired tasks. Each session continued for approximately one hour, and its goal was to familiarize each rater with the two rating rubrics and how to implement them. The two rating rubrics were shared via Google Drive and relevant instructions and information about the rating process were conducted through a teleconference, after which the researcher sought a consensus regarding the rubrics, focusing on areas where the three parties held differing opinions. The researcher then provided each rater with the same representative language samples at each

proficiency level, and asked them to rate them using the rubrics. After they had finished rating these samples, offline and independently of each other, they shared their views and experience via another teleconferencing session. Then, each used a minimum of three more recordings to practice further, and were allowed to ask the researcher questions at any time. This process continued until the two raters achieved a high degree of agreement in rating.

Rating. Each rater was sent recordings to be rated in batches via Google Drive, as well as Excel grading forms for each individual test-taker. Both raters were also required to record all the scores on their own forms. During the scoring process, rating was conducted independently and without any discussion between the raters. To ensure data security, once a rater had finished rating one task, she was not allowed to access that batch of data again. The rating of interactive tasks was based on the degree of openness from low to high: open role-play task 5 (inviting a friend to a party and borrowing a coffee machine), open role-play task 4 (getting to know each other at "Chinese corner") and situational topic discussion (casual talk while waiting at the bus stop). As the degree of openness increases, the difficulty degree of rating increases. Thus the raters first rate the less open tasks, and then rate the more open tasks after they become more familiar with the analytical rating rubric.

Raters' online interviews. After the rating process had been completed, a brief online interview was conducted with each rater. These interviews, along with the notes the raters took during the rating process, provide important evidence regarding their understanding of the rating rubrics. The online interviews included the following questions:

1. What is your opinion of the rater-training process, especially in terms of its clarity and effectiveness?

2. How did you ensure that your scores properly reflected the right test-takers' performance in the paired speaking tasks?

3. What difficulties did you experience during the rating process, and do you have any suggestions for modifying the rating rubric or the rating process?

Data Analysis

Table 8 presents supporting analyses for each of this study's research questions.

Quantitative analysis. Descriptive statistics were first calculated to confirm whether the test data were normally distributed. Central tendency, distribution, and dispersion were examined using the mean, standard deviation, minimum, maximum, skewness, and kurtosis of each score. Repeat measures ANOVA were conducted twice to see whether there are interaction effects between level and task, and level and rating category. The means of the solo-task scores and language components were also calculated for the three proficiency levels, to reveal how the test's difficulty differed across the participants' levels of competence.

Classical test theory (CTT) was used to answer research question 2, regarding the reliability and consistency of ratings. The Spearman-Brown prophecy formula (Brown, 2012) was used to calculate the inter-rater reliability of the two raters, including the reliability of all tasks, and the rating reliability of each category in the rubrics. CTT was also used to examine the internal consistency reliability of individual tasks and the entire test, which were calculated using Cronbach's alpha (Brown, 2012), to establish the extent to which different categories in the rubrics measured the same construct together.

Table 8

Supporting Analyses for Each Research Question

Research questions	Literature review	Quantitative	Qualitative
1. How effectively do the three paired speaking tasks developed in this study reflect Chinese learners' L2 pragmatic competence in interaction? To what extent do these tasks strike a balance between standardization and authenticity?	Paired speaking test		Opened-ended questionnaire, needs analysis of the personal language-use domain
2. When using an analytical rubric with interactional features, to what extent can raters ensure the reliability and consistency of their rating?		Classical test theory (inter-rater reliability and internal consistency reliability); Pearson correlation analysis	Raters' perspectives from their online interviews and their rating notes
3. What features useful for distinguishing between varied levels and tasks are identifiable in test-takers' paired test discourse? How much can those distinguishing interactional features deepen our understanding of the developmental trajectory of Chinese learners' L2 pragmatic competence in interaction?	Interactional competence; conversational organization	Descriptive statistics	Discourse analysis
4. To what extent are the findings from mixed methods design reliable and how can they enhance the validity of the future assessment of Chinese learners' L2 pragmatic competence in interaction?		Classical test theory (inter-rater reliability and internal consistency reliability)	Discourse analysis

In addition, Pearson correlation analysis was used to assess how much the test-takers' L2 pragmatic competence in interaction was related to their language proficiency levels as assessed by the three solo speaking-proficiency tasks.

Qualitative analysis. DA was used to explore in detail the quality of in-test discourse data, in light of prior work on conversational organization (e.g., Sacks et al., 1974) and Celce-Murcia's (2007) communicative competence model. Using an adapted form of the symbol system developed by Gail Jefferson (Atkinson & Heritage, 1984), the paired speaking test discourse was carefully transcribed for sequential analysis (for the transcription conventions, see Appendix E). As mentioned earlier, a large number of studies have shown that discursive approaches such as DA are the most effective for analyzing interactions (McNamara et al., 2002). The specific steps used in the present study followed ten Have's (2007) "data exploration strategy", Pomerantz and Fehr's (1997) "systematic steps" and Galaczi's (2004, 2008, 2014) "topic analysis".

Summary

This chapter has summarized the current study's methods of data extraction and data analysis; its participants' characteristics; its instruments, including the background questionnaire, the three solo proficiency tasks, the paired interactive tasks, and the rating criteria for all four tests; the specific procedure used, including a pilot study, rater training, test administration, rating process, and raters' online interviews. It has also outlined the analytical methods utilized to answer each research question. The next chapter will summarize the results of both quantitative and qualitative analysis.

CHAPTER 4

RESULTS

This chapter first presents the results of the quantitative analysis. Descriptive statistics, inter-rater reliability and internal consistency reliability, repeated measures ANOVA results and Pearson correlations are discussed.

Second, the chapter presents the results of the qualitative analysis. These results are related to the analysis of international students at the college level in China and Chinese language teachers' language learning and teaching needs in the personal language use domain, the analysis of raters' views towards rating, and DA of test performance discourse.

Quantitative Analysis

Descriptive statistics. In order to gather information on the distributions of measured variables, preliminary analysis was conducted. Table 9 lists the descriptive statistics for the individual scores for the five categories of the analytical rating rubric within each task of the three paired speaking tasks. This table also lists the total scores for all the three tasks together.

The average score of the five categories for each individual task ranged from 2.01 (topic management for task 2: inviting a friend to a Christmas party) to 2.58 (situation response for task 3: situational topic discussion). The standard deviation (*SD*) ranged from 0.33 (situation Response for task 1: knowing each other in a Chinese corner) to 0.58 (topic management for task 2: inviting a friend to a Christmas party). The lowest score for each category was 1 and the highest score was 3. Values more than twice the

standard error of skewness (*ses*) are probably skewed to a significant degree (Brown, 1997). Thus the acceptable range of skewness values of this study were from -0.52 to 0.52, indicating the categories of "language use" and "turn-taking organization" were

Table 9
Descriptive Statistics for the Rating Categories

Category	Task	*N*	Mean	*SD*	Min.	Max.	Skewness	Kurtosis
Language use	1	90	2.38	0.45	1.50	3.00	-0.17	-1.16
	2	90	2.23	0.52	1.00	3.00	-0.18	-0.70
	3	90	2.34	0.48	1.25	3.00	-0.31	-0.77
	All 3	90	2.32	0.44	1.25	3.00	-0.25	-0.72
Situation response	1	90	2.56	0.33	1.75	3.00	-0.25	-0.94
	2	90	2.46	0.55	1.00	3.00	-1.08	0.77
	3	90	2.58	0.36	1.75	3.00	-0.45	-0.59
	All 3	90	2.53	0.29	1.58	3.00	-0.60	0.55
Turn-taking organization	1	90	2.38	0.50	1.25	3.00	-0.45	-0.68
	2	90	2.33	0.47	1.25	3.00	-0.32	-0.83
	3	90	2.40	0.51	1.50	3.00	-0.35	-1.16
	All 3	90	2.37	0.43	1.50	3.00	-0.36	-0.93
Sequence organization	1	90	2.50	0.44	1.25	3.00	-0.53	-0.47
	2	90	2.16	0.54	1.00	3.00	-0.32	-0.43
	3	90	2.48	0.44	1.50	3.00	-0.25	-1.28
	All 3	90	2.38	0.39	1.42	3.00	-0.34	-0.84
Topic management	1	90	2.31	0.48	1.50	3.00	-0.33	-0.99
	2	90	2.01	0.58	1.00	3.00	-0.09	-0.97
	3	90	2.41	0.48	1.25	3.00	-0.63	-0.34
	All 3	90	2.24	0.44	1.25	3.00	-0.29	-0.72

normally distributed in terms of skewness, while task 2 of "situation response" (-1.08), task 1 of "sequence organization" (-0.53), task 2 of "topic management" (-0.63) were negatively skewed to a significant degree. Values more than twice the standard error of kurtosis (*sek*) are probably different from mesokurtic to a significant degree. (Brown, 1997). Thus the acceptable range of skewness values of this study were from -1.04 to 1.04, indicating the categories of "situation response" and "topic management" were normally distributed in terms of kurtosis, while the task 1 of "language use" (-1.16), task

3 of "turn-taking organization"(-1.16) and task 3 of "sequence organization" (-1.28) were significantly non normal in terms of kurtosis.

For the three tasks combined, the overall means of the five rating categories ranged from 2.24 (topic management) to 2.53 (situation response). The standard deviation ranged from 0.29 (situation response) to 0.44 (language use & topic management). The lowest score for the three tasks together was 1.25 and the highest score was 3. Based on the acceptable range of skewness value mentioned above (-0.52, +0.52), except the category of "situation response" (-0.60), the distribution of all scores combined under the other four rating categories was normal. And according to the acceptable range of kurtosis value mentioned above (-1.04, +1.04), the entire distribution of all scores combined under the five rating categories was normal.

In order to investigate the development of the three tasks across three different levels (low, middle, and high-level), the means of the three tasks across three different levels was calculated. As shown in Table 10, the means of these three tasks increased

Table 10
Task Means across Levels

	Low	Mid	High	All levels
Task 1-role play 1	1.88	2.47	2.80	2.39
Task 2-role play 2	1.82	2.18	2.68	2.24
Task 3-discussion	1.90	2.48	2.82	2.44

with the level of test taker's L2 pragmatic competence in interaction. In other words, the average score of each task in the high-level group was higher than that in the mid-level group, and the score in the mid-level group was higher than that in the low-level group. While Task 2 (invite a friend to a Christmas party) had the lowest mean score, Task 3 (situation topic discussion) had the highest mean score.

The means of the five rating categories in the analytical rating rubric across the three different levels (low, mid, high-level) (listed in Table 11) was calculated as well.

Table 11
The Means of Rating Categories across Levels

	Low	Mid	High	All levels
Language use	1.81	2.33	2.77	2.32
Situation response	2.27	2.54	2.78	2.53
Turn-taking organization	1.87	2.42	2.79	2.37
Sequence organization	1.95	2.39	2.78	2.38
Topic management	1.74	2.27	2.69	2.24

The average level of each category in the analytical rating rubric increased with the level of test taker's L2 pragmatic competence in interaction. These values indicated that the average score of the high-level group in these five categories was higher than that of the mid-level group, while the average score of the mid-level group was higher than that of the low-level group. The three tasks were combined, the category of the "situation response" had the highest average score, and the category of "topic management" was the lowest. In the high-level group, it displayed that candidates' "topic management", "language use", and "turn-taking organization" are all noticeably lower. The categories of "turn-taking organization" and "sequence organization" showing the interactional features were higher than the category of "language use". The categories of "turn-taking" and "sequence-organization" of the mid-level also scored for higher than the category of "language use". The highest score in the high-level group is the "turn-taking organization", but it could be found that the five categories scored very similarly.

Repeated measures ANOVA. Repeated measures ANOVA was performed twice. For the first time, level was treated as one factor and task as the other repeated-

measure factor. For the second time, level was also treated as one factor and rating category as the other repeated-measure factor.

The sample size was 90, which is relatively large. Before running the repeated measures ANOVA, the assumptions were checked for both sets of data. The Q-Q plots indicated that the variables deviated slightly from normality. ANOVA assumes that the data is normally and independently distributed. However, research shows that ANOVA is robust to moderate deviations in normality (Glass, 1972). The data were also checked for univariate outliers and multivariate outliers. No outlier was found. Then Mauchly's Test of Sphericity was checked to assess whether variances were equal. The results indicated Sphericity was violated for both sets of data. The Huynh-Feldt correction was used to show the source tables.

The ANOVA source table for scores by competence level and task is shown in Table 12. The *p* values indicate the main effect for task was significant at $p<.01$, and the interaction effect between task and level was not significant. The power statistics show that there was sufficient power to reject the null hypothesis regarding task (power was

Table 12

ANOVA Source Table for Scores by Competence Level and Task

Source	*SS*	*df*	*MS*	*F*	*p*	Partial Eta sq	Power
Within-Participants Effects							
Task	2.01	1.89	1.07	18.68	0.000	0.177	1.00
Task*Level	0.46	3.78	0.12	2.12	0.084	0.047	0.60
Error (Task)	9.37	164.31	0.06				
Between-Participants Effects							
Level	36.05	2	18.03	295.32	0.000	0.087	1.00
Error	5.31	87	0.06				

1.00, greater than .80)(Brown, 2007), but insufficient power to detect a task by level interaction (power was .60, lower than .80). The partial eta^2 values can be interpreted as percentages of variance associated with the task, the task and level interaction, and error (Brown, 2008). Stating with task, the value of 0.177 means that 17.7% of the variance is accounted for by task, whereas the task and level interaction accounts for 4.7%, and the error accounts for 8.7%. The results indicate that task in the test was a "main effect", and it was significantly different across the three levels.

The ANOVA source table for scores by competence level and rating category is shown in Table 13. The *p* values indicate that the main effect for rating category was significant at $p<.01$, and that the interaction effect between rating category and level was also significant at $p<.01$. The power statistics show that both the rating category (power was 1.00, greater than .80), and rating category and level interaction (power was 1.00, greater than .80) had sufficient power to reject the null hypothesis and declare a significant difference (Brown, 2007). The partial eta^2 values can be interpreted as the percentage of variance associated with the rating category, the rating category and level

Table 13

ANOVA Source Table for Scores by Competence Level and Rating Category

Source	*SS*	*df*	*MS*	*F*	*p*	Partial Eta sq	Power
Within-Participants Effects							
Ratingcategory	4.24	3.67	1.16	42.80	0.000	0.330	1.00
Ratingcategory*Level	2.34	7.33	0.32	11.83	0.000	0.214	1.00
Error (Ratingcategory)	8.61	318.83	0.03				
Between-Participants Effects							
Level	53.89	2	26.95	261.15	0.000	0.857	1.00
Error	8.98	87	0.10				

interaction, and error (Brown, 2008). Starting with rating category, the value of 0.330 means that 33.0% of the variance is accounted for by rating category, whereas the rating category and level interaction accounts for 21.4%. This indicates that rating category effect is more important than the rating category and level interaction effect to explaining variance. However, it is worth noting that level error accounts for 85.7% of the variance, which is much more than the above two effects. This error variance is most likely due to a high correlation between rating category and level.

Figure 1 shows a significant interaction effect for rating category by competence level. It indicates that four of the five significantly different rating categories (language use, turn-taking organization, sequence organization, and topic management)

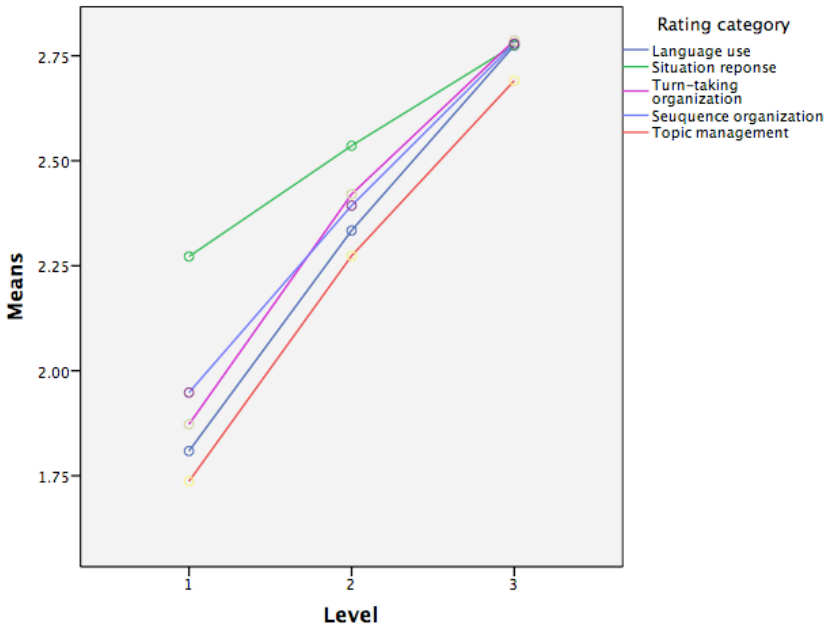


Figure 1. The scoring by competence level and rating category

were not systematically different with regard to competence level. The rating category of situation response had a much higher mean overall, implying that this rating category was

not as distinguishable as the other four rating categories among different competence levels. It is worth noting that Figure 1 shows turn-taking organization crossed sequence organization slightly, which means turn-taking organization resulted in higher score for middle and high competence level candidates than sequence organization.

Follow-up one-way ANOVAs indicate that all the rating categories in the rating rubric as a "factor" are statistically significant differences (all at $p < .01$) across three competence levels.

Reliability estimation. Inter-rater reliability was calculated for the rating categories across all tasks, all tasks combined, and the entire paired interactive tasks; and internal consistency reliability computed for each rating category across all tasks, those categories with interactional features across tasks, and the whole paired interactive tasks. Correlation analyses were then used to investigate the extent to which learners' L2 Chinese pragmatic competence in interaction was related to their Chinese-language proficiency levels.

Inter-rater reliability. Table 14 presents the inter-rater reliability results for the five rating categories associated with each of the paired speaking test's three tasks, as estimated using the Spearman-Brown prophecy formula (Brown, 2012). As Brown (2012, p. 65) explained, "when averaging the two raters' scores (or adding them) before making a decision based on them, the reliability of the two sets of ratings taken together becomes pertinent." The reliability results ranged from as low as 0.19 for situation response in task 1, to as high as 0.80 for both language use and turn-taking organization in task 3.

As a rule of thumb, a Spearman-Brown result of 0.80 or higher indicates sufficient reliability, and 0.90 or higher, good reliability. However, in exploratory studies,

Table 14

Inter-rater Reliability for the Five Rating Categories (Tasks Considered Separately)

Rating category	Task	Spearman-Brown prophecy result
Language use	1	0.74
	2	0.73
	3	0.80
Situation response	1	0.19
	2	0.65
	3	0.48
Turn-taking organization	1	0.75
	2	0.45
	3	0.80
Sequence organization	1	0.67
	2	0.59
	3	0.66
Topic management	1	0.63
	2	0.62
	3	0.75

thresholds as low as 0.60 are not uncommon (Cooper & Schindler, 2014). In this case, if 0.60 is considered the minimum acceptable result, 4 of the 15 task/category pairings fell below this threshold. They were: situation response in task 1 (0.19) and task 3 (0.48), and turn-taking organization (0.45) and sequence organization (0.59) in task 2.

Next, Spearman-Brown prophecy formula was again used to calculate inter-rater reliability for: (1) each of the five rating categories, with the three tasks considered as a single unit; and (2) the test as a whole, without regard to such categories. As shown in Table 15, the first of these two tests resulted in a considerably higher worst score (0.46), again for situation response; and a somewhat lower best score (0.76), again for language use. The three categories with interactional features – that is, turn-taking organization, sequence organization, and topic management – all cleared the acceptable threshold of 0.60, with scores of 0.62, 0.65, and 0.69, respectively. The inter-rater reliability for the

entire test, meanwhile, was 0.77 – only slightly lower than the sufficient reliability threshold of 0.80.

Table 15

Inter-rater Reliability for the Five Rating Categories (All Tasks Combined) and for the Entire Test

Rating category	Spearman-Brown prophecy formula
Language use	0.76
Situation response	0.46
Turn-taking organization	0.62
Sequence organization	0.65
Topic management	0.69
All	0.77

Internal consistency reliability. To investigate internal consistency reliability, Cronbach's alpha values were estimated for (1) each of the five rating categories; (2) the categories related to the measurement of interactional features; and (3) the entire test. The results are shown in Table 16.

Generally, Cronbach's alpha values of at least 0.80 are held to indicate good reliability, and from that level down to 0.70, adequate reliability (Hair, Black, Babin, & Anderson, 2010). By convention, however, a more lenient cutoff point of 0.60 is

acceptable in exploratory studies. As such, only the category of situation response (0.46) could not be deemed suitable for retention, while three of the remaining four rating categories exceeded the value for good reliability. These were language use (0.89), turn-taking organization (0.85), and topic management (0.84). The collective Cronbach's alpha estimate for the three interactional categories (i.e., turn-taking organization, sequence organization, and topic management) was 0.96, the same as the estimate for the entire test.

Table 16

Internal Consistency Reliability for Each of the Five Rating Categories; the Three Rating Categories with Interactional Features; and Overall

Rating category		Coefficient Alpha	Cronbach's Alpha if item deleted
Language use	Task 1	0.89	0.83
	Task 2		0.88
	Task 3		0.84
Situation response	Task 1	0.46	0.24
	Task 2		0.58
	Task 3		0.33
Turn-taking organization	Task 1	0.85	0.75
	Task 2		0.84
	Task 3		0.78
Sequence organization	Task 1	0.77	0.61
	Task 2		0.77
	Task 3		0.70
Topic management	Task 1	0.84	0.75
	Task 2		0.87
	Task 3		0.74
Turn-taking organization		0.96	0.94
Sequence organization			0.93
Topic management			0.95
Overall		0.96	

Based on the above Cronbach's alpha estimates, it can be seen that the internal consistency of the analytical rubric built on the five categories is relatively high, and that its five categories generally measure the same construct, though this is especially true of the three related to interactional features. The situation response category had the lowest Cronbach's alpha estimate (0.46), and would – in the specific case of task 2 – have had the lowest value of Cronbach's Alpha if an item were to be deleted from it (see Table 16).

Correlation analyses. To investigate the extent to which Chinese-language learners' L2 pragmatic competence in interaction was related to their Chinese-language proficiency level, two Pearson correlation analyses were conducted. The first compared the participants' performance on the paired speaking tasks used to measure L2 pragmatic

competence in interaction against their performance on the solo speaking tasks used to measure their Chinese-language proficiency levels. The second analysis compared the results of the same paired speaking tasks against the candidates' Chinese-language proficiency levels as assessed by their own instructors.

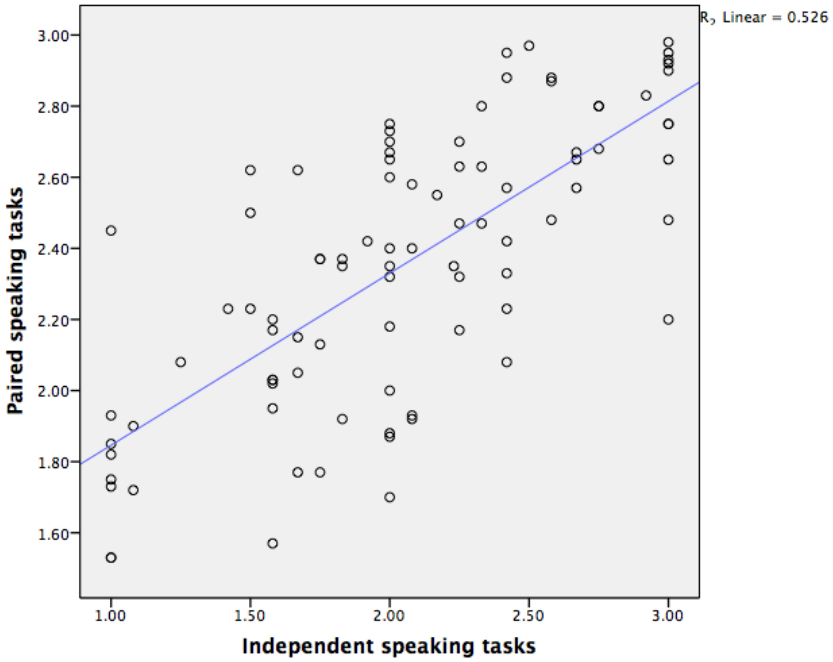


Figure 2. Relationship between the paired and solo speaking tasks

The correlation between the paired and solo speaking task results was found to be 0.73 (p<0.001), with an r^2 value of 0.53, meaning that 53% of the variance in the paired task scores could be explained by the solo task scores. The correlation between the paired speaking tasks and the Chinese-language proficiency levels as assessed by the participants' instructors was 0.81 (p<0.001). In this case, the r^2 value was 0.66, indicating that 66% of the variance in the scores of the paired speaking tasks could be explained by the test-takers' Chinese-language proficiency levels as assessed by their instructors. The

strength of these correlations was relatively high. Figure 2 further illustrates the relationship between this study's paired and solo speaking task results.

Qualitative Analysis

Transcribing and coding. The raters' opinions of the rating process as expressed in both their online interviews and their rating notes, were transcribed and coded. The transcription and coding results were then re-checked by the researcher approximately one month later to ensure their reliability. This process established that the intra-coder reliability (Brown, 2001) was greater than 90%.

For the online interviews with raters. These individual online interviews mainly discussed three aspects of the rating process: rater training; how best to score the two candidates in a paired speaking test; and difficulties the raters encountered, along with any other feelings or suggestions about the rating process that they had. The results of these interviews are summarized below.

First, regarding the rater training, Rater 1 felt that its purpose was clear and that its content was explained thoroughly. The time allocated to gaining an understanding of the rubrics and to do the sample rating were adequate. Rater 2 mentioned that the most important and helpful thing in the rater training was the typical samples for each competence level that were provided to them.

Second, in terms of properly scoring both test-takers in each paired speaking test, Rater 1 mentioned that she first sought to identify the two participants by gender, name, and voice, and then took notes in the process of listening to the recording to ensure that what the two people said could be distinguished clearly. These notes were essential to her rating process, as her scores were summarized from them. She added that the key

prerequisite for rating was familiarity with the descriptions of each level in the analytical rubric, and restated the critical importance of taking notes while listening, especially to record speakers' performance in terms of the rubric's descriptions (e.g., pauses and features of the turn). She also said she believed that if the differences (e.g., voice) between the two people were relatively large, it rendered the whole process considerably easier. Similarly, Rater 2 mentioned that she constantly needed to confirm the participants' identities while rating, and that developing a method of judging which person was which had been very time-consuming. Sometimes, she said, she even went back to previously rated tasks to re-identify speakers. For example, in role-play 1, everyone said their name, whereas some of the same participants failed to mention their names in role-play 2 and/or the situational topic discussion. In the role-play 2, Rater 2 said that she often had to rely on the coffee-machine discourse itself to distinguish the person who wanted to borrow the coffee machine from the person who owned it; and in the situational topic discussion, she generally identified the participants according to where they said they came from. However, on rare occasions it was sometimes still too difficult to determine which candidate was which, and in such cases, she made a comment to that effect in her rating notes.

Third, with regard to difficulties encountered, feelings, and suggestions, Rater 1 felt that the rating experience was very pleasant and that no particular aspect of it needed to be revised. She also mentioned how interesting some of the test-takers' conversations were. The main difficulty she encountered was rating the situational discussion task, and when the two participants' voices were very similar, it became even more difficult. In addition, she felt that in the same task, some candidates talked too much, and sometimes

in ways that were out of keeping with the topic requirements, which added a further challenge to the rating process. In such cases, however, she felt that simply lowering the participants' scores was an appropriate response. In the same context, Rater 2 again noted her trouble with identity confirmation, and suggested that everyone should talk about their names at the beginning of each task, as this could save future raters considerable time that would otherwise be spent identifying them. She also said that the analytical rubric was useful, especially those categories particularly related to L2 pragmatics in interaction (turn-taking organization, sequence organization, and topic management). For example, turn-taking organization refers to whether the second speaker fully understood the previous speaker and responded to him/her properly rather than just randomly talking about some favorite topic; and Rater 2 felt that this was important as a gauge of whether the two speakers could really interact with each other – a question ignored by many language educators. She added that she believed learning a language was not only about vocabulary and grammar, but also about how to speak, and mentioned that even native speakers could have problems in interaction, which might be related to factors other than language proficiency. Lastly, Rater 2 mentioned that in the paired speaking test format, the two speakers might affect each other's performance in some extreme situations: for example, if one person spoke well, yet the other totally failed to understand. However, the recordings generated as part of this study revealed no serious problems of this kind, and generally indicated that the pairings of test-takers worked well, with both participants communicating and no huge differences between them.

For the raters' notes. To track how raters actually used the analytical rubric for interactional features, the researcher asked each of them to record her own reasons for the

scores assigned, immediately next to those scores. Up to a point, this evidence reflected the raters' understanding and application of the rubric, and confirmed the reliability of the rating process. To better gauge the overall picture, the researcher quantified the analysis results based on two rounds of coding, one month apart, as discussed above; the results are presented in Tables 15 through 20. The raters' notes reflected that they understood the rubric and revealed their scoring foci when using the rubric to rate each category, with the numbers referring to how many times each item was mentioned. Despite using the same rubric, however, the raters' respective foci were very different.

Table 17
Summary of Coding Results, Raters' Notes for "Language Use"

			Rater 1	Rater 2
Language use	Overall	Task 1	n/a	3
		Task 2	3	5
		Task 3	n/a	5
	Pronunciation	Task 1	n/a	10
		Task 2	n/a	3
		Task 3	n/a	9
	Grammar	Task 1	n/a	3
		Task 2	n/a	n/a
		Task 3	n/a	12

In rating the category of "language use" (in Table 17), the two raters' scores were mainly based on three subcategories: overall, pronunciation, and grammar. Rater 1 basically did not record her ratings for this category, whereas Rater 2 not only did so, but also listed specific grammatical errors.

As shown in Table 18, the two raters' understanding of the "situation response" category was also divided into three main subcategories: omissions or wrong information, off-topic discussions, and manners that were inappropriate to the situation. Rater 1 paid

Table 18

Summary of Coding Results, Raters' Notes for "Situation Response"

			Rater 1	Rater 2
Situation response	Missing or wrong information	Task 1	29	2
		Task 2	26	10
		Task 3	2	n/a
	Off topic	Task 1	1	n/a
		Task 2	4	2
		Task 3	3	1
	Inappropriate manners	Task 1	n/a	n/a
		Task 2	1	n/a
		Task 3	n/a	n/a

special attention to the rating of this category, especially during the two role-play tasks, in which she marked a large number of candidates as omitting information, giving wrong information, or going off topic, but Rater 2 rarely mentioned this category and only occasionally recorded something related to it.

In the category of "turn-taking organization" (Table 19), the raters' understanding was mainly divided into four subcategories: naturalness, pauses, turn

Table 19

Summary of Coding Results, Raters' Notes for "Turn-taking Organization"

			Rater 1	Rater 2
Turn-taking organization	Pause	Task 1	9	3
		Task 2	2	3
		Task 3	3	n/a
	Naturalness	Task 1	21	1
		Task 2	4	1
		Task 3	n/a	n/a
	Turn length	Task 1	4	n/a
		Task 2	6	n/a
		Task 3	1	n/a
	Overlapping/interruption	Task 1	3	1
		Task 2	1	n/a
		Task 3	n/a	n/a

length, and overlapping/interruption. This was broadly based on the rating rubric, though the element of naturalness had been added by the raters. Rater 1 paid special attention to the naturalness of the turn-taking, as well as to the candidates' pauses, and whether turn lengths were excessive. Rater 2, in contrast, made few records regarding this category.

Table 20

Summary of Coding Results, Raters' Notes for "Sequence Organization"

			Rater 1	Rater 2
Sequence organization	Overall	Task 1	n/a	n/a
		Task 2	3	n/a
		Task 3	n/a	n/a
	Understanding of previous turns	Task 1	3	2
		Task 2	6	2
		Task 3	n/a	2
	Preferences	Task 1	2	n/a
		Task 2	3	4
		Task 3	n/a	n/a
	Response tokens	Task 1	1	n/a
		Task 2	2	n/a
		Task 3	n/a	n/a
	Pre-sequencing	Task 1	n/a	n/a
		Task 2	n/a	2
		Task 3	n/a	n/a

As shown in Table 20, the two raters treated "sequence organization" as comprising five subcomponents, that is, overall sequence organization, understanding of previous turns, preference organization, response tokens, and pre-sequencing. Neither rater recorded much information about candidates' performance in this category. In addition, Rater 1 sometimes used vague expressions: for example, mentioning that a test-taker "did not have a concept of sequence", without specifically pointing out that what was wrong with that person's sequence organization.

The two raters understood "topic management" (Table 21) to include six subcategories: topic initiation, topic development, topic shift, topic termination, incomplete topic, and overall performance. Rater 1 had fairly meticulous records for these subcategories, especially whether each topic's initiation was natural, and its ending

Table 21
Summary of Coding Results, Raters' Notes for "Topic Management"

			Rater 1	Rater 2
Topic management	Initiation	Task 1	6	4
		Task 2	19	4
		Task 3	n/a	2
	Development	Task 1	1	14
		Task 2	2	n/a
		Task 3	2	8
	Shift	Task 1	3	n/a
		Task 2	4	n/a
		Task 3	3	n/a
	Termination	Task 1	4	n/a
		Task 2	4	n/a
		Task 3	2	n/a
	Incomplete	Task 1	n/a	n/a
		Task 2	4	n/a
		Task 3	1	n/a
	Overall	Task 1	n/a	3
		Task 2	n/a	n/a
		Task 3	n/a	n/a

not abrupt. Rater 2, on the other hand, had obviously focused most of her attention on topic development, and especially so during role-play 1 and the situational topic discussion. Though she also noted whether the openings of conversations were abrupt, she made no records relating to the shift, termination, and incompleteness subcategories.

As shown in Table 22, both raters' notes also mentioned some matters beyond the scope of the rubric. For example, both had made records of whether the interactional

pattern was dominated by one person. Rater 2 placed special emphasis on the "content" of test-takers' performance, and also recorded whether their language was authentic or not; information about the examinees' personalities; and whether the two paired candidates had similar voices.

Table 22
Summary of Coding Results, Raters' Notes for "Other Elements"

			Rater 1	Rater 2
Other elements	Dominance	Task 1	1	2
		Task 2	9	9
		Task 3	6	13
	Content	Task 1	n/a	13
		Task 2	n/a	10
		Task 3	n/a	17
	Authenticity	Task 1	n/a	n/a
		Task 2	n/a	5
		Task 3	n/a	n/a
	Personality		n/a	1
	Voice similarity		n/a	1

On the whole, analysis of the raters' notes revealed that Rater 1 paid more attention to the use of the analytical rubric for interactional features than Rater 2 did, while Rater 2 focused instead on content, authenticity, and personality. In the first two role-play tasks, Rater 1 made detailed records regarding her scoring of the situation response. In rating role-play 1 and the situational discussion tasks, Rater 2 focused on the candidates' topic development. It is also worth noting that, in the context of rating the situational topic discussion task, both raters paid the most attention to the rating of content, and Rater 1 did not attach as much importance to situational responses as she did during the role-play tasks. In addition, both raters' interest in patterns of conversational dominance raises the possibility that this issue may be worthy of further study.

Discourse analysis. Samples of in-test discourse arising from all three tasks were randomly chosen by the researcher as illustrations of the examinees' performance in the high-, middle-, and low-competence groups. However, for purposes of this phase of analysis, such levels were determined by the scores assigned to each pair by the two raters, while the language teachers' pre-speaking test assessments of the students' language proficiency served only as a reference. The sequence of speaking-task DA analysis was role-play 1, then role-play 2, and lastly the situational discussion task.

Of these three tasks, role-play 2 was found to have the lowest degree of openness, and the situational discussion task the highest. Since the first part of role-play 1 consisted of mutual introductions – a fixed mode – and its second part was more open, two excerpts from role-play 1 will be shown, to represent these two parts. Thus, a total of four excerpts for each level will be provided to illustrate the test-takers' L2 pragmatic competence in interaction. Again, all transcriptions were reviewed one month after they were made, to ensure their reliability, and agreement between the two sets of transcriptions exceeded 90%. Transcriptions and recordings were used simultaneously for DA purposes. In both role-play tasks, the prompts were the same for all three proficiency levels, whereas in the situational topic discussion task, there were separate topics for each such level: specifically, *hobbies* for the low-proficiency group (with prompts including reading books, watching movies, playing video games, doing exercise, and traveling); *countries* for the middle-proficiency group (with prompts including shopping habits, means of transportation, recreation and entertainment, environmental issues, and educational modes); and *urban livability* for the high-proficiency group (with prompts including environmental pollution, social security, quality of life, and the speed of

economic development). The test-takers did not need to cover every prompt for the discussion task. The results of the analysis are presented below in five categories: language use, situation response, turn-taking organization, sequence organization, and topic management.

Discourse analysis: high-competence group

Language use. This category refers principally to whether the test-taker has abundant language to deal with the interaction's L2 pragmatics without difficulties. Judging by their four excerpts, the high-competence test-takers had mastered a wide range of language; were free to express themselves in interactions without difficulties; and produced language accurate enough that it did not create obstacles to communication. In terms of the range of their language, the high-competence test-takers exhibited ample knowledge of vocabulary and grammatical structure, using not only standard high-frequency words, but also new Internet terms such as "*学霸* (a learning tyrants)" (turn 30 in Excerpt 3). They could also use sentences to express specific pragmatic meanings: for example, to answer the previous speaker's question (turn 2 in Excerpt 1). To Ban's question "Are you Chinese?", the normal answer would be "yes" or "no", but Zeng did not follow the rules, and used a question ("Do you think I look like Chinese?") to express the meaning "I'm not Chinese". At the same time, he lived up the atmosphere. In addition, high-competence test-takers used discourse devices not only in their expressions confined to sentences, but also in longer discourses, as a means of expressing their opinions more deeply. Examples of this included "*是一个* (is one)" (turn 9 in Excerpt 4); "*然后另外一个…就是* (then another one is that)" (turn 19 in Excerpt 4); "*还有...方面* (also has another aspect)" (turns 25 and 27 in Excerpt 4); and "*然后还有另外一个...的因素就是*

Excerpt 1: High-competence pairs

Knowing each other in a Chinese corner (part A). b: Ban, z: Zeng

1	b 啊，不好意思，你是中国人吗? *Excuse me, are you Chinese?*	2	z 你觉得我像中国人吗？哈哈~= *Do you think I look like Chinese? Haha*
3	b =对，我觉得你像中国人，因为我，那个，你的脸好像你是中国人，呵呵~ *Yes, I thin you look like Chinese, because your face looks like Chinese, hoho.*	4	z 不好意思，我不是中国人的，啊= *I'm sorry, I'm not Chinese.*
5	b =那你是哪国人?= *Then what country are you from?*	6	z =啊，我是越南人，你呢? *Ah, I'm Vietnamese. How about you?*
7	b 啊，我是泰国人，我叫班龙。 *Ah, I'm Thai. My name is Long Ban.*	8	z 班龙，嗯，你好。 *Long Ban, eh, hello.*
9	b 你好。 *Hello.*	10	z 认识你很高兴。 *Nice to meet you.*
11	b 你叫什么名字？ *What's your name?*	12	z 我叫泉耀。 *My name is Quanyao.*
13	b 哦 Oh.	14	z 曾泉耀= Quanyao Zeng.
15	b =曾泉耀,我是班龙,上班的班,龙*的龙 *Quanyao Zeng, I'm Long Ban. Ban as ban in shangban (go to work). Long refers to * .*	16	z 嗯，我是曾，曾经的曾那个曾，泉是矿泉水的泉， *Well, I'm Zeng, Zeng as ceng in "cengjing" (once), Quan as quan in "kuangquan shui" (mineral water).*
17	b 嗯 *Hm.*	18	z 耀是荣耀的耀。 *Yao is as yao in "rongyao" (glory).*
19	b 你是来中国是学（.）做什么？ *What are you doing when you come to China?*	20	z 嗯，我来是为了学语言的。 *Well, I came to learn the language.*
21	b 啊，学语言，那你是刚刚*是新学生，还是已经学了，还- *ah, learn the language, then you are a new student, or have already learned, but also-*	22	z 我是大一的学生，我今年刚来的，你呢？ *I am a freshman student, I just came this year, and how about you?*
23	b 我，我是刚，今年的学生，刚刚来的。 *I, I am just this year's student, just arrived.*	24	z 啊，所以- ah, so-
25	b 所以还不认识，有没有，还没有朋友。 *So I don't know (anyone). I don't have friends.*	26	z 你学什么专业的？ *What is your major?*
27	b 我是汉语专，汉语国际教育。 *I'm majored in Chinese international education.*	28	z 嗯。 *Yes.*
29	b 你? *How about you?*	30	z 我是音乐厅的学生。 *I am a student of the concert hall.*
31	b 哦 *Oh*	32	z 嗯 *Hm.*
33	b 很厉害！ *Awesome!*	34	z [英语学院的 *College of English*
35	b [专业配音的, 呵呵~ *The professional voice actor, hoho*	36	z 过奖，过奖。 *You flattered me.*

(and another factor is)" (turn 57 in Excerpt 4). In addition, test-takers at this competence level were able to use language strategies. For example, when learning each other's names, both Ban and Zeng explained their own names using various commonly used

words, so that they could communicate better and increase the likelihood that their own names would be remembered (turns 15, 16, and 17 in Excerpt 1).

In terms of language use, this competence group's four excerpts were consistently accurate, with few vocabulary or grammatical issues. For example, "*以后* (afterwards)" (turn 23 in Excerpt 3) should be "*之后* (later)", and the rhetorical question "*我哪不会去呢* (I definitely will go)" (turn 28 in Excerpt 3) had a word-order problem (the right order would be "*我哪会不去呢*"). Such errors were rare and did not tend to affect the speaker's overall understandability, and therefore could be ignored by raters.

Situation response. This category refers to whether a candidate can consciously and properly navigate the situation required by the task. In addition to the accurate use of language, sociolinguistics (Leech, 1983) – including test-takers' perceptions of society and their sensitivity to situations – constitutes an important component of learners' L2 pragmatic competence in interaction. The high-competence test-takers were highly sensitive to, and consciously performed, the required contextual tasks. For example, the conversation in Excerpt 1 was an effective depiction of a typical situation of new friends meeting for the first time. In this case, Zeng and Ban completed the task according to its requirements: Zeng spoke to Ban first, and then the two asked about each other's backgrounds and current activities. Similarly, following the instructions, Huang invited Wang to attend a Christmas party (turns 1, 3, 5, 7, 9, and 11 in Excerpt 3), and borrowed Wang's coffee machine (turns 33 and 35 in Excerpt 3). Wang's task was to tell Huang she had something to do on the day, though she was free decide whether to accept Huang's invitation or not. Wang chose to go, but said she would have to leave early (turns 20, 22, and 24 in Excerpt 3). In addition, as required, Wang explained that her

Excerpt 2: High-competence pairs
Knowing each other in a Chinese corner (part B). t: Ti, a: Ai

1	t 六七年, *Six or seven years,*	2	a 嗯： *hm*
3	t 去过哪些地方呢? *Where have you been?*	4	a 在武汉吗? *Is it in Wuhan?*
5	t 中国,哪? *Where in China?*	6	a 中国，啊，济，济南，[聊城 *China, ah, Jinan, Liaocheng*
7	t [啊, 济南我也去过= *ah, Jinan I have also been to*	8	a =啊，北京, 啊，你去过济南，啊，为什么去过济南? *ah, Beijing, ah, you have been to Jinan, ah, why have you been to Jinan?*
9	t 去找朋友，呵呵~ *To see a friend, hoho.*	10	a 男朋友吗? *Your boyfriend?*
11	t 呵呵~, 对, 呵呵~ *Hoho, right, hoho.*	12	a 呵呵呵~ *Hohoho*
13	t 他以前在那边读书。 *He used to study there.*	14	a 哦：= *Oh.*
15	t =然后我经常去那边找他。 *Then I often go to see him.*	16	a 哦，他为什么不来找你呀？ *Oh, why didn't he come to see you?*
17	t 他偶尔也会来，因为我时间比他多，呵呵~ *He came occasionally, because I have more time than he does.*	18	a 哦，好，啊（.） *Oh, OK, ah.*
19	t 你觉得那边怎么样? *What do you think about there?*	20	a 嗯，我觉得那边儿挺好的，比这里好一些，[你觉得呢? *Um, I think there is good and better than here. What do you think?*
21	t [嗯, 对，对，但是，呃：，那边没有地铁啊，很麻烦。 *Well, yes, yes, but, uh:, there is no subway over there. It's very troublesome.*	22	a 没关系呀，地铁，我们国家没有地铁。 *It's okay, subway, our country has no subway.*
23	t [也是。 *That's right.*	24	a [还是能活着啊。 *Still alive.*

coffee machine was broken, and then helped Huang to find an alternative solution (turns 36, 38, and 42 in Excerpt 3). In other words, both excerpts demonstrated the accurate completion of all required tasks.

Excerpt 3: High-competence pairs
Inviting a friend to a Christmas party and borrowing a coffee machine.
h: Huang, w: Wang

1	h 欸，王艳，呃，[你圣诞节 *Hi, Yan Wang, you Christmas*	2	w [hey，girl
3	h 有什么打算吗？ *What are your plans?*	4	w 呃，这圣诞节啊，嘶，嗯，对呀，我也正在想想呀，我该做什么啊。 *Hey, this Christmas, hey, well, right, I'm thinking about it too, what am I supposed to do?*

5 h 哦，是这样的，我好，我刚好正在准备一个圣诞节晚会。

Oh, yes, I'm fine. I'm just preparing for a Christmas party

6 w 嗯

hm

7 h 就（.）会在十二月二十四日举办的，就下午四点半。

will be held on December 24, just after 4:30 in the afternoon.

8 w 嗯

hm

9 h 然后可能晚上十点结束

and then can finish at 10 pm.

10 w 嗯

hm

11 h 你要不要参加？[可以一起

Do you want to participate? [Can be together

12 w [哎呀, 听起来很有意思啊!

Ah, it sounds very interesting!

13 h 对啊。

Yes

14 w 我也对此, 我都很感兴趣。

I am also very interested in this.

15 h 嗯

hm

16 w 你是说，你刚才是说几点来的?

You mean, what time did you say?

17 h 下午四点半，应该我们没有课了吧。

4:30 in the afternoon, we should have no class.

18 w 下午四点半，对，不是刚好下课吗？

4:30 pm, right, isn't it just the break after the class?

19 h 对对，所以可以跟我一起去。

correct, so you can go with me.

20 w 可以的，反正我的时间还挺有限的。

Yes, I have limited time.

21 h 嗯

Hm

22 w 因为我那天，嘶，呃，七点，呃，*要去约会呢*，呵呵呵呵~

Because I am going on a date at seven o'clock.Hoho.

23 h 呵呵~，那没事，那你可以四点半去，“以后”七点如果你有事的话，你就可以先走。

Oh, that's okay, then you can come at half past four. After seven o'clock, if you have something to do, you can go first.

24 w 好啊，四点半，还有，啊，两个，那个，半小时的时间嘛，[时间够长了

Well, half past four, and two and a half is long enough.

25 h [嗯，反正圣诞会北大所有学生都会参加。

Well, all students of Peking University will attend.

26 w 对啊=

Right

27 h =所以你可以多多交一些新的朋友。

So you can make more new friends.

28 w 当然可以啊，就，啊，凭你自己来邀请我，我，*我*<哪不会去呢>?

Of course I will, invite me by yourself, how come I won't go?

29 h 呵呵~

Hoho

30 w 就咱班，就像你咱班学霸。

You are straight A student in our class.

31 h 呵呵~，[没有没有

Hono, no no.

32 w [这样辛苦, 呵呵~，邀请我，呵呵~

You made such efforts to invite me, hoho.

33 h (.) 呃，呃，然后还有一件事，不知道你可不可以帮我的忙?

Uh, then there's one more thing, I don't you know if you can help me?

34 w 嗯，什么呢?

Well, what?

35 h 就因为我好像需要一个咖啡机，你可不可以借我你的咖啡机?

Just because I seem to need a coffee machine, can you borrow me your coffee machine?

36 w 呃，可以的，诶，等一下，哎呀，怎么办啊? 就我今天早上我，我也要用我的咖啡机，反正它，它坏了欸，你-

Uh, yes, uh, wait, oh, what can I do? I used it this morning. The coffee machine, anyway, it's broken, you -

37 h -坏了吗?

- is it broken?

38 w 可以那个，我想想哦，好像咱班一个从美国来的谁，马克，他有诶。

Yes, I think, oh, it seems like someone who has come from the United States, Mark, he has.

39 h 啊，是吗？

Yeah, right?

40 w 嗯

Hm

41 h [他有

He has

42 w [对，就我们, (.) 好像在我们一单元，1012 的房间，可以过去见一下，他人很好。

Yes, let's (go). It looks like we're in the same unit, room 1012, we can go over to check. The guy is nice.

43 h 呃，[好啊好啊

Uh, okay.

44 ……

In addition, the high-competence test-takers could respond appropriately to specific situations. For example, in the case of Excerpt 1, both Zeng and Ban were told that they did not know each other, and their discourse reflected this social distance: both were very polite and courteous, and both said "*不好意思* (excuse me)" (turns 1 and 4 in Excerpt 1).

In Excerpt 2, Ti and Ai were told that for purposes of the task, they were good friends. In keeping with this instruction, their conversation reflected a close relationship in which they could ask personal questions. For example, Ai asked Ti if she had gone to Jinan to see her boyfriend (turn 10 in Excerpt 2); Ti answered "yes" (turn 11 in Excerpt 2); and then the two girls laughed together in a way that seemed very intimate. In short, based on the tone as well as the content of Excerpts 2 and 3, the apparent social distance between the two pairs of high-competence interlocutors was appropriate to the test's imaginary situations.

Turn-taking organization. The four excerpts randomly chosen to represent the high-competence group show that the test-takers at this level were able to engage in turn-taking naturally and smoothly. Usually, there were no gaps between the turns, which were well connected and neither excessively long nor short.

According to ten Have (2007), as noted previously, the two key characteristics of "conversation" are that only one person speaks at a time, and that the gaps and overlaps marking changes of speaker are very small. However, when real interactions are very enthusiastic, this is often signaled by "latches/overlaps", which are indicative of cooperation (Tannen, 1982) and fusion (McCarthy, 2010). In this context, it is also worth remembering Sacks et al.'s (1974) dictum that speaker shift can be achieved in three

Excerpt 4: High-competence pairs

Situational discussion about urban livability. y: Ya, m: Men

1 y 你，你毕业之后打算去哪里发展？
Where do you plan to go after graduation?

2 m 啊，这个啊，(.) 这个：太不好说了。我还有，但，其实我还有一些‘想’，一些想法，是：，我毕业以后想去，想在联合国工作，在联合国的那个，搞：环境啊，还是发展的那项目 UNTP。呃：，所以还不太确定，那我再想想，可能是中国啊，还是，可能是美国啊。但是这个还，还没想好。但是，我想，当然想在一个：，我特别，我特别看重安全，安全的地，对。
ah, it is hard to say. I have some thoughts. Some of my thoughts are: After I graduate, I want to go to work in the United Nations. At the United Nations, I'm going to do something related to the Environment, or the UNTP project. Oh, so I'm not sure. Then I will think again. It may be in China, or it may be in the United States. But this is still not decided. However, I think, of course, want to be in one: I especially value safety.

3 y 那你觉得中国，你，现在北京安全吗？
Do you think China, Beijing now safe?

4 m 嗯，我觉得，我觉得，挺安全的，觉得北京挺安全的，可以啊，晚上的时候，可以随便走一走啊。那个我觉得（.）在，在整个儿在亚洲那个，啊，中国，还是，主要是北京是挺安全的。但：，那个，也，也要考虑到环境那方面，因为，啊：，想，想找一个地方发展，也要，要包括很多因素，环境，呃：，北，呃，中国在环境方面还有一些很大的问题，我觉得是啊。
Hm, I think, I think it's very safe. I feel that Beijing is very safe. You can, in the evening, you can just walk around. That I think in, in the whole Asia, ah, China, or mainly Beijing is very safe. However, that, also we must consider the environment aspect, because, ah, I want, want to find a place to develop my career. Also, many other factors included, the environment, uh, Bei, uh, there are some big problems in China in terms of the environment, I think so.

5 y 对啊
Right

6 m 嗯，[嗯
Hm, hm

7 y [其实对我来说，这个，嗯：，工作的，呃，工作方面
In fact, for me, work aspect

8 m 嗯
Hm

9 y 是一个很重要的因素。
is a very important factor.

10 m 嗯，嗯
Hm, hm

11 y 我哪里找到比较好的工作，我就去哪里发展。
Where I can find a better job, I will go there.

12 m 嗯
Hm

13 y 但是，[呃
But,

14 m [哈哈~，真的？
Haha, really?

15 y 呵呵~，对-
Hoho, right.

16 m 会考虑到哪些方面, [或是
will consider about what, or

17 y [这是一个很重要的因素
This is an important factor.

18 m 嗯
Hm

19 y 然后另外一个很重要的因素就是，如果以后我要跟我老婆一起有一个孩子，
Then another important factor is if I want to have a child with my wife in the future,

20 m 嗯=
Hm

21 y =然后环境
Then environment

22 m 环境
Environment

23 y 也是一个很‘dao’，呃，[很重要的因素
It is also an important factor.

24 m [很重要的
Very important.

25 y 还有
Also

26 m 嗯
Hm

27 y 教育方面

Educational aspect

28 m 教育方面, 啊

Educational aspect

29 y 对，然后我，我现在觉得我毕业之后可能先在一个中国的比较大的城市，

Right, then I, I now think that after graduating I may first be in a relatively large city in China,

30 m 嗯

Hm

31 y 一，呃，对，发展，比如说深圳，

One, uh, right, development, for example, like Shenzhen.

32 m 嗯

Hm

33 y 我觉得深圳<比北京比较好>，

I think Shenzhen is better than Beijing

34 m 哦=

Oh

35 y =因为深圳工作的机会=

Because of the chance of working in Shenzhen

36 m =也好

Also good

37 y 也有，然后环境比较好，然后是一个很(.) 现代化的城市。

the environment is better, and it is a very modern city.

38 m 看来你对深圳的印，呵~ -

It seems your impress[sion of Shenzhen, ho

39 y 对

Right

40 m 的理解是，是很-

Your understanding is

41 y 对

Right

42 m 很深的啊

Very deep

43 y 对

Right

44 m 你在那边呆了多久？

How long have you been there?

45 y 呵呵，我在那边呆了啊（.）几个月，[*

Oh, I stayed there for a few months (.) months,

46 m [嗯？几个月就，就决定在, 在那边发展了？呵呵~

Within a few months, already decided to live there to develop?

47 y 对，因为我老婆的，呃，父母就[在那边，然后

Yeah, because my wife's parents are [over there

48 m [就在那边，哦:

just over there

49 y 嗯，对

Hm, right

50 m 嗯，嗯，嗯

Hm

51 y 然后啊，生孩子之后我还是要回国，回德国

Then after I have a baby, I still want to go back to Germany.

52 m 嗯=

Hm

53 y =因为我觉得德国的这个，呃，环境比较好，

Because I think Germany's environment is better

54 m 嗯

Hm

55 y 教育方面我觉得也，呃（.）

Education I think also

56 m 也相当[可以了

is also quite

57 y[比较好, 对，然后还有另外一个很重要的因素就是食品安全，

It's better, yes, then there is another very important factor is food safety

58 m 嗯

Hm

59 y 对我来说，恩(.)，我自己吃什么不是很重要的，我现在已经长大了，

For me, what I eat for myself is not very important. I am now grown up.

60 m 嗯

Hm

61 y 但是对一个很小的孩子来说，呃，他吃什么，呃，很重要，所以我：觉得生孩子之后我还是要回，回德国，

But for a very young child, what he eats, is very important, so I feel like I have to go back to Germany after I have kids.

62 m 嗯

Hm

63 y 然后，可能去法兰克福或者慕尼黑，

Then maybe go to Frankfurt or Munich

64 m 哦:

Oh

65 y 比较，比较大的城市，因为那边的工作机会比较多。

Relatively, relatively big cities, since there are more job opportunities in larger cities

66 m 啊，看来我们的角度也不太一样，因为现在我又，嘶，我又没老婆，我又，所以我考虑的那，我考虑的那些事情也是就是对我，对我自己而言。所以说，呃，我现在找的是（.）比较，要，要在比较安，安全的，比较安全，环境比较好的-

It seems that our perspective is not the same, because now I don't have a wife. And the things I'm thinking about are also to me, to myself. So what I'm looking for now is a place safer with a better environment.

main ways. In the order of their prevalence in actual conversation, these are: the next speaker being selected by the previous speaker ("other-selection"), the speaker selecting him- or herself ("self-selection"), or the current speaker continuing to speak. All this being said, if there is a long pause between speakers, or the current speaker is interrupted, a turn cannot be deemed "good". But conversely, if the current speaker skillfully invites others to participate in the conversation, it may indicate that he/she has good L2 pragmatic competence in interaction; and active self-selection that enables someone to take the floor at the right time can also be considered a sign of such competence.

In Excerpt 1, both candidates were very willing to contribute to the interaction, so the speakership changed frequently and turn-taking was fast. Thus, there were multiple latches (turns 2-3; turns 4-5; turns 5-6; and turns 14-15). In Excerpt 3, both Huang and Wang also actively interacted with each other, again resulting in multiple overlaps (turns 1-2; turns 11-12; turns 24-25; turns 31-32; and turns 41-42). In all four excerpts from the high-competence group, there were basically no pauses between the turns, in keeping with their close logical connections. Although individual interruptions occurred (e.g., Zeng at turn 24 in Excerpt 1, and Ya at turn 15 in Excerpt 4), they did not affect the overall fluency of turn-taking.

In all four of the high-competence pairs' excerpts, speaker shift took two forms – other-selection and self-selection – with the former being more common (just as in real-world situations generally). Slef-selection occurred, for example, in turns 1, 11, and 19 in Excerpt 1, all of which involved Ban asking Zeng a question, thus naturally identifying Zeng as the speaker for the next turn. Similar examples were also found in all three of the other excerpts (e.g., turn 3 in Excerpt 2; turn 35 in Excerpt 3; turn 3 in Excerpt 4).

Notably, listeners in the high-competence group employed self-selection as the next speaker, showing their initiative in participating in the interaction. For instance, in turn 21 in Excerpt 2, when Ai told Ti that Jinan was better than Wuhan, Ti was eager to express her differing point of view – "Jinan has no subway, which is inconvenient" – and began making this statement before Ai had a chance to ask her for her opinion. Another example was provided by turn 46 in Excerpt 4: when Men heard that Ya had been in Shenzhen for only a few months, and felt Shenzhen was good, Men was eager to express his doubts about this, and thus began to speak before turn 45 was over. However, self-selection was not the norm, presumably due to widespread conversational conventions that forbid interrupting someone or overlapping with the current speaker, as discussed above.

Sequence organization. The second core concept of conversational organization is sequence organization, a category that involves three main factors: the degree to which a test-taker comprehends the previous turn; whether he/she is able to use different response tokens; and whether he/she properly designs turns to express preferred and non-preferred alternatives. As mentioned previously, the basic building block of a sequence is an AP, which refers to the turn-taking by two speakers performing FPP and SPP, where a particular type of FPP requires a specific, corresponding type of SPP (Schegloff, 1968). This means that in order to interact smoothly, that is, without communication obstacles, the listener must understand the meaning of the previous speaker's utterance. All eight test-takers represented in the excerpts from the high-competence group could fully understand each other's turns, generated appropriate responses to them, and had no communication difficulties or misunderstandings.

Another important facet of sequence organization is made up of response tokens, which are of four main types: confirmation of previous information; repetition of part or all of the other party's speech; provision of assessment; and use of signal words to take the floor. Of these, the first – confirmation – constitutes the most basic sign that the listener is listening. It appears in all four excerpts, for instance, as "*嗯* (hm)" (turn 28 in Excerpt 1; turn 2 in Excerpt 2) ,"*哦* (oh)" (turn 31 in Excerpt 1; turn 14 in Excerpt 2; turn 5 in Excerpt 3; turn 6 in Excerpt 4), and laughter (turn 29 in Excerpt 3). Some listeners also repeat some or all of the previous speaker's words, again as an indication that they are listening; but in the case of such repetition, the degree of interaction is higher than when simply confirming the words of the other party. For example, in Excerpt 4, Men repeatedly echoed Ya's words or phrases, such as in turns 22, 24, and 28. Providing one's own assessment, meanwhile, is also an arguably more convincing sign that one is paying attention than merely confirming the information is, given that a speaker can use the latter technique without actually understanding what the other party has said (Goodwin, 1986). Such assessments can be brief or extended, a typical brief one being to agree or disagree. A simple assessment, in Excerpt 2, turn 18 "*哦，好* (oh, good)", indicated that Ai understood and accepted Ti's claim. In Excerpt 3, turn 26, Wang expressed his understanding and agreement by saying "*对啊* (right)". More extensive assessment, however, is an even more effective demonstration of the listener's understanding of the speaker's words and also showing her contributions to the interaction. For example, in Excerpt 3, after Huang completed turn 11, Wang evaluated the activities introduced by Huang as interesting in turns 12 and 14, and further expressed her intention to not hinder the party-invitation action. Lastly, some people used signal words (Gardner, 2006;

Jefferson, 1993) to indicate that they wanted to take the floor. For example, in Excerpt 1, Zeng (turn 24) used "*啊* (ah)" and "*所以*(so)" to indicate that he wanted to speak, but was interrupted by Ban (turn 25), and the attempted floor-taking was unsuccessful. Similarly, in Excerpt 2, Ti (turn 7) said "*啊* (ah)" and "*济南我也去过* (I have been to Jinan)" without waiting until the former speaker had completely finished the turn, thus indicating that she wanted to speak.

In Excerpt 4, Men's use of response tokens was very prominent, and included all four types of tokens mentioned above: e.g., in turn 6 (confirmation), turn 24 (repetition), turns 38, 40, and 42 (assessment), and turn 66 (using the signal words "*啊* [ah]" and "*看来* [it appears]"). Through frequent use of various response tokens, Men demonstrated strong L2 pragmatic competence along with an enthusiasm about participating in the interaction. All four of the randomly selected excerpts from this competence group likewise indicated that test-takers used all the above-mentioned response tokens frequently and properly, including more complex ones, meaning that – as listeners – they actively monitored the content of their interlocutors' speech and negotiated and communicated with them during their interaction.

Another important facet of sequence organization, preference (Pomerantz, 1984; Sacks, 1987), refers to the speaker designing a turn in a way that suits the recipient, so as to minimize the threat of losing face (Sacks & Schegloff, 1979). The most obvious criterion for a preferred structure in a specific context is that the turn shape not be marked by delays, mitigative devices, or explanations. While dealing with dispreferred structures, the eight high-competence candidates in these four excerpts exhibited a strong command of preference structure. All could use simple pre-sequences to indicate the topics to be

discussed, depending on what they wanted to accomplish; and based on their needs to complete certain social actions, some implemented complex multi-turn pre-sequences as well, and/or used post-sequences to further highlight their intentions. To effectively interact with others, speakers in this group also used pauses (e.g., turn 33 in Excerpt 3), mitigative devices (e.g., "*不知道你可不可以帮我* [don't know if you can help me]", turn 33 in Excerpt 3, and "*你可不可以借我* [Can you lend me]", turn 35 in Excerpt 3), accounts (e.g., "*就因为我好像* [Because I seem to]", turn 35 in Excerpt 3).

The four excerpts also clearly show that high-competence test-takers could correctly implement the preference sequences required by the test, including invitations, requests, and turning-down of requests; and that they might also use preference sequences not required by the test, such as compliments (turn 33 in Excerpt 1) and disagreement (turn 21 in Excerpt 2). Taking the invitation and turning-down of a request (in Excerpt 3) as an example, before issuing the invitation in turn 11, Huang used complex multi-turn pre-sequences (turns 1, 3, 5, 7, and 9) to introduce and forecast it. Huang first asked whether Wang had some plan for the coming Christmas (turns 1 and 3). Wang's reply indicated that she had not arranged anything (turn 4), thus clearing the first pre-invitation hurdle. Huang then described the activities she was planning (turn 5), and to facilitate Wang's acceptance of the invitation, Huang also told Wang the party's start and end times in a very earnest tone (turns 7 and 9). Then, after the invitation action had taken place, a post-sequence continued to support it (turns 17 and 19). And once Huang learned that Wang would need to go somewhere else after 7:00 p.m. on the day of the party (turns 22 and 24), she was active in stressing her own flexibility in this regard ("*可以先走* [can go first]", turn 23) to further encourage Wang to come. Huang also used a

psychological strategy of indicating her own considerateness, such as by saying Wang could make friends at the party (turn 25, 27). After this series of actions, Wang readily accepted the invitation (turns 28, 30, and 32) and the invitation action was completed successfully. Huang's request regarding the coffee machine (turns 33 and 35) was quickly agreed to by Wang (turn 36), but then she remembered that her coffee machine was broken, so used the pre-sequence "*等一下，哎呀，怎么办* (wait, oh, what can I do?)", before providing an explanation: "it was broken in the morning". She then used a post-sequence to actively help Huang to solve the problem (turns 38 and 42), thereby further reducing the harm that might have been caused by the turning-down of the request, and making it easier for Huang to accept (turn 43). In other words, Wang successfully completed this dispreferred action.

Topic management. This category includes whether learners can naturally start new topics, develop their own and others' topics, achieve smooth topic transitions, and naturally bring discussion of a topic to a close. The high-competence test-takers demonstrated high competence in this area, and were especially confident about developing topics, regardless of whether they or their partners had initiated them.

Ideally, rather than being abrupt, the initiation of a topic will seem natural, that is, be linked to the interlocutors' identities, the conversational context, or topics that were raised earlier. The four excerpts from the high-competence group of test-takers indicated that all these students could initiate topics naturally and smoothly. For example, in Excerpt 1 – Ban and Zeng at the Chinese corner – Ban commenced chatting with Zeng by asking questions related to the setting (turns 1, 3, 5). Once the two partners had learned

more information about each other, Ban naturally and smoothly opened another new topic, by asking Zeng what there was to do in China (turn 19).

In the sphere of topic development, listeners' minimal assessment responses are sometimes regarded as extensions of topics, despite also being response tokens. However, to avoid this categorical overlap, the present study does not treat assessment, confirmation, or other response tokens as topic development. Rather, its category of topic development includes: continuing to explore; asking back; helping the other person to complete his/her meaning; and further developing the topic.

Continuing to explore mainly refers to the speaker or listener wanting to obtain more detailed or accurate information. This was very common in the discourse of the sampled high-competence test takers. For example, Ban learned that Zeng studied language in China in turn 21 (Excerpt 1). He continued to explore whether he had just arrived, or had been studying for some time. This was comparable to turn 4 in Excerpt 2, turn 16 in Excerpt 3, and turn 3 in Excerpt 4.

Asking back means that the listener, after answering the speaker's question or talking about his/her own feelings, asks the other person the same question or about similar feelings, thereby actively inviting the other person to participate in the topic. For example, after answering a question regarding her feelings about Jinan (turn 20 in Excerpt 2), Ai asked the same question back to Ti. Turns 22 and 29 in Excerpt 1 provide another example.

Helping the other person complete the meaning of his/her expression may occur regardless of whether the other person has encountered difficulties. In Excerpt 4, for example, Ya (turn 55) paused slightly when trying to express the benefits of education in

Germany, so Men made a statement intended to complete what Ya wanted to say (in turn 56). Elsewhere in Excerpt 4, on the other hand, Men (turn 36) helped Ya to express the idea that "Shenzhen has good job opportunities" despite the fact that Ya was not having any trouble articulating this idea. Both of these examples indicate that Men's L2 pragmatic competence in interaction was strong.

Lastly, in-depth topic development can refer to either the sentence level or the discourse level. In the four excerpts from the high-competence group, all parties exhibited the ability to critically develop their own and others' topics, that is, to maintain independent thinking and personal attitudes and perspectives on various issues rather than merely accepting each other's views and opinions. For example, in turn 21 of Excerpt 2, Ti stated that she did not like Jinan as much like Ai did, because it had no subway, to which Ai responded (in turns 22 and 24) that this was not a big problem, as her entire country had no subways, but its people still lived well.

In the first three excerpts from the high-competence group, topic development was mainly at the sentence level (e.g., turn 3 in Excerpt 1; turns 22 and 24 in Excerpt 2; and turns 25 and 27 in Excerpt 3). In Excerpt 4, however, it extended to the discourse level: with multiple sub-topics being developed very comprehensively. Ya initiated the topic "Where do I want to go after graduation?" in turn one, and both partners shared more or less equally in its development and extension over the next 65 turns (turns 2-66), in a manner very similar to natural conversation. Both partners demonstrated an ability to develop not only their own sub-topics, but also each other's. In Men's first turn (turn 2), he developed and talked about: (1) the difficulty of answering the question; (2) the fact that he wanted to work for the United Nations; (3) that he was unsure where he would

work; and (4) the importance he assigned to safety. Ya further explored Men's development in turn 3, and after Men's detailed response (turn 4), Ya began to contribute detailed and in-depth views using discourse connectors, as also mentioned above in the discussion of language use. He talked about his future career's "working aspects" (turns 7, 9, 11, 13, 15 and 17), "environmental aspects" (turns 19, 21 and 23), and "educational aspects" (turns 25 and 27), as well as his "after graduation short-term arrangements" (turns 29, 31, 33, 35, 37, 39, 41, 43, 45, 47 and 49), "long-term arrangements after having children" (turns 51, 53 and 55), "food safety" (57, 59 and 61), and "specific places he wants to go" (turns 63 and 65). Both parties exhibited a high level of language competence and L2 pragmatic competence in interaction, demonstrating a "high degree of participation" (Tannen,1982) and good "interspeaker coordination" (Hutchby & Wooffitt, 1998).

Topic transition can be usefully conceived of as a stepwise process (Sacks, 1992), in which links with previous topics can be established using connective words, and new aspects or summaries of the original topic can serve as pivots to new topics. In Excerpt 1, turn 21, "learning language" can be regarded as a pivot, as it connected both back to turn 20 ("to come to China to learn the language") and to turn 21's new question regarding how long Zeng has been studying. Elsewhere in Excerpt 1, Zeng asked Ban about "professionals" in turn 26, echoing turn 20, and forming a semantic relationship with language learning to achieve natural transition. In addition, in Excerpt 4, Men summarized the views previously expressed by Ya in turn 66 by proposing that his own life and Ya's were totally different, thus indicating the end of this topic, and preparing for the start of a new one.

Not all topics will have an obvious termination. Sometimes, assessment tokens such as “great” and “good” can be signs of an intention to terminate a topic. For example, in Excerpt 3, Wang used turns 28 and 30 to make jokes, compliment Huang, and express her acceptance of the invitation. She then marked the end of the invitation as a topic, and appeared to be on the verge of launching into a new topic, using a stepwise approach to topic change that could also be considered a pre-closing sequence. Another example was provided by turn 66 in Excerpt 4, in which (as noted above) the speaker prepared for the new topic by summarizing the previous one. In other words, all the high-competence candidates exhibited a strong ability to close topics smoothly rather than abruptly.

Discourse analysis: middle-competence group

Language use. The four excerpts randomly selected from the middle-competence test-takers demonstrated that they had a language reserve adequate to daily communication and basic L2 pragmatic competence in interaction. Although on the whole their communication and understanding was unimpaired, their language included various errors and inaccuracies, and sometimes was needlessly complicated. Such problems included all aspects of pronunciation, vocabulary and grammar. The pronunciation problems included both inaccurate tones and wrong initials, such as “*新* (new)” (turn 25 in Excerpt 5). Common words were sometimes misused, e.g., “*可是* (but)” which should have been “*现在* (now)” (turn 10 in Excerpt 6), and “*有名* (famous)” which should have been “*流行* (popular)” (turn 19 in Excerpt 8). Even in some very common sentence patterns, grammar mistakes were made, including omission of the particle “*的*” when using the structure “*是...的* (indicates judgement)” (turn 11 in Excerpt 5), and incorrect use of “*对...有什么看法* (opinions toward)” to mean “*对...怎么样* (how well

Excerpt 5: Mid-competence pairs
Knowing each other in a Chinese corner (part A). a: A, b: Ban

1 a 欸，朋友你好。

Hi, friend, hello.

2 b 你好。

Hello.

3 a 呃，我叫艾米力，我是阿富汗人，我刚刚来中国，你可以介绍你的名字。

My name is Mili Ai. I'm Afghan. I just came to China. You can introduce your name.

4 b 可以啦，认识你很高兴。我叫班达，我也，我来自日本, 呃，我叫啊，“piao gen”（.），呃，认识你很高兴。

Yes, I am very happy to meet you. My name is Da Ban. I am from Japan. My name is "piao gen", oh, I'm very glad to meet you.

5 a 呃，我也是，你可以介绍，你可以告诉我你为什么来中国。

Me too, you can introduce. You can tell me why you came to China.

6 b（.）我呀，我在，呃，日本的时候一直想来中国,有一天来中国学习，哎，呃，嗯，汉语，因为现在，嗯，在世界上汉语是最重要的，有好多人，呃，用这个语言，呃，说话，呃，所以，呃，我也想有一天来到中国然后开始学习汉语。嗯，你呢，你为什么, 呃，来中国学习汉语，你的专业是中文吗？

As for me, I was in, uh, I have always wanted to come to China when I was in Japan. I wanted to go to China one day to study Chinese. Well, uh, hm, Chinese, because now, hm, Chinese is the most important language in the world. There are so many people, uh, use this language, uh, to talk. Uh, so, uh, I also want to come to China one day and start learning Chinese. Hm, how about you, why did you come to China to learn Chinese, are you majored in Chinese?

7 a 是的,我的专业是（.）中文，我是哈理工的学生。

Yes, my major is Chinese, and I am a student at Harbin Institute of Technology.

8 b 啊：

Ah:

9 a 啊，[我

Ah, I

10 b [太好了！

Great!

11 a 因为我的爸爸<是做生意>，所以我想学习中文

Because my dad is doing business, so I want to learn Chinese

12 b 嗯

Hm

13 a 毕业之后我想, 呃，帮助我的爸爸，还有, 呃：，做生意，还有要是可以，我就‘投’资在中国。

After graduation, I thought, help my dad do business. And if it's okay, I'll invest in China.

14 b 啊，[太好了！

Ah, great!

15 a [你

You

16 b 你是哈理工的, <哈理工的学的>？

You study at Harbin Institute of Technology?

17 a 是的=

Yes

18 b =我也是

Me too

19 a 你也是?

Are you?

20 b 对=

Right

21 a =[我怎么

How come I

22 b [一个月, 一个月以前来到了=

A month, arrived a month ago

23 a =你刚刚[来的?

You have just come

24 b [你是‘新’生吗?

Are you a new student?

25 a 我不是‘新’生，<而是我是大二>

I am not a 'new' student, but rather I'm a sophomore.

26 b 大二的。

Sophomore.

27 a 对啊

Right

28 b [那太好了！

That's great!

29 a [你的中文, 很不错=

Your Chinese is rather good.

30 b =还行,……

Not too bad.

somebody treats others)" (turn 1 in Excerpt 6). There were some blended sentences. For example, "*在学工商管理硕士生* (study MBA student)" mixed the two common sentence patterns "*在学工商管理* (studying for my MBA)" and "*是工商管理硕士生* (MBA student)" (turn 7 in Excerpt 6).

Sometimes, complex sentences were misused: for example, "*刚刚 2009 就开始了汉语* (since 2009 started Chinese)" should have been "*直到 2009 才开始学习汉语* (until 2009, when I had just begun to learn Chinese)" (turn 18 in Excerpt 8). There were also word-order problems, such as "*中国教师很多去了教汉语* (Chinese teachers are going there a lot)" (turn 16 in Excerpt 8); incomplete sentences like "*我们的以前同学* (our previous classmates)" (turn 6 in Excerpt 7); and unclear sentences, such as turn 14 in Excerpt 6 and turn 7 in Excerpt 7.

Situation response. The middle- competence students' in-test discourse indicated that most of them were able to grasp the required situations and could complete the test's basic tasks. Sometimes, however, they were unable to perceive differences in context changes and/or to respond to them appropriately. For example, in Excerpt 5, A and Ban were designated as people who did not know each other: that is, the social distance between the two was relatively large, and they should have been as courteous as possible when communicating. However, in turns 3 and 5, A rudely used affirmative sentences in the place of questions, telling Ban "*你可以介绍你的名字* (you may state your name)" and "*你可以告诉我你为什么来中国* (you may tell me your reasons for coming to China)". Conversely, in Excerpt 7, Xie and Fu were instructed to portray good friends, yet used turns 1 and 2 to say "hello" in a polite manner, indicating that they were not

Excerpt 6: Mid-competence pairs
Knowing each other in a Chinese corner (part B). t: Tian, b: Bai

1	t 那<你对中国怎么样>? *What do you think about China?*	2	b 还好啊，但是有点儿不习惯，不习, 不习惯。呃，有很多，不识，不认识的东西。（.）然后气候也没有那么好，天气也有点不同，跟我国家不一样，冬天太冷了，夏天太热了。<“我”觉得怎么>，你呢？你觉得中国怎么样？ *It's okay, but I'm not used to it, not used, not used to. There are many things that I don't know. Then the climate is not so good and the weather is a bit different. Unlike my country, winter is too cold and summer is too hot. Ho[w do I think? What about you? What do you think about China?*
3	t 我，我觉得还可以，因为在武汉，跟我们：城市一样，差不多，空，呃，天气一样，就现在：= *I think it is okay, because in Wuhan, like our city, almost, air, uh, the weather is the same. Right now*	4	b =很冷 *Very cold.*
5	t 在我们国家也很冷，呃：（.）所以我喜欢，我没问题。 *It is also very cold in our country. So I like it. I have no problem.*	6	b (.) 嗯:(.) 你来中国, 呃，学习什么专业呢？ *Hm. You came to China, what do you study?*
7	t 哦，我<在学工商管理硕士生>。 *Oh, I'm an MBA student.*	8	b 啊：，那我才学了一年汉语，我的水平还没有那么高，所以我觉得先不要学习什么专业，因为我觉得我水平还是不够。 *Well, I only learned Chinese for a year. My level is not that high, so I don't think I should study any majors first because I think my proficiency is not enough.*
9	t 啊，这样“吧”，<那我不一>，不跟你，不一样，因为我们用英语，呃，讲课，所以我的专业用英语，中文我学了，为了生活。 *Ah. In this way, then I am differ[rent, not like you, different from you, because we use English, uh, to lecture, so my major is taught in English. I've Chinese learned for life.*	10	b （.），嗯，那就好，那*，我以前的专业就是教育，教育。啊，我学了一年汉‘语’，然后通过学习考试，然后今年我应该学习我的专业。但是，嗯，我开始学习我的专业，我觉得太难了，一年的，一年学习的汉语不够，所以我换成汉语，“可是”在学习汉语。 *Well, that's good, my previous major was education. I studied Chinese for a year and then passed the exam. Then this year I should study my major. But I started to study my major. I find it too difficult. I don't have enough Chinese to study in a year. So I changed to Chinese, but I was learning Chinese.*
11	t 啊，你你，在住，你住在学校里面还是外面? *Do you live in school or outside?*	12	b 嗯，我住在学校里面，在宿舍里，你呢? *Well, I live in the school, in the dormitory, how about you?*
13	t 嗯，你们宿舍方便吗? *Well, is your dormitory convenient?*	14	b （.）还行吧，很*，我喜欢。 *OK, it's very, I like it.*

fully aware of how to initiate a topic naturally in such circumstances. In addition, in turns 12 and 14, Xie gave completely contradictory replies: first, apparently forgetting the instruction that his coffee machine was broken and he could not lend to Fu, and then remembering this, yet failing to make any reasonable corrections or transition.

Turn-taking organization. The middle- competence students had few if any big gaps between their turns, but their turn speed was not fast, and apart from in Excerpt 5, there were relatively few cooperative overlaps/latches between turns. Speaker shift was mainly achieved via other-selection, as few appeared to want to actively take the floor using the self-selection approach.

The greater number of overlaps/latches in Excerpt 5 may have resulted from it having been centered on common life situations that both speakers were familiar with. Yet, in turns 13-24, this amounted to both parties constantly interrupting each other and always failing to completely finish their own turns. This had a strongly negative influence on the effectiveness of the interaction. The other three excerpts, on the other hand, had more loosely connected turns, marked by pauses that varied in frequency but were always brief; examples included turns 6, 10 and 14 in Excerpt 5; turns 9 and 12 in Excerpt 7; and turn 2 and 10 in Excerpt 8.

In addition, the speaker shift was mainly achieved through the most common way as the current speaker selecting the next one. Few listeners actively selected themselves to be the next speakers. It could be seen that candidates at this level were more likely to wait for the speaker to pass the speakership to him/her, but he/she lacks the enthusiasm of actively expressing his/her views on certain events and information.

Sequence organization. For the most part, the eight middle- competence learners in Excerpts 5-8 understood the meaning of the previous turns, but there were some cases where they misunderstood and/or failed to provide appropriate responses. Among all types of response tokens, confirmation and repetition – that is, the two tokens that do not necessarily express understanding – were used most often, while assessment tokens were

Excerpt 7: Mid-competence pairs
Inviting a friend to a Christmas party and borrowing a coffee machine. f: Fu, x: Xie

1	f 呃，你好，你听说我们，呃, 那个十二月二十四日有圣诞节晚会？ *Hello, you heard about our Christmas party on December 24th?*	2	x 你好，我听说了，但是我不知道(.) 谁参加，有什么活动呢？ *Hello, I heard, but I don't know. Who is involved and what activities are there?*
3	f 嗯，我想，我想请你参加这个活动。 *Well, I think, I would like to invite you to participate in this event.*	4	x 你可以给我介绍有谁参加，有什么活动？ *Can you tell me who is there and what activities are there?*
5	f 呃，可以(.) 那个，“十二个月”二十四号周四下午四点半(.) 呃，会在晚上十点前完，呃，那个，101 教室，明白？ *December,24, Thursday, 4:30 p.m. Yes, it will be finished by 10:00 pm, in 101 Classroom, understand?*	6	x 明白，<我们的以前同学> *Understood, our previous classmate*
7	f *	8	x 嗯，明白 *Hm, understood*
9	f（.）啊，我有一个问题。 *Ah, I have a question.*	10	x 嗯 *Hm*
11	f 我，我的咖啡机坏了，正好我听说~你有一个，你能不能借给我？ *My coffee machine is broken, I heard that ~ you have one, can you lend me?*	12	x (.) 哎呀，我(.)可以借给你，但是我不能来参加这个活动= *I can lend it to you, but I cannot go to this event.*
13	f =[为什么？ *Why?*	14	x [我非常想，但是很抱歉，我七点，晚上有别的活动，约会，所以我不能(.)去你的。还有一个不好的消息，我的咖啡机坏了。 *I want to go very much, but I'm sorry, I at seven o'clock, have other activities at night, dating, so I can't go to your party. Another bad news is that my coffee machine is broken.*
15	f 呵呵~，哎呀，真糟糕，哎呀！怎么办呢？…… *It's bad, oh! What should I do?*		

rare. In terms of preferences, accounts and mitigative devices were used occasionally, but all pre-sequences were simple, and no multi-turn pre-sequence or post-sequences were used.

In most cases, this group of learners could understand the meaning of the previous turn, and communicate without hindrance. However, listeners sometimes failed to receive signals or did not understand their meanings. For example, Da (in turns 2, 4, 6 and 8 in Excerpt 8) expressed his opinion in great detail and showed a good ability to develop the topic, but his interaction was ineffective because it did not actually constitute a response to the question Wu had asked him in turn 1, that is, “What's the difference between education methods in your country [Armenia] and in China?” Da replied by

talking about the difficulties he had encountered in learning Chinese and adjustments to his mindset, making it fairly clear that he had not understood Wu's meaning. However, Wu apparently did not understand Da's meaning either, as he did not take the initiative to repair Da's misunderstanding, instead just saying, "*知道了* (I knew)" (turn 9). Thus, Excerpt 8 was an invalid interaction from turn 1 to turn 9. Another example of an unanswered question occurred in Excerpt 6, when Bai asked whether Tian lived on or off campus. Rather than responding, Tian just continued to ask for new information.

The middle- competence participants' response tokens were mainly of three kinds: confirmation, repetition, and assessment. Of these, confirmation was the most common, and included "*嗯* (hm)" (turn 12 in Excerpt 5), "*哦*(oh)" (turn 7 in Excerpt 6), "*明白* (understood)" (turn 6 in Excerpt 7), and "*是，是* (yes)" (turn 5 in Excerpt 8). Some of these test-takers were also willing to repeat what others had said: for example, "*大二* (sophomore)" (turn 26 in Excerpt 5), and "*教育方式* (educational mode)" (turn 2 in Excerpt 8). It can be seen that, while this group of candidates could use these two methods of reply accurately in terms of their placement within the discourse, they did not do so in a way that clearly demonstrated an understanding of their partners' utterances. Indeed, Da's discourse taken as a whole showed conclusively that, despite his repetition of Wu's question's keyword "educational style", he did not actually understand Wu's meaning. In addition, some assessment tokens were not used properly. For example, Ban often used the simple assessment "*太好了* (great)" (e.g., in turns 10 and 14 in Excerpt 5) but sometimes misused it (e.g., in turn 28).

The middle- competence candidates did not use preference structures often: only two of the four excerpts contained any, and of those two, one (Excerpt 5) included just a

Excerpt 8: Mid-competence pairs
Situational discussion about country comparison. w: Wu, d: Da

1 w 嗯，那在亚美尼亚，跟中国对比（.）从教育方式有什么不同呢？
Well, in Armenia, how educational style is different from Chinese educational style?

2 d (.) 啊，教育方式。
Ah, educational style.

3 w 是，是，是
Yes.

4 d 嗯（.）有很多不同的方面，比如在我的国家<先最最>，第一个最重要的，<我们学习是亚，亚美尼亚语>。
There are many different aspects, such as in my country the first and most important thing is that we study Armenian.

5 w 是，是
Yes

6 d <这就是我们学习中文>，所以我们有中文学习，这是对我，呃，比较难，比较“重”，因为我，怎么说，我知道的词不太多，你知道吗?
This is we are learning Chinese, so we have Chinese studies. This is for me, uh, more difficult, more heavy. Because I, how to say, I don't know too many words, you know?

7 w 嗯
Hm

8 d 但是我不放弃，我很“放松”，就因为这不是我的母语，我来这儿学习汉语，所以我不要那么紧张，你知道吗?
But I don't give up. I'm relaxed, because it's not my mother tongue. I'm here to learn Chinese, so I'm not so nervous. Do you know?

9 w 嗯哼，知道了。
Uh-huh, I got it.

10 d （.）你呢？
How about you?

11 w 在印尼跟中国的教育方式也有区别的，但是区别不大。
There is also a difference in the way of education between Indonesia and China, but the difference is not significant.

12 d [嗯
Hm

13 w [因为可能我觉得印尼也是在东南亚的一圈。
Because I think Indonesia is also a circle in Southeast Asia.

14 d 嗯(.) 对，因为<你看你们的国家在>，印尼，在泰国，在越南*早已经开始学习汉语，你知道吗？
Yes, because Indonesia, Thailand, and Vietnam have already started learning Chinese very early. Did you know?

15 w 嗯
Hm

16 d 在，呃，在这些的国家，呃，（.）中国已经很，呃，早以前，<中国教师很多去了教汉语>，你知道吗?
In these countries very, long ago, Chinese teachers went to teach Chines, do you know?

17 w 是，是
Yes.

18 d 但在亚美尼亚，呃，就是（.）刚刚 2009 “就”开始了汉语，越来越“有名”，这样。
But in Armenia, we just started Chinese in 2009, and it becomes more and more "famous".

19 w 哦
Oh.

single compliment (turn 29). The rest were all in Excerpt 7, in which Fu used a simple pre-sequence (turn 1) ask if the other party had heard about the Christmas party, thus laying a foundation for one of the conversation's two main topics. Xie responded by asking about the specific activities that would be involved. Next, in turn 3, Fu issued an

official invitation in the form of an assertive sentence with a modal verb "想 (want to)", that is, "*我想请你参加这个活动* (I would like to invite you to participate in this event)", but did not provide any relevant information about the party, including the information that Xie had asked for. Then, before Xie had replied, Fu issued another request, in a manner that seemed quite rushed. Having explained that his coffee machine was broken and saying, "I heard that you have one", Fu in turn 11 used the mitigative devices "*能不能* (can or not)" in an interrogative sentence to request it. Xie responded with willingness ("I really want to"), but then apologized ("I'm sorry") and gave an explanation involving "another event" before the real turn-down action ("can't go", turn 14). Suddenly, Xie then added that he could not lend Fu the coffee machine, thus contradicting turn 12, and did not provide a timely post-sequence by way of remedying the fact that he had just promised to lend coffee machine a few moments before.

Topic management. In terms of topic initiation, the middle- competence candidates' performance was unexceptional; there were no abrupt topic starts, but nor were efforts made to establish connections between topics and the prevailing situations. In terms of topic development, this group exhibited an ability to develop their own and other's turns, even at the discourse level in some cases. However, the topics of discussion were mainly descriptive or introductory, and most members of this group did not use multiple turns to develop their own opinions on a topic. When it came to topic transitions, these test-takers' use of the step-wise approach was poor, with some sudden shifts to new topics, and a frequent lack of signs that a topic had ended.

In terms of topic initiation, as mentioned above, there was little use of the test-scenarios' environments, though most of the time the topic arose in a natural-seeming

rather than an abrupt way (e.g., turn 1 in Excerpt 7 and turn 1 in Excerpt 8. As for topic development, further inquiry, asking back, co-operation in turn completion, and in-depth development were all reflected in the middle- competence group's discourse, but their frequency was relatively low. After embarking on a new topic, the middle- competence candidates only rarely continued to ask questions (e.g., turn 19 in Excerpt 5). Some test-takers used asking back to invite their partners to continue discussing a topic, for example, in turn 6 in Excerpt 5, and turn 10 in Excerpt 8. Occasionally, someone helped to complete his/her partner's turn, such as turn 4 in Excerpt 6.

Candidates at this level also showed an ability to develop topics to discourse level: e.g., turns 6, 11 and 13 in Excerpt 5; turns 2 and 10 in Excerpt 6; and turns 4, 6, 8 14, 16, and 18 in Excerpt 8). This ability applied to both their own turns (e.g., turn 6 in Excerpt 5 and turn 2 in Excerpt 6) and to others' turns (e.g., turn 8 in Excerpt 6, and turns 14, 16, and 18 in Excerpt 8). For example, in turn 10 in Excerpt 6, after Tian (turn 9) mentioned that his coursework was taught in English, whereas learning Chinese was for use in daily life, Bai continued to develop this topic to discourse level in turn 10 by speaking of his own situation.

Across the four middle- competence excerpts, it can be seen that topic shifting seldom followed the stepwise approach. In Excerpt 6's turn 6, Bai used China as a pivot, connecting the topic of impressions of China with that of studying in China; but more often, there was a lack of transitions at this level. For example, in Excerpt 5's turn 11, after Bai talked about a broad swathe of his own experience, Tian did not comment or extend this, instead moving directly to a wholly new and unrelated topic.

Lastly, in the sphere of topic termination, the middle-competence group did not use assessment tokens or pre-closing sequences to terminate current topics and shift to new ones. This was consistent with the suddenness of their topic shifts that was already noted.

Discourse Analysis: low-competence Group

Language use. The low-competence candidates had insufficient language reserves for the paired interactive tasks, and often encountered obstacles in their expression and communication. Thus, it can be said that their L2 pragmatic competence in interaction was limited. Language errors were very common, even when they were delivering quite simple content. Long pauses were also noticeable among all four pairs of students from this group, but especially in Excerpts 10-12 (e.g., turn 12 in Excerpt 10; turn 12 in Excerpt 11; and turn 9 in Excerpt 12). In addition, the content delivered by these candidates was extremely basic, its accuracy was low, and errors were very common.

All the problems of pronunciation, vocabulary, and grammar that were identifiable in the middle- competence students' excerpts appeared in the lower-level students' excerpts as well, but with an even higher frequency. With regard to pronunciation in particular, some had difficulty pronouncing the initials *sh*-, *zh*-, and *q*-, such as in "*是* (is)" (turn 20 in Excerpt 9; turn 7 in Excerpt 11); "*这* (this)" (turn 2 in Excerpt 10); "*去* (go)" (turn 15 in Excerpt 10); and "*说* (speak)" (turn 15 in Excerpt 11). Their vocabulary mistakes included selection of the wrong elementary words, for instance, "*哪个* (which)" rather than "什么 (what)" in turn 5, Excerpt 9; "*管理* (management)" rather than "*经理* (manager)" in turn 22, Excerpt 9; "*带* (bring)" rather

Excerpt 9: Low-competence pairs
Knowing each other in a Chinese corner (part A). g: Ge, a: A

1	g 你好。 *Hello.*	2	a 你好。 *Hello.*
3	g 我叫格罗斯，你呢? *My name is Ge Luosi, how about you?*	4	a 啊，我叫阿道芙。 *Ah, my name is A Daofu.*
5	g 你是“什么”国家(.)的？ *Which country are you from?*	6	a 我是坦桑尼亚人。 *I am a Tanzanian.*
7	g 好“啦”，[我是 *Okay, I am*	8	a [你呢? *How about you?*
9	g 我是贝宁人。 *I am from Benin.*	10	a 啊 *Ah.*
11	g 啊，你来中国，<什么时候>? *When did you come to China?*	12	a 啊，<去年来中国>。 *Ah, last Year came to China.*
13	g 啊，去年来中国。 *Ah, came to China last year.*	14	a 对 *Right*
15	g 中国好吗？ *How is China?*	16	a 中国好。 *China is good.*
17	g 啊，你，所以你喜欢中国？ *Ah, you, so you like China?*	18	a 我喜欢中国，对，你呢？ *I like China, right, how about you?*
19	g 啊，我也喜欢中国，中国好，中国，我喜欢，我喜欢，嗯，你（.）现在<做‘什’么在中国>? *Ah, I also like China. China is so good. China, I like it. What do you do in China?*	20	a 嗯，什，我‘是’学生。 *Hm, I am a student.*
21	g 嗯 *Hm*	22	a 我学习“经理”, 你呢？ *I learn management, how about you?*
23	g 啊，我也是学，生，我学习，啧，国际贸易。 *I am also a student. I study international trade.*	24	a 国际贸易，啊，很好。 *International trade, ah, very good.*
25	g 对 *Right.*		

than “*接待* (host)” in turn 14, Excerpt 11; and “*对* (right)” rather than “*有* (have)” in turn 6, Excerpt 12. The many grammatical problems in these excerpts involved misuse of the preposition “*在* (in)” (turn 11, Excerpt 9); basic structure, e.g., “*是...的* (indicating judgement)” (turn 12, Excerpt 9); word-order in adverbials relating to place (e.g., in turn 19, Excerpt 9); random use of the possessive case particle “的” (e.g., in turn 7, Excerpt 10); unnatural-seeming utterances like “*你听说得对* (what you heard is right)” (turn 16, Excerpt 11); mistakes in using modal verbs that indicate possibility, such as “*可以* (can)”

Excerpt 10: Low-competence pairs
Knowing each other in a Chinese corner (part B). y: Ye, l:Li

1	y 那(.)你对中文，中国的印象怎么样? *How do you feel about Chinese language and China?*	2	l (0.3) 很好，<我觉得中国很好的地方>，在‘这’里我可以好好学习。 *Very good. I think China is a good place. I can study well here.*
3	y 嗯，也（.）你：有没有中国的朋友? *Do you have any Chinese friends?*	4	l（.）我还没有啊(.)你可以当我的新‘朋’友吗? *I haven't got one. Can you be my new friend?*
5	y 当然可以，你是第一次来中国吗? *Of course, is it the first you came to China?*	6	l 对 *Right.*
7	y 那（.）有哪些地方你很喜欢，你很<喜欢哪些的地方>? *Where do you like?*	8	l (0.3)嗯（0.5） *Hm*
9	y 你喜欢中国哪些“方便”? *What aspect of China do you like?*	10	l（.）呃(.) *Hm*
11	y 方面 *Aspect*	12	l(.)第，第一个是我很喜欢中国的习惯，在这‘里’很多人很努力学习（0.4）文化也很好（.），呃，你觉得怎么样呢? *First, I like Chinese habits very much. Many people study very hard here. The culture is also very good. Oh, what do you think?*
13	y 我也这么觉得咱们的（.），咱们的文化也很有意思，我觉得学中文也很有意思。 *I also think that our culture is also very interesting, I think learning Chinese is also very interesting.*	14	（1.3）
15	y 那我们‘去’跟，那我们去找(.)交新朋友吧。 *Then let's go to find new friends.*		

employed to mean “*可能* (be likely to)” (turn 18, Excerpt 12); and erroneously using “*很* (very)” as a predicate (turn 23 in, Excerpt 12).

Situation response. Excerpts 9-12 also showed that the low- competence test-takers were not able to complete their tasks as required, or respond appropriately to specific situations when dealing with slightly more complex context tasks. In Excerpt 11, Tang wanted to invite friends to attend a Christmas party and to borrow a coffee machine from Da. According to the instructions for this scenario, Da had other things to do on the day of the Christmas party, and her coffee machine was broken, and thus she should have been – at best – hesitant both about accepting the invitation and lending Tang the machine. However, Da cheerfully accepted the invitation in turn 6, and in turn 18, accepted Tang’s coffee-machine request.

Excerpt 11: Low-competence pairs
Inviting a friend to a Christmas party and borrowing a coffee machine. t: Fu, d: Da

1	t 嗯（0.3）圣诞节快到了。 *Christmas is coming.*	2	d 嗯，[对。 *Hm, right.*
3	t [嗯，我（.）我（.）我和我的:同学有一个，啊，圣诞（.）晚会，啊，我邀请你参加。 *Well. I and my classmates will have a Christmas party. I invite you to attend.*	4	d 哦，真的吗？啊，你太可爱了,谢谢。 *Oh, really. Ah, you are too cute, thank you.*
5	t 呵~ *Ho*	6	d 我, 当然来看一看。 *Of course, I will take a look.*
7	t 好的，啊，那个我们的（.）晚会是（0.5）呃，‘是’（0.5）呃，今天。 *Okay, ah, that our party is today.*	8	d 今天，啊。 *Today, ah.*
9	t 哈哈哈~ *Haha*	10	d 啊，这么快，呵呵~，几点开始? *Ah, so fast, hehe. What time does it start?*
11	t 今天周四~，今天是周四，啊，是四点半。 *Today is Thursday, ah, it's half past four.*	12	d 四点，啊：啊，我（.）三点半，呃（.）下班，我希望我（.）可以（.）来，我（.）希望我（.）来。呵呵~ *Four o'clock. I will finish my work at 3:30 p.m. I wish I could come.*
13	t 呵呵~，好的，啊（0.3）[呃 *Hoho, good.*	14	d [但是，呃（.） 我(.)可能，我应该，呃(.)带呃，“接待” 什么（.）吃的东西吗？ *However, what should I bring? Something to eat?*
15	t 啊，我听‘说’了，你有一个（.）咖啡机。 *Ah, I heard you have a coffee machine.*	16	d <你听说得对>。 *What you heard is correct.*
17	t 呵呵~（.）你可以把你的咖啡机- *You can (lend) your coffee machine-*	18	d 啊! 但是咖啡机, 这么*重*，我是女孩儿。 *Ah! But coffee machine, so heavy, I'm a girl.*
19	t 哈哈~ *Haha*	20	d 我没有[这么结实 *I'm not that strong.*
21	t [我, 我可以来你的房间帮你。 *I can come to your room to help you.*	22	d *哦*，哈哈哈~ *Oh, haha*

In the same excerpt 11, there was a long pause in turn 1 before Tang stated that “Christmas is coming”, and another in turn 7 before he said “Christmas. The party is today.” Taken together, these pauses indicated that Tang had been unable to figure out how to express the specific timing of the party at the outset of the conversation, and thus failed to respond to the situation as required.

Turn-taking organization. The low- competence candidates’ turn-taking was neither fast nor smooth, and featured very few cooperative overlaps/latches, with Excerpts 9 and 10 having none at all. Other-selection was used in almost every case, and self-selection was very rare. Occasionally, the speaker would signal that it was the

listener's turn to speak, but the listener could not take the floor, so the speaker had to continue talking: as occurred in Excerpt 10, when Ye designated Li as the next speaker in turn 13, but in turn 14, Li was not able to contribute anything. After a long pause, Ye had to continue to speak at turn 15. This suggests that the initiative of the low- competence test-takers to participate in these interactions was not high.

Unsurprisingly, there were frequent pauses in the selected test discourse at this competence level, sometimes short (e.g., turns 4 and 10 in Excerpt 10; turns 8 and 20 in Excerpt 12), and sometimes very long (e.g., turns 2 and 8 in Excerpt 10; turns 6 and 15 in Excerpt 12). Notably, Excerpt 10 contained an extraordinarily long pause (turn 14). This is not to suggest, however, that interruptions never occurred (see, for example, turn 17 in Excerpt 11 and turn 11 in Excerpt 12).

In addition, regarding the selection approach of a new speaker, candidates at this level did not actively take the initiative to speak. Basically, they were designated only by the previous speaker. Moreover, the current speaker passed the speakership to the listener, however, in some situations, the listener did not have the ability to speak anything, and the current speaker had no choice but to continue to speak. As in excerpt 10, Ye designated Li as the next speaker in turn 13, and then in turn 14, Li was not able to contribute to the turn. After a pause for a long time, Ye had to continue to speak at turn 15. Through the method of speaker shift, it can also be seen that the initiative of the low-level test takers to participate in the interaction was not high.

Sequence organization. The sampled low- competence candidates sometimes encountered serious obstacles in understanding the meanings of previous turns, to the point that normal interaction was completely derailed. Their use of response tokens

Excerpt 12: Low-competence pairs
Situational discussion about hobby. z: Zhou, c: Cui

1	z 嗯（.）你呢，你每天（.）打乒乓球吗? *How about you, do you play table tennis every day?*	2	c 不是，一般周末的时候 *No, usually on the weekend.*
3	z [嗯： *Hm*	4	c [打乒乓球 *Play table tennis.*
5	z 嗯 *Hm*	6	c (0.4) 乒乓球“有”身体很好。 *(Playing) table tennis is good for body health.*
7	z （.）嗯 *Hm*	8	c （.）读书呢？我觉得（.）读书没有动作，对身体不好。 *Reading? I think reading does not move and is not good for the body.*
9	z 不是，啊（.）啊，读书，读书的时候（.）呃，我不做，不不（.）动作，可是（.）这时候我（.）心‘里’，呃，心里很高兴，还有呃（.）气氛很好。 *No, ah, ah, reading, while reading. Uh , I don't do, don't, don't do actions. But, at this time, inside of my heart, I am, uh, very happy. In addition, uh, the atmosphere is very good.*	10	c *嗯：* *Hm*
11	z （.）所以对精神（.）精神健康- *So for mental health -*	12	c 啊 *Ah*
13	z 很好。 i*s good.*	14	c 嗯 *Hm*
15	z (0.3)可是我觉得（.）打乒乓球（.）呃（.）这个（.）呃（.）这个没有意思。 *But I think playing table tennis, this is not interesting.*	16	c 啊：有意思，呃，虽然做运动的时候， *Ah, interesting, uh, although when he dose exercises,*
17	z [嗯 *Hm*	18	c [有, 做运动的时候（.），呃，<可以危险>。 *Yes, when you do exercise, uh, may be dangerous.*
19	z 嗯 *Hm*	20	c （.）但是我（.）锻炼身体，所以冬天的时候，<还,还没感冒了>。 *But I do exercises, so in the winter, I haven't caught a cold yet.*
21	z 啊： *Ah*	22	c 啊 *Ah*
23	z <这是好>，[很好 *This is good, very good.*	24	c [嗯 *Hm*
25	z 很好，运动的好处，嗯。 *Well, the benefits of sports. Hm.*		

included confirmation, repetition, and simple assessment (e.g., “*很好* [very good]” in turn 24, Excerpt 9 and turn 25, Excerpt 12), but apart from confirmation, their frequency of appearance was low. In terms of preference organization, some candidates avoided using dispreferred structures even when they were required by the task, while others used dispreferred structures without any turn-shape changes.

A typical example of failure to understand the meaning of a previous turn occurred in Excerpt 10, when Li had three turns (8, 10, and 14) that did not correspond to

the meaning of Ye's preceding statements, thus rendering effective interaction impossible. This prompted Ye to end the interaction prematurely, in turn 15. Preference organization was mainly shown in Excerpts 11 and 12. The former used two actions: invitation and request. For the invitation, Tang directly used assertive sentences and did not employ any mitigative devices, saying "*我邀请你参加* (I invite you to join)" (turn 3). When asking to borrow Da's coffee machine, Da interrupted him. Da was supposed to engage in two turn-down actions, but she avoided both of them, instead choosing to simply accept the invitation and agree to lend Tang the machine. In Excerpt 12, Zhou and Cui both used disagreement actions. Rather than using any devices to change the turn-shapes, however, the two participants just directly opposed each other. For example, Cui mentioned in turn 8 that reading books was not good for the body, to which Zhou simply said "*不是* (no)" in turn 9. Zhou then said (in turn 15) that playing table tennis was not interesting, in response to Cui's revelation that she habitually played it (in turns 4 and 6). In other words, the interaction was blunt and the two partners seemed to take no notice of face-saving issues.

Topic management. In terms of topic initiation, the low- competence candidates did not have prominent problems. However, they did exhibit serious problems in developing topics, and such topic development as they did engage in was limited to the lexical and sentence levels. Topic shifts were fast and not conducted in a stepwise fashion, and there were no signs of topics' termination.

In the sphere of topic development specifically, these test-takers used a simple returning question token, "*你呢* (how about you)" (e.g., turn 8 in Excerpt 9), and only occasionally explored somewhat beyond it, such as in turn 17, Excerpt 9: "*所以你喜欢中*

国 (So you like China)". All four excerpts lacked in-depth development, and topic development above the sentence level simply did not occur.

Because the low- competence students did not engage in in-depth topic development, their topics were numerous and shifted rapidly. The questions they asked each other tended not to make use of connectives. In Excerpt 9, for example, Ge asked "Where are you from?", "When did you come to China?" and "How do you feel about China?" in turns 5, 11, and 15, respectively; but although these topics had a clear logical relationship to one another, the candidates did not have the ability to connect them verbally.

A Synopsis of the DA Results

From the detailed analysis of the 12 in-test discourse excerpts presented in Chapter 4, above, it can be seen that the three groups of candidates adjudged to have distinct L2 Chinese competence levels prior to the paired speaking test could be clearly distinguished from one another in all five of that test's rating categories. The details of these inter-group differences are set forth below.

Table 23

Summary of Distinguishing Features of Language Use, by Competence Group

	Range	Accuracy
High	• Ample knowledge of vocabulary and grammar structures • Always used the discourse level	• Consistently accurate
Mid	• Adequate knowledge of vocabulary and grammar structures • Sometimes used the discourse level, but most of the time only the sentence level	• Various language errors, especially when expressing complex content
Low	• Insufficient reserves of language • Could only use the simple sentence level	• Highly frequent language errors, even when content was straightforward

Language use. As indicated in Table 23, the distinguishing feature of high-competence test-takers in the language-use rating category was that they had mastered ample knowledge of vocabulary and grammar structures, and were thus consistently able to express themselves freely at both the sentence and discourse levels, accurately and without obvious difficulty. The candidates in the middle-competence group had language reserves adequate to their day-to-day exchanges. Occasionally, they were able to express themselves at the discourse level without apparent difficulty, but most of the time they were more comfortable doing so at the sentence level. Small errors could also be identified throughout their in-test discourse. Lastly, the low-competence students had very small language reserves, which usually enabled them to perform only limited interactions, and they thus often encountered communication problems. They lacked the ability to express themselves using complex sentence structures, let alone at the discourse level; and their language errors were extremely frequent, even in simple content.

Situation response. As shown in Table 24, the distinguishing features of situation response are as follows. High-competence examinees had a full understanding of the situational information, were highly sensitive to situations, and consciously produced the appropriate responses required by the task. Middle-competence candidates, for their part, were moderately sensitive towards most of the required situations, but sometimes were unable to perceive or respond appropriately when confronted with a new situation that differed only slightly from a previous one. Lastly, the low-competence students simply did not have adequate sensitivity to situational information; and even when they demonstrated an *understanding* of the required situation, they often forgot to deal with (or perhaps actively avoided dealing with) its more complex aspects.

Table 24

Summary of Distinguishing Features of Situation Response, by Candidates' Competence Levels

Level	Sensitivity	Appropriateness
High	• High sensitivity to all situations • Consciously perform as required	• Appropriate response to specific situations
Middle	• Moderate sensitivity to most required situations • Can complete the basic requirements	• Cannot appropriately respond to some situational changes
Low	• Inadequate sensitivity to many situations • Sometimes forgetting or avoiding responding to situations that are more complex	• Cannot deal appropriately with slightly more complex situations

Turn-taking organization. The distinguishing features of turn-taking organization, as set forth in Table 25, can be described as follows. The high-competence candidates were able to smoothly perform turn-taking, with cooperative overlaps/latches. Their most frequently used mode of speakership shift was other-selection, but active self-selection was also common. The middle-competence test-takers' turn-taking sometimes

Table 25

Summary of Distinguishing Features of Turn-taking Organization, by Candidates' Competence Levels

Level	Pauses, overlaps/latches	Approach to speakership shift
High	• Very smooth • Frequent use of cooperative overlaps/latches	• Commonly use other-selection • Also frequently use active self-selection
Middle	• Some small pauses • Infrequent use of cooperative overlaps/latches	• Mainly use other-selection • Active self-selection is rare
Low	• Not smooth • Prominent, lengthy gaps	• Mainly wait for other-selection • Sometimes the current speaker has to continue to talk

featured small pauses, and their use of cooperative overlaps/latches was comparatively rare. In terms of speakership shift, the members of the middle group mainly waited for other speakers to select them rather than engaging in active self-selection. Lastly, the low-competence participants' turns were not smooth, and were marked by prominent, lengthy gaps. While this group mainly waited for other-selection, there were sometimes serious obstacles to communication by the selected speaker, which forced the current speaker to continue talking to fill the resultant gap.

Sequence organization. The distinguishing features of the participants' sequence organization can be summarized as follows (see also Table 26). The members of the high-competence group exhibited full understanding of each previous turn. They

Table 26

Summary of Distinguishing Features of Sequence Organization, by Candidates' Competence Levels

Level	Understanding of the previous turn	Response tokens	Preference structures
High	• Full understanding	• Frequently use all types, including the most complex • Can use creative assessment	• Use both simple and complex multi-turn pre sequences and post-sequences • Comprehensively use delays, mitigative devices, accounts, etc.
Middle	• Sometimes fail to understand the previous turn, resulting in invalid communication	• Often use tokens show understanding • Can use simple self-designed assessments	• Can use simple pre-sequences • Can use delays, mitigative devices, accounts, etc.
Low	• Often fail to understand the previous turn	• Often use tokens that do not indicate understanding • Can use simple formulaic assessments	• Avoid using dispreferred actions • Do not change the turn-shape

frequently used all kinds of response tokens, and could use more complex tokens such as signals that they wished to speak. They were also able to make creative comments. When conducting dispreferred actions, they could use both simple and complex multiple-turn pre-sequences and post-sequences, and showed a comprehensive mastery of delays, mitigative devices, accounts, and so forth. The middle-competence candidates, meanwhile, sometimes appeared to not understand previous turns, and this resulted in invalid communication. They often used response tokens to show understanding, and could use very simple self-designed assessments as well, that is, they cannot only use the formulaic language. When a dispreferred action was completed, the middle-competence candidates tended to use a simple pre-sequence, though delays, mitigative devices, accounts, and so forth were also sometimes observed. Lastly, in the low-competence group, the students often failed to understand the previous turn. They used response tokens that did not indicate understanding, and their assessments were simple and formulaic. To save each other's face, some avoided using dispreferred actions even if these were part of the task requirements, and/or avoided any method that would have changed the turn-shape.

Topic management. The distinguishing features of the three competence-groups' topic management is summarized in Table 27. The members of the high-competence group could use simple methods of topic development, such as further inquiries, asking back, helping the other person complete his/her turn, and so forth. They also used multi-turn structures, which deepened to discourse level via in-depth development of speculative thinking. They used the stepwise approach to topic shifts, e.g., finding a pivot, building a semantic relationship, and making a summary of the previous meaning; and

used assessment or pre-closing to end some topics. The middle-competence candidates could use simple self-designed patterns to ask back, but were considerably less likely to explore further. Their topic development sometimes extended into the deep-discourse level, but most of it was limited to declarative or introductory content. They seldom used either the step-wise approach to topic shifts, or assessment or pre-closing sequences to end the current topic. Finally, the low-competence test-takers exhibited very weak topic-

Table 27

Summary of Distinguishing Features of Topic Management, by Candidates' Competence Levels

Level	Topic development	Topic shift	Topic termination
High	• Can use simple topic development • Use multi-turn structures that deepen to discourse level with in-depth development of speculative thinking	• Use the step-wise approach to shift topics	• Sometimes use assessment or pre-closing sequences
Middle	• Can use simple self-designed patterns to ask back • Less likely to further explore others' turns • Can use the discourse level, but generally with content that is merely a description or an introduction	• Seldom use the step-wise approach to shift topics	• Seldom use assessment or pre-closing sequences
Low	• Very weak development, limited to the word- and simple-sentence levels • Can use simple formulaic patterns to ask back	• Frequent topic shifts; no ability to build connections between topics	• Abrupt topic terminations

development abilities, which were generally confined to the lexical and syntactic levels; and they could only use simple, formulaic patterns to ask back. The topics they discussed changed frequently, and they had little or no ability to build connections between topics; and thus, their topic terminations were always abrupt.

Summary

This chapter presented the results from the quantitative and qualitative analyses.

From the quantitative analyses, descriptive statistics showed that the individual scores for the five rating categories of the analytical rating rubric for the three paired interactive tasks deviated moderately from the normal distribution. The results of the repeated measures ANOVA indicated that the interaction between tasks and levels was not significant, but that the interaction between rating categories and levels was significant. The follow-up one-way ANOVA showed that the rating category as a "factor" was significantly different across levels. The inter-rater reliability estimates for 4 of the 15 task/category pairs were below the minimum acceptable threshold. They were the situational response in tasks 1 and 3, and the turn-taking organization and sequence organization in task 2. The three categories with interactional features (turn-taking organization, sequence organization, and topic management) all cleared the acceptable thresholds. The inter-rater reliability of the entire test was only slightly below the sufficient reliability threshold. The internal consistency of the analytical rating rubric was relatively high which indicated the five rating categories measured the same construct, and that this was especially true for the three categories associated with the interactional features. The category of situational response in task 2 had the lowest value. Pearson correlation analyses indicated that the correlation between paired interactive tasks and

solo tasks and the correlation between the paired interactive tasks and the levels of Chinese proficiency assessed by the participants' instructors were both relatively high.

From the qualitative analyses, the analysis of raters' online interview showed that both of the two raters believed that rating training was helpful and effective, taking raters' notes was essential during the scoring process, and that the main difficulty encountered was distinguishing candidates in the same group with similar voices. Overall, the analysis of the raters' notes indicated that Rater 1 focused more on rating the interactional features than Rater 2 did, while Rater 2 focused more on content, authenticity, and personality. In the first two role-play tasks, Rater 1 detailed her score on the situational response. In rating role-play 1 and situational discussion tasks, Rater 2 focused on rating the topic development. Through a detailed discourse analysis of the 12 excerpts from in-test discourse, it could be seen that the five rating categories of the analytical rating rubric showed distinguishing features across different competence levels. These distinguishing features were summarized in tables.

These results will be combined and used to analyze the four research questions in the next chapter.

CHAPTER 5

DISCUSSION

The present research has established that the proposed paired speaking test (comprising two open role-play tasks and a situational topic discussion) and its analytical rating criteria are suitable for application to both diagnostic and achievement testing in the classroom. Scores from this paired speaking test and the analytical rating criteria, and a cross-section of in-test discourse, were also used to investigate the distinguishing interactional features across competence levels in the personal language-use domain, yielding fresh understanding of trends in the development of Chinese-language learners' L2 pragmatic competence in interaction.

To serve the above purposes, however, the interpretation of test scores needs to be meaningful for all stakeholders. In other words, when a test-taker receives a certain grade, that grade should clarify – not only the candidate him- or herself, but to other relevant parties – what types of tasks relevant to L2 pragmatics in interaction he/she can successfully perform. As such, the tasks included in the test, and its rating criteria's diagnostic information, are particularly important to score interpretation.

Target Domain and Task Design

Research Question 1: *How effectively do the three paired speaking tasks developed in this study reflect Chinese learners' L2 pragmatic competence in interaction? To what extent do these tasks strike a balance between standardization and authenticity?*

The first research question addresses the domain of target-language use, and the design of tasks appropriate to eliciting test-takers' performance in that domain. The current research proceeds from an assumption that, if such assessment tasks' scores are to

be meaningful, the performance they observe should reflect the language used in real life. In other words, the combination of task type and task content should be capable of eliciting performance that is representative of the candidate's L2 pragmatic competence in interaction that is required by the target language-use domain.

Target domain. Because the boundary of the personal language-use domain is unclear, this study clarified it via needs analysis, which included the perspectives of two groups – international students, and Chinese-language teachers at universities – and covered common personal relationships, frequently used language functions, places where interactions occurred, common situations, topics discussed with friends and strangers, and speech acts. The results established that, among both target groups, the personal language-use domain was held to refer to language used by friends, family members or strangers (whether in public or private) to exchange opinions, engage in casual discussions, address life problems, and perform specific speech acts. As such, it was distinct from the language used in the workplace, academia, and commercial transactions.

The common social situations most frequently mentioned in connection with the personal domain by the needs-analysis participants included casual chat, informal discussion, and inviting friends to participate in activities. The common discussion topics mentioned in this context included all aspects of Chinese life, personal information, personal perceptions, and values; and the most frequently mentioned speech acts from the personal domain included invitations, requests, questions/answers, and apologies.

Task type. This dissertation's literature review indicated that, as compared to OPI, paired oral tests can elicit more diverse interactional features (e.g., Ducasse &

Brown, 2009; Wang, 2015), a more symmetrical interactional pattern (e.g., Galaczi, 2008; Iwashita, 1998; Kormos, 1999; Lazaraton, 2002; Taylor, 2001), and a wider range of language functions and roles (e.g., Galaczi et al., 2011; Skehan, 2001). Moreover, participants in paired tests are given more opportunities to demonstrate their conversational skills (e.g., Brooks, 2009; O'Sullivan, Weir, & Saville, 2002), and such tests can thus provide better oral-language sampling than other types of speaking tests (Skehan, 2001). Within the category of paired speaking tests, meanwhile, extended discourse (Taguchi & Roever, 2017) – which includes both open role-play tasks (Youn, 2013, 2015) and topical discussions (Galaczi, 2004, 2014) – has been found well suited to testing interaction-relevant constructs.

Many previous studies have utilized open role-play, topical discussion, and decision-making tasks to test candidates' L2 pragmatic competence in interaction in the personal language-use domain. However, the results of the pilot study of the current study indicated that, while open role-play tasks were effective in this regard, topical discussion and decision-making tasks did not elicit language suitable for assessment of this domain, mainly because turn lengths were too long and turn-taking too slow, as compared to natural daily conversations. Therefore, to foster more natural, easy, and casual conversation, the present researchers added a familiar situation to the topical discussion, in which the candidates were also instructed to continue their roles from role-play task 1. Specifically, having become friends at a "Chinese corner" event on campus, the two now often go out together, and today are waiting at a bus stop on the way to go shopping: a natural, immersive context for discussion of the topic and for the generation of language appropriate to the personal language use domain. In other words, the

combination of topical-discussion and role-play task elements was found to be highly effective.

Task design. In any test, it is important to ensure standardization; and to test L2 pragmatic competence in interaction, the authenticity of tasks should be maximized. In this study, to achieve standardization, the role-play tasks were given fixed requirements regarding their situations, personal relationships, topics to be discussed, and specific speech acts. In open role-play task 2, for example, the two characters are good friends; the situation and topic are an invitation to attend a Christmas party after Chinese class and a request to borrow a coffee machine. The situational topic discussion also featured fixed requirements regarding the situation, personal relationship, and general discussion content. For example, advanced test takers were instructed to portray good friends going shopping together and talking about the city's livability while waiting at the bus stop; and their discussion prompts included environmental pollution, social security, life convenience, economic development, and residents' friendliness.

All tasks aimed to elicit verbal interactions that were authentic, that is, as similar as possible to naturally occurring conversations. Therefore, for purposes of standardization, all the tasks incorporated a particular degree of openness, from low to high: with role-play task 2 featuring low openness; role-play task 1, medium openness; and situational topic discussion, high openness. Role-play 2 had the lowest degree of openness, due to information not being shared with the participants. For example, A invites B to attend a Christmas party, but B can choose to participate or not. Role-play 1 is more open than role-play 2, due to its standardized situation, that is, the participants knowing each other and engaging in a free-form talk about their impressions of China.

Lastly, the situational topic discussion – as an extended discourse – had the highest degree of openness, with nothing other than prompts to control the candidates' discussion.

Thanks to this standardization of the test design, all candidates' interactions could be meaningfully compared. Among the 12 excerpts discussed above, the performance of the three proficiency groups varied greatly in terms of turn-taking organization, sequence organization, and topic management; and DA results confirmed that their in-test discourse displayed the characteristics of natural discourse structure. It should be borne in mind, however, that the foundation for the tasks' authenticity and standardization was laid by the clear description of the target domain arrived at via needs analysis, which can thus be deemed an effective method for ensuring that task designs will be effective at measuring constructs in a given target language-use domain.

Rating Reliability

Research Question 2: *When using an analytical rubric with interactional features, to what extent can raters ensure the reliability and consistency of their rating?*

This study's second research question concerns how to design rating rubrics and conduct rater training in such a way that the scores assigned to the performance elicited by the speaking test accurately reflect each candidate's L2 pragmatic competence in interaction. The design of rating rubrics is foundational to rating, and it is essential that such rubrics incorporate appropriate, measurable interactional features. At the same time, ensuring that the raters receive a sufficient quantity of high-quality training is also vital to the rating process as a whole.

Rating rubric design. Well-designed rating rubrics can help ensure raters' scores are consistent, and that the relevant rating categories are adequately reflected. The

analytical rating rubric developed for purposes of this study was based on a thorough review of the theoretical and assessment literature. Since, in the context of Chinese-language education, L2 pragmatic competence in interaction is a brand-new assessment construct, an analytical rubric was selected for use in the current study, as providing the most detailed information about each candidate's performance. As previously noted, based on the theories of interactional competence and conversational organization, along with the findings of previous assessment studies involving interaction (e.g., Youn, 2013), five rating categories were included: (1) language use; (2) situation response; (3) turn-taking organization; (4) sequence organization; and (5) topic management. Among these, three – turn-taking organization, sequence organization, and topic management – are the most relevant rating categories for L2 pragmatic competence in interaction. Of the other two, language use is an indispensable resource for candidates' completion of interactive tasks, while situation response reflects that L2 pragmatic competence in interaction cannot be separated from the social dimension.

The results of internal consistency testing indicated that the developed analytical rating rubric was reliable. More specifically, the entire test was found to have high internal consistency, and the three of its five rating categories that most directly measured L2 pragmatic competence in interaction (see above) had a high internal consistency as well. This indicated that the rubric categories, and especially those three, were measuring the same construct. It is worth noting, however, that the category of situation response had a considerably lower consistency value than the other four categories: an anomaly requiring further exploration.

Raters' reliability. Raters' performance is as important as well-designed rating criteria, and relies upon the provision of an adequate quantity and quality of rater training. Rater training for the developed test battery's two main parts – that is, the solo tasks and the interactive tasks – was conducted in two separate one-hour sessions. The main goals of these sessions were to instill each rater with: (1) a familiarity with and understanding of the rating rubrics; (2) an ability to correctly grade examples of the different competence levels with particular scores, across multiple rounds of rating; and (3) the confidence to actually begin rating students' work independently, after achieving a high level of inter-rater consistency. Qualitative analysis of the raters' online interviews revealed that both of them had a positive attitude towards this training, feeling generally that the purpose of the training was clear, that its content was explained clearly, and that the time allocated to it was sufficient. They praised the quality of the analytical rating rubric, especially in those categories with interactional features; and said they felt that the training program's most important element was the practice application of the rating rubrics to samples of students' speech.

Quantitatively, throughout the test, inter-rater reliability of the two raters was far in excess of the minimum acceptable value (0.6), albeit slightly lower than the sufficient value (0.8). Among the five rating categories, the inter-rater reliability value for language use was the highest, while the three categories with the interactional features all reached the minimum acceptable value. However, inter-rater reliability for the situation-response category did not reach the minimum acceptable value. Overall inter-rater reliability was acceptable, though not sufficient for a main study.

To further investigate why inter-rater reliability was relatively low, the two raters' contemporaneous rating notes were examined closely, mainly for evidence of how well they understood and applied the analytical rubric's rating categories (for the interactive tasks). This qualitative analysis revealed that the two raters followed the rating rubric to differing degrees: with Rater 1 following it much more closely than Rater 2 did. Indeed, it seemed that Rater 2 did not consciously make reference to the rating rubric when rating at all – for instance, making almost no notes regarding the categories of situation response and turn-taking organization, in contrast to Rater 1, who made a large number of notes regarding both of these categories. When rating the category of topic management, the two raters again showed different priorities, with Rater 1 paying more attention to topic initiations and endings, while Rater 2 focused more on topic development. This marked divergence in the two raters' levels of compliance with the rating rubric, in combination with their different foci when rating the same categories, might be sufficient to explain their insufficient inter-rater reliability.

In summary, although the rubric was found to have high reliability, it could be further improved based on the DA results. Although the raters reported in their online interviews that they understood the rating rubric, and that their rater training had been effective, the empirical results render such statements questionable.

The Measured Construct

Research Question 3: *What features useful for distinguishing between varied levels and tasks are identifiable in test-takers' paired test discourse? How much can those distinguishing interactional features deepen our understanding of the developmental trajectory of Chinese learners' L2 pragmatic competence in interaction?*

To achieve this study's main research goal of deepening scholarly understanding of Chinese learners' L2 pragmatic competence in interaction via the interactional features displayed in their paired speaking test discourse, the test in question and its rating rubric were both designed to maximize the visibility of interactional features in the test discourse of L2 learners at all competence levels. A deep understanding of a new construct can be achieved through longitudinal research, but such an approach can be costly in financial as well as time terms. Thus, a cross-sectional approach was selected for the current deep exploration of the construct of L2 pragmatic competence in interaction.

Previous researchers who have assessed interaction-relevant constructs (e.g., Galaczi, 2014) have delineated three major categories of interactional features: topic development (degree of topic development; topic extension of "own" and "other" topics), listener-support moves (backchannelling; confirmation of comprehension), and turn-taking management (in a no-gap-no-overlap manner; following an overlap/latch; following a gap; pause). Though previous studies incorporating this framework have contributed to our understanding of interaction-relevant constructs, the present research suggests that they have not done so comprehensively or systematically enough. Specifically, the detailed results of DA of 12 excerpts of in-test discourse in Chapter 4, above, not only identified additional components – not limited to interactional features – that were required to complete the paired speaking test's interaction-based tasks, but also distinguished new interactional features within the three major interactional rating categories (i.e., turn-taking organization, sequence organization and topic management). Moreover, DA revealed that all the rating categories were distinguishable across three

different competence levels, a finding that was confirmed via quantitative analyses (descriptive statistics and repeated measures ANOVA).

New components. The two new components referred to above were language use and situation response. Each is dealt with in turn below.

Language use. Linguistic competence is a fundamental component of communicative competence and the cornerstone of social interaction (e.g., Celce-Murcia, 2007). Without linguistic competence, in other words, a person cannot complete interactive tasks. In the current study, differences in language use were mainly reflected in two subcategories – range and accuracy – meaning that the participants also had varying reserves of vocabulary and different levels of knowledge of grammatical structure. In particular, the low-competence candidates could only use simple sentences; the intermediate candidates could use more complex ones; and the advanced candidates alternated flexibly between simple and complex sentences as specific situations warranted, also expressing themselves at the discourse level when necessary.

The three groups also performed very differently in terms of accuracy. The low-competence test-takers encountered obstacles caused by their low language reserves even when expressing simple content, which seriously diminished the effectiveness of their interaction. The intermediate group, in contrast, had language reserves that were adequate to the completion of mundane interactions, but their language accuracy was low enough that they experienced various kinds of language problems that impacted the smoothness of their interaction somewhat negatively. The members of advanced group, thanks to their rich language reserves, could accurately comprehend and communicate a broad array of information, and achieved high-quality interaction.

Pearson correlations further supported the idea that the development of L2 pragmatic competence in interaction was closely related to the development of linguistic competence. Specifically, the L2 pragmatic competence in interaction measured in this study was not only strongly congruent with the results of the same candidates' solo oral tests rated by the two raters, but also with their language-competence levels as assessed by their own language teachers prior to the paired speaking test.

Situation response. Because interaction requires a minimum of two participants, it cannot be separated from the social realm. As such, from a sociopragmatics perspective (Leech, 1983; Thomas, 1983), social perception and sensitivity to one's situation are critical components of language learners' L2 pragmatic competence in interaction.

The present study's results indicated that situation response could usefully be divided into two subcategories: sensitivity and appropriateness. The researcher performed multiple comprehension checks during the test to ensure that the candidates understood the task requirements. This process established that candidates at the low-competence level were not sensitive enough to situations, and either forget about or consciously avoided responding to the test's more complicated situations. The middle-competence candidates had some sensitivity to their situations, but could not perceive subtle differences between one situation and another; in role-play task 2, for example, the instructions made it clear that the relationship between the two participants was one of close friendship, and yet some candidates used greetings more suited to people they did not know, or whom they had not seen in a long time. The high-competence candidates, however, were highly sensitive to the situation and could consciously produce appropriate responses to all of the test's situations.

During the rating process, raters' notes recorded missing or erroneous messages, off-topic comments, and bad manners as factors affecting candidates' situation-response scores, since all interfered with the smooth progress of in-test interactions.

Interactional features. In addition to the above two rating categories, the results showed that it was readily possible to make distinctions between and among the three rating categories with interactional features. These are discussed further in the three subsections that follow.

Turn-taking organization. The rating category of turn-taking organization can be divided into three subcategories: pauses, overlaps/latches, and approaches to speakership shift. In terms of both pauses and overlaps/latches, the low-competence learners' turn-takings were usually not smooth, with large, obvious gaps, whereas their intermediate-level counterparts gaps tended to be small; and advanced learners often used cooperative overlaps/latches, rendering their turn-taking both fast and smooth. When it came to speakership shift, the low-competence candidates often just passively waited for the previous speakers to let them speak – and sometimes, even after they were appointed to speak, they were limited by their lack of linguistic competence from doing so. Such cases were generally marked by very long pauses, and in some instances, the previous speaker was left with no choice but to continue speaking. Candidates with moderate competence levels also usually waited for the previous speakers to ask them to take the floor, but sometimes they took the initiative to indicate that they wanted to speak. High-competence candidates, however, were always flexible: either waiting for their partners to assign them to speak, or taking the initiative to speak, depending on the dynamics of the specific situation and their personal inclinations. In other words, it could be seen that

the higher their linguistic competence levels, the stronger L2 Chinese leaners' initiative in interaction was.

Sequence organization. The category of sequence organization can also be broken down into three types of interactional features: understanding the previous turn, response tokens, and preference structures. In first of these subcategories, there was a clear, positive correlation between L2 learners' competence levels and their understanding of the previous speakers' turns. The data also indicated that the response tokens used by L2 learners became more complex in structure, and their content showed deeper understanding, as one moved up the competence levels. Notably, low-competence learners could only use simple formulaic assessments, while middle-level learners could use simple self-designed assessments, and advanced learners were able to use creative assessments for specific situations. Lastly, when using preference structures, it was found that the more complex the structure, the higher the level of linguistic competence it required. Thus, some low-competence learners avoided using dispreferred structures even when they were required to do so by the task instructions – but the same reaction was rare to nonexistent among intermediate and advanced learners. Likewise, at each higher level of competence, the sampled L2 learners were more likely to use various devices (e.g., delay, mitigative devices, accounts) to help each other not to lose face in communication; to be more conscious of their use of such devices; and to use increasingly complex pre- and post-sequences to further reduce the adverse effects of dispreferred structures on interaction.

Topic management. The topic management category can also be divided into three main interactional features: topic development, topic shift, and topic termination. In

the first of these three subcategories, the low-competence learners exhibited no ability to develop complex topics, instead sticking closely to the vocabulary or simple-sentence levels, and using only formulaic patterns when asking similar questions back to the other speaker. Candidates at the intermediate level, in contrast, were able to design their own ways of asking such questions, yet usually did not explore each other's turns in depth, usually staying at the sentence level, and only occasionally operating at the discourse level when describing or introducing things. The high-competence candidates, for their part, were able to use simple topic-development approaches flexibly, and to develop in-depth, discourse-level topics based on specific content.

In terms of topic shift, the low-competence learners were able to speak on a variety of topics, but perhaps only because they did not (or perhaps could not) develop any one topic in depth, and their topic shifts were abrupt and seemingly unmotivated. Learners at the middle and high competence levels, in contrast, tended to develop their topics deeply; and the advanced learners paid particular attention to cohesion, often using the stepwise approach to connect topics.

With regard to topic termination, it should be borne in mind that even native speakers do not always use pre-sequences to indicate that topics are about to end. But that being said, certain topic-termination differences between lower-level and higher-level learners were still discernible in the present study's data. For example, the low-competence group was more likely than the other two groups to terminate topics abruptly, while the middle- and high-competence groups preferred to use assessment or pre-closing sequences to end topics.

Developmental trajectory. Based on the above-mentioned general interactional features, the developmental trajectory of Chinese learners' L2 pragmatic competence in interaction can be summarized by the five elements shown in Figure 3: that is, frequency, proactivity, complexity, content and coherence.

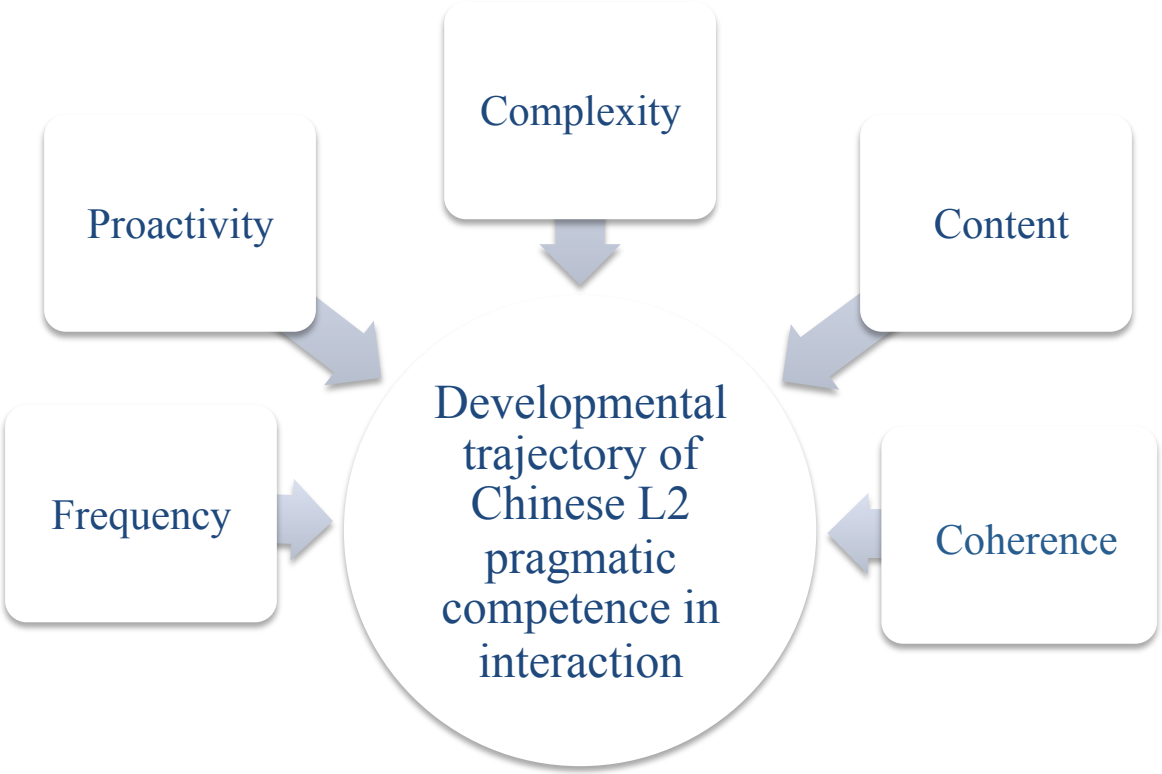


Figure 3. Five elements of the developmental trajectory of Chinese learners' L2 pragmatic competence in interaction

Among these, *frequency* means that the higher a learner's competence level, the more often he/she will interact. This is mainly because the highest-level learners' turn-takings have no gaps, and are generally marked by cooperative overlaps/latches and smoother turn-takings. *Proactivity* means that more-competent learners tend to be more active in their interactions: for example, more likely to want to take the initiative to speak,

rather than passively waiting to be selected by other speakers. *Complexity* refers to the observation that, the higher a learner's competence, the more he/she is able to control complex structures and use creative language. For instance, when evaluating others, the sampled low-competence learners only used simple formulaic assessments, whereas their high-competence counterparts demonstrated an ability to use creative assessments. Similarly, low-competence learners usually avoided using complex sequences related to preference structures, whereas high-competence ones appeared quite comfortable with doing so, using pre-sequences and post sequences to reduce the harm that dispreferred structures would otherwise cause to their interactions; and in the sphere of topic development, high-level learners' topic development moved between the vocabulary, sentence and discourse levels depending on the needs of the specific situation, whereas low-level learners did not move beyond the vocabulary and simple-sentence levels. *Content* refers to the fact that, at higher levels of competence, the subject matter that learners include in their interactions becomes more detailed and profound. For example, high-level learners in this study exhibited higher degrees of understanding of their partners' previous turns. Likewise, learners limited by their low levels of language competence could only develop topics in simple ways, and while intermediate learners encountered fewer such difficulties, only advanced learners could express their own insights in an organized manner and without any obvious obstacles. Lastly, *coherence* refers to the fact that, at higher levels of competence, the sampled L2 Chinese learners tended to pay special attention to the relationships between the words they were using. For example, low-competence learners shifted between topics very quickly, not applying any connectives, whereas almost all high-competence test-takers were able to use the

stepwise approach to change topics in a more natural-seeming way. And low-level learners often ended their topics abruptly, in contrast to their high-level counterparts, who tended to provide assessments or pre-closing sequences to avoid this.

In sum, it would appear that, as L2 pragmatic competence in interaction develops, learners are more active, and their cognitive abilities are more capable of dealing with faster turn-taking, more complex structures, and the more coherent delivery of deeper content. As such, they become more effective speakers and listeners, which in turn may allow them to contribute more to their interactions and have more influence over them. The present study's findings indicate that the construct of Chinese learners' L2 pragmatic competence in interaction is a broad one, incorporating all the important structures of conversational organization; and thus, they provide important empirical support for the body of literature that has argued for a broader definition of interactional competence (Ducasse & Brown, 2009; May, 2009).

The Reliability of Mixed Methods Approach

Research Question 4*:To what extent are the findings from mixed methods reliable and how can they enhance the validity of the future assessment of Chinese learners' L2 pragmatic competence in interaction?*

Qualified mixed methods research is a systematic and legitimated combination of both quantitative and qualitative analysis methods. By combining quantitative and qualitative analysis methods, the strength of each is realized and the weakness of each is minimized. This question aims to reveal how the quantitative and qualitative analysis methods combine together to enhance the reliability of the results of the current study, and to bring inspiration to future assessment research. This study is a sequential mixed

methods design (Teddlie & Tashakkori, 2006), specifically using a quantitative analysis method followed by a qualitative analysis method.

Legitimation techniques. Three legitimation techniques (Brown, 2013; Onwuegbuzie & Johnson, 2006) are used to investigate to what extent this mixed methods design is reliable and has implications for future assessment studies. These legitimation techniques are weakness minimization, convergence, divergence, and clarification.

Weakness minimization. Both quantitative and qualitative studies have their own shortcomings. In a mixed methods design, researchers should carefully evaluate these shortcomings in order to minimize them. The limitation of the DA approach in this study is that the generalization of the findings could be questioned. Quantitative analysis methods, such as descriptive statistics, reliability estimations, etc., cannot show how to operationalize the construct in the assessment and further deepen the understanding of the construct. In this case, the mixed methods design can compensate for the limitations of the two sides, thus improving the reliability of the research results. 90 Chinese L2 learners participated in the study and the personal language use domain was selected as the target domain for the test. Well executed sampling can increase confidence in the generalizability of quantitative research results. The DA analysis indicates how to operationalize the construct in the assessment. In addition, using the DA approach to analyze in-test discourse improved understanding of more distinguishing interactional features. Furthermore, the developmental trajectory of L2 pragmatic competence in interaction can be revealed from a cross-section perspective.

Convergence. Convergence refers to the notion that data from different sources come together to offer similar conclusions. Data from different sources provide different information, thus enhancing the reliability of research results. For example, in this study, the conclusions of statistical analyses and DA analysis both show that the five rating categories in the analytical rating rubric are distinguishable at three different competence levels. The results of statistical analyses can be generalized, but it is not possible to show how the interactional features in the five rating categories are distinguishing. However, a detailed description of distinguishing features can be seen in the results of the DA analysis.

Divergence. Divergence refers to the fact that data from different sources can be combined to get conflicting conclusions, which requires further exploration. Data from different sources that lead to dissimilar results is worthy of attention. Further exploration can lead to interesting findings regarding contextual effects. Quantitative and qualitative analysis methods can offer more perspectives and make research results more robust and reliable. For example, in this study, three rating categories with interactional features showed high reliability through reliability estimates. However, in DA analysis, it showed some distinguishing interactional features different from the analytical rating rubric. The results indicate that the rating rubrics based on a large body of literature might have good reliability, but still cannot fully reflect the whole picture of in-test discourse. The analytical rating rubric can be revised based on the DA research results to further improve its reliability.

Clarification. Clarification refers to the use of additional sources to explain the conclusion drawn from the existing data. Such additional sources can offer additional

explanation and conclusions, while enhancing confidence in the results. In this study, the statistical results showed that when rating the category of "situation response", the inter-rater reliability of the two raters was very low, and the internal consistency of the rating category was also low. In the DA analysis, the rating rubric was valid for measuring this category. Based on the above two data sources, it can be concluded that there was a problem with the rating of this category. The problem was not the rating rubric, but rather the raters. The current data could not provide detailed information on how the raters rated, so a new data source is needed to explain this conclusion. Raters' notes can serve as the new source. Their notes informed us that the two raters had different degrees of compliance with the rating rubric. One of the raters left a large number of records of how she rated this category. However, another rater had basically no records in this category, which can indicate one rater placed great emphasis on rating this category, while another rater ignored this category while rating. These differences between raters led to this inconsistency in evaluating "situation response" and thus a low inter-rater reliability and internal consistency. Therefore, the new data source can help explain the conclusions and make the research results more reliable.

Insights for future assessment studies. The above four types of legitimation techniques indicate that the mixed methods design of this study can enhance the reliability and persuasiveness of the research findings, and bring instruction and enlightenment to future related assessment research, such as how to further improve the reliability of the analytical rating rubric and raters' rating.

Revising the rating rubric. Based on the divergence technique of different data sources, the analytical rating rubric (see Appendix D) based on the literature review is

somewhat different from DA's research findings, and the analytical rating rubric can be modified to improve internal consistency reliability.

The category of turn-taking organization should be modified. In the DA study, the distinguishing characteristics of turn-taking organization are whether there are "cooperative overlaps/latches" and the frequency of using them. They appear frequently in high-level groups, not very frequently in mid-level groups, and not at all in the low-level groups. Another distinguishing feature is the speakership shift method. The high-level group commonly uses both the other-selection and active self-selection approaches. The intermediate group rarely uses the active self-selection approach, and the low-level group does not use the active self-selection approach. Another feature of the low-level group is that other-selection approach may be unsuccessful because the other party's level is low and cannot contribute to the turn. Therefore, the current speaker has no other choice but to continue talking. In the study, we can find that the turn length is not distinguishable. Higher competence level learners don't necessarily use longer turns than lower level counterparts.

In the DA findings, tasks showed different distinguishing characteristics as well: the lowest openness role-play task 2 is suited for measuring the candidate's situation response and preference organization structures. The more open role-play task 1 and the most open situational topic discussion are more suited for measuring candidates' competence to develop topics. Thus role-play and situational topic discussion tasks should use two different analytical rating rubrics. The current five rating categories can be retained in the role-play tasks. However, the situational topic discussion task, observed from DA findings, dose not contain as many specific situations as the role-play tasks do.

Therefore, based on raters' feedback, "content" could be a category more suited than "situation", which may further increase the internal consistency reliability.

Strengthen the rater training. Raters' notes, a new data source mentioned in the clarification technique, show the understanding and application of two raters to the analytical rating rubric. The two raters had different degrees of adherence to the analytical rating rubric, and the rating foci of each rating category were different as well, both of which may result in a low inter-rater reliability. In the future assessment, rater `training can be improved by providing more examples, paying attention to sample selection, and training raters to take standardized notes. Each of these means of improving future assessment will be briefly discussed below.

First, specific examples can be provided to raters. In the rating process, the raters may encounter great difficulties, especially when they rate unfamiliar constructs. It is necessary to help the raters fully understand the new concepts. Providing concrete examples can help raters understand the abstract descriptors in the rating rubrics, and also help the raters understand the key points of grading.

Second, researchers can pay attention to the randomness of the sample. In the rater training of this study, raters had two rounds of time to use the samples to rate, and in the second round the two reached a high level of agreement, and after which the two raters began a separate rate. The six samples used by raters obviously fell into the low, medium, and high levels. However, in the process of real rating, distinguishing levels was much more difficult. Therefore, in the process of practice, particular attention should be given to the randomness of samples, avoiding the selection of samples with obvious discrimination.

Last, raters should be trained to take unified standardized notes. This study found that the raters' notes could be used as an effective tool to view the raters' understanding and scoring rationale. In the process of training, raters should know how to take standardized notes so that the notes are more reliable for further interpretation. During training, raters' understanding of the rating rubric and the rationale for scoring should be checked immediately and periodically afterward. Once misunderstanding or inconsistencies are found, immediate adjustments can be made to avoid larger problems in the real rating process.

Summary

This chapter answered the four research questions using the results of the study. The first research question addressed the domain of target language use and the appropriateness of tasks to eliciting test-takers' performance in that domain. This study defined the personal language use domain based on the needs analysis. According to the pilot study results, the topic discussion and role-play tasks were considered to be effective task types in eliciting natural like conversations in the domain. Tasks were designed to balance authenticity and standardization. The second question was about how to design rating rubrics and conduct rater training. The design of the analytical rating rubric was based on an extensive literature reviews. The internal consistency estimates indicated that the rubric was reliable. But the distinguishing results of DA indicated that each rating category of the rubric could be further revised. Although the raters expressed their positive opinions towards the rater training in the online interviews, their rating notes showed their understanding and application of the rating rubric, and their rating foci were different. The third question was the core issue of the study, which was to deepen

understanding of the new construct (L2 pragmatic competence in interaction) and to understand the developmental trajectory of this construct from a cross-sectional perspective. The results showed that as L2 pragmatic competence in interaction improves, learners are more active, cognitive abilities are better able to handle faster turn-takings and more complex structures, and are more capable of coherent delivery of deeper content. As a result, they become more effective speakers and listeners, which in turn may enable them to contribute more to their interactions and have a greater impact on them. The fourth question mainly answered the contribution of the mixed methods design to the study. The results demonstrated that the mixed methods design could mitigate the weaknesses of both the quantitative and qualitative approaches, offer similar conclusions from different angles, and lead to more interesting findings through different conclusions. New data sources could be used to better interpret the existing conclusions. The mixed methods design enhanced the persuasiveness of the current research findings and provided practical suggestions for the future assessments of Chinese learners' pragmatic competence in interaction. The discussion in this chapter provided an important basis for the conclusion of the next chapter.

CHAPTER 6

CONCLUSION

Many gaps exist in the interaction relevant assessment studies. Based on literature review, paired speaking tests should be used to study L2 pragmatic competence in interaction, especially in languages other than English. And the language test domain should be extended beyond academic purposes and more diverse task types should be designed. In addition, more attention should be paid to distinguishing interaction features across competence levels.

This study seeks to fill these gaps. To do so, it has three main purposes. First, within the Mandarin Chinese context, to examine the developmental trajectory of Chinese learners' L2 pragmatic competence in interaction via speaking test performance discourse, and deepen understanding of L2 pragmatic competence in interaction. Second, to design appropriate open role-play and situational topic discussion tasks in the paired speaking format and ensure that they are sufficient to elicit distinguishing interactional features. Third, to design a reliable analytical rubric for assessing Chinese learners' pragmatic competence in interaction.

The findings realized the study purposes and thus fulfilled the research gaps. Findings indicated that the open role-play and situational topic discussion (extended discourse) tasks were effective in eliciting the interactions needed to assess the construct, since they were designed in a way that balanced standardization and authenticity. The analytical rubric was found to have high internal consistency reliability, and the results of DA indicated that it could be improved further. The inter-rater reliability of the two raters was slightly insufficient. The detailed DA results of the 12 excerpts revealed more

distinguishing interactional features in the three main rating categories (turn-taking organization, sequence organization, and topic management), and other components (language use and contextual response) as well. The five elements of the developmental trajectory of Chinese learners' L2 pragmatic competence were summarized as frequency, initiative, complexity, content, and coherence. Specifically, with the development of L2 pragmatic competence in interaction, learners become more active and capable of handling faster turn-takings, more complex structures, and delivering deeper content in a coherent manner. The results of the mixed method approach were reliable and could help to improve both the analytic rating rubric and quality of future rater training.

Implications

The findings of this study have implications for both language assessment and instruction, including: (1) how to define the personal language use domain; (2) how to develop the appropriate rating rubrics for assessing interaction relevant constructs; (3) how to improve the quality of rater training; and (4) what is the developmental trajectory of the L2 pragmatic competence in interaction?

Personal language use domain. The basis of language assessment is to select the appropriate target language use domain. Most prior research focuses on language for academic purposes. However, the personal language use domain often encountered in daily life is not well defined by CEFR. A needs analysis indicated that the personal language use domain refers to the language used between friends, family members or strangers (either in public or private settings) to exchange opinions, perform casual discussions, solve life problems, and implement specific speech acts. Therefore, it is quite different from the language used in workplaces, academia, and business transactions. The

social situations, discussion topics, and mentioned speech acts were also summarized. This can be used as a reference for those who are interested in choosing this target domain as their test domain, and can lay the foundation for further modification of this target domain in the future.

Rating rubric development. In the speaking assessment, the development of appropriate rating rubrics is critical, and it is also a prerequisite to ensure the raters' rating reliability. A rating rubric can be developed based on the test discourse result. It is called a data-driven rating rubric, the approach of which is very time consuming (Youn, 2015). Fully analyzing the data and then making rating rubrics in every oral test is not realistic, especially for large-scale speaking tests. In such cases, a rating rubric developed from the findings of previous research and theoretical knowledge could be more practical. This study detailed the distinguishing interactional features across competence levels, which could be useful for developing rating rubrics with interactional features in language assessment studies.

Rater training. Rater training is another important process in rating speaking tests. The goal is to ensure that the raters fully understand the rating rubrics and that their ratings are reliable and consistent with other raters. This study found that the raters encountered unexpected difficulties when rating new constructs. In order to improve the quality of future rater training, examples related to the rating rubric descriptors could be provided to the raters. In addition, it was also found that training raters for making standardized rating notes could help observe their inconsistencies in rating. Especially in the process of rater training, if raters' notes are different, adjustments can be made as soon as possible to avoid continued differences in the following real rating.

Developmental trajectory. Based on a detailed analysis of distinguishing interactional features across competence levels, this study summarized the developmental trajectory of L2 pragmatic competence in interaction. Five elements were incorporated: frequency, proactivity, complexity, content and coherence. After leaners' competence is improved, he/she will be more inclined to actively participate in the interaction, and can use more complex structures to express more profound and coherent content. This discovery can help language teachers pay attention to design exercises that enable students to have more interactions. Teachers need to encourage students to actively participate in interactions, consciously use more complex sentences, make efforts to express more profound insights, and pay attention to use connective words.

Limitations

Several limitations merit discussion.

Non-normal distribution. Based on the range values given by two times of standard errors of skewness and kutosis, some of the data moderately deviated from the normal distribution. The normal distribution is the basic assumption of many statistical analyses, such as repeated measures ANOVA. Although research shows that ANOVA is still robust to moderate deviations in normality (Glass, 1972), the results of the statistical analyses should be interpreted cautiously. The mixed method design, which provides consistent evidence from different angles, reduced the magnitude of this limitation.

Participants' identity. This study used the paired test format. Candidates, with similar proficiency level, were paired into the same group by their Chinese language teachers. Many candidates were of the same gender. In some groups, students' pronunciations were similar. The test was audio recorded, and it was found that two test

takers with similar pronunciations could make it difficult for raters to identify their names. However, the raters indicated in their online interviews that such situations were not common and did not have a significant impact on the candidates' scores.

Raters' notes. It is found that the information provided by the rates' notes was helpful in explaining the insufficient inter-rater reliability of the two raters. However, the role of the raters' notes was not thought of before the completion of their rating, so the raters were not trained on how to consistently record rating notes. This may have impacted the reliability of the rating notes. However, this study also provided data about raters' online interviews, which offered more information from multiple perspectives to reduce the impact of data limitations.

Suggestions for Future Research

This study filled the research gaps detailed above. However, L2 pragmatic competence in interaction is a new research field, and more relevant research is urgently needed. The following three aspects are particularly worthy of exploring.

Video recording. According to Ducasse and Brown's (2009) model on interactional features, non-verbal interpersonal communication (gaze and body language) is one of the three categories. This research focused on the interactional features consisting of words and sentences, and thus only audio recording was made. In fact, test takers frequently use accents, intonations, etc. to express particular communicative functions in interaction. For future research assessing L2 pragmatic competence in interaction, researchers may consider making video recordings so that all information related to interactional features can be documented. Interactional features should also be analyzed from the perspectives of non-verbal interpersonal communications and prosody.

Raters' perspective. As mentioned earlier, the raters' notes helped to reveal their understanding and application of the rating rubric. The summary of the distinguishing interactional features and the developmental trajectory of the L2 pragmatic competence in interaction were based on the DA results of participants' test discourse. Raters, who listened to the recordings and scored candidates from beginning to end, would have their own opinions. Raters' notes also revealed that they had their own preferences when rating a category. Understanding the views of the raters will be helpful for a comprehensive understanding of the construct.

Paired group effect. This study used the paired speaking format as the type of test. Different from OPI, this test had no tester to interaction in testing. Two candidates were assigned to one group. The ideal grouping situation was that two test takers have the same or similar levels. All of the groupings were completed by their Chinese language teachers. After testing, it was found that the two test takers of some groups had different language proficiency levels. Raters mentioned in the online interview that candidates were performing well, and there were very few cases where two candidates could not interact equally. However, an interesting research question is to what extent a learner interacts differently when he/she is assigned to another candidate with/without similar language proficiency levels?

Summary

This chapter summarized the study. It summarized the significance of the study and its filling of the research gaps. It indicated the implications of the study in the personal language use domain, rating rubric development, rater training and the development trajectory of the construct. The limitations of the study were also mentioned,

including the data moderately deviating from normality, the raters' issues distinguishing speakers in the audio recording, and the validity of the raters' notes. Finally, it summarized aspects of the research that need to be further explored. These aspects included more comprehensive research about the interactional features through a video recording, understanding the new construct from a different angle by studying raters, and investigating the language proficiency effect between the two candidates within the same group on their performance in-test discourse. It is hoped that the construct of L2 pragmatic competence in interaction can be more thoroughly understood and assessed in future studies.

Appendix A：Background Questionnaire in Chinese

bèi jǐng diào chá wènjuàn
背景调查问卷

yīngwénmíng zi
1. 英文 名字： _____

xìngbié
2. 性别： _____

niánlíng
3. 年龄： _____

mǔ yǔ
4. 母语： _____

guójí
5. 国籍： _____

zài zhōngguó jūzhù niánshù
6. 在 中 国居住年数： _____

zài zhōngguó jūzhù guò de chéngshìhuòzhěchéngzhèn
在 中 国居住过的 城市 或者 城镇 ： _____

shòu jiàoyù jīnglì
7. 受 教育经历：

mùqián de xuéxí qíngkuàng huàquān běnkēshēng shuòshìshēng bóshìshēng qí tā
目前的学习 情况 （画圈）： 本科生 硕士生 博士生 其他： ____

mùqián de zhuānyè
目前的 专业： _____

zuìgāo xuélì
最高学历： _____

huò dé zuìgāo xuélì shí suǒzài guójiā
获得最高学历时所在国家： _____

zhōngwén xuéxí jīnglì
8. 中文 学习经历：

xuéxí zhōngwénniánshù
学习 中 文年数： _____

liè chūsuǒyǒucān jiā guò de zhōngwénxué xí xiàngmù bāokuò zài zhōngguó jí qí tā guó jiā
列出所有参加过的 中 文学习 项 目（包括在 中 国及其他国家）：

_____

_____

mùqián xuéxí zhōngwénsuǒzài dàxué
目前学习 中 文所在大学： _____

mùqiánsuǒxuézhōngwénkèchéng jí bié
目前所学 中文 课程 级别： _____

zhōngwén yǔ yán cè shì jīng lì
9. 中 文语言测试经历：

nǐ yǐ qiánkǎoguò ma
你以前考过HSK吗? _____

rúguǒ kǎoguò kǎoshìniánfènshì
如果考过，考试年份是：_____

zuìgāo bǐshì shuǐpíngshì
最高笔试 水平 是： _____

zuìgāokǒuyǔshuǐpíngshì
最高口语 水平 是： _____

nǐ cānjiā guò qí tā zhōngwénshuǐpíngkǎo shì ma
你参加过其他 中文 水平 考试吗？ _____

rú guǒcān jiā guò qǐng liè chūsuǒyǒu de kǎo shì míngchēng
如果参加过，请列出所有的考试 名 称 ： _____

měi yígè kǎoshì de niánfèn
每一个考试的年份 : _____

měi yígè kǎoshì de zuìgāoshuǐpíng
每一个考试的最高 水平 : _____

lián xì fāngshì kě xuǎnzé de
10. 联系方式(可选择的)：

yóuxiāng
邮箱 ： _____

wēixìn
微信： _____

Appendix A (continued): Background Questionnaire (in English)

1. English name: _____
2. Gender: _____
3. Age: _____
4. Native language: _____
5. Nationality: _____
6. Time spent living in China: _____

   Cities or towns of China where you have lived: _____
7. Educational experience:

   Current academic status (circle): Undergraduate MA Ph.D. other: _____

   Current main subject of study: _____

   Highest degree you have earned: _____

   Country in which you earned your highest degree: _____
8. Chinese language-learning experiences:

   Time spent learning Chinese: _____

   List all the Chinese language programs you've attended (both in China and other countries): _____

   _____

   Currently taking Chinese at: _____ (University)

   Level of your current Chinese-language class: _____
9. Chinese-language testing experiences:

   Have you taken HSK before? _____

   If yes, test year(s): _____

   the highest level of your written ability: _____

   the highest level of your speaking ability: _____

   Have you taken other Chinese-language proficiency tests: _____

   If yes, list all: _____

   test years of each: _____

   the highest level of each: _____
10. Contact information (optional):

    Email: _____

    WeChat: _____

Appendix B：Survey of International Students' Needs for Chinese Use in the Personal Domain, from Their Teachers' Perspectives (in Chinese)

从教师的角度了解留学生在"个人领域"中使用中文的情况

1.引言

我是夏雪，现为夏威夷大学东亚系的一名在读博士生。为了获得我的博士学位，我需要完成一个研究项目。我的研究目的是要测试不同水平汉语学习者的互动能力。为了更好地设计测试，我需要从您的角度了解留学生在个人领域中使用中文的情况。所以您的参与对我来说非常重要。

2.意向书

步骤: 你需要简单地填写个人背景信息，然后需要回答几个开放式的问题。这些开放式的问题主要是关于在个人领域中会常用到的情景及话题等。

时长: 完成整个问卷大概需要15到20分钟。

利益及风险: 参加这个问卷调查，对您来说可能没有直接的利益。但是，调查的结果会帮助研究人员更好地了解个人领域的对话情况。对您来说，参加这个问卷调查几乎没有任何风险。

隐私及保密性: 在任何时候，您的回答都不会与个人识别信息相联系。您的回答会和其他人的回答放在一起做分析和报告。整个程序都是匿名的。只有研究小组可以看到相关资料。

自愿参加: 是否参加本研究项目完全是自愿的。您可以在任何时候退出该项目。如果您停止参与该项目，也不会有惩罚或者损失。

报酬: 没有报酬。

问题: 如果您对本研究有任何问题，请用邮箱 xuexia@hawaii.edu 给我发邮件。您也可以跟我的导师，王海丹博士联系，她的邮箱是 haidan@hawaii.edu 。如果您对作为一个研究参与人所拥有的权利有问题，请联系夏威夷大学人类研究审查组，邮箱为：uhirb@hawaii.edu。

*** 我已经阅读并理解上述信息。我同意参与这个问卷调查，并且允许研究人员使用跟上述内容有关的信息。**

同意 _____
不同意_____

3. 关于“个人领域”的介绍

根据《欧洲现代语言教学大纲》，外语使用领域主要分为：个人、公共、教育、职业等四个领域。其中“教育”、“职业”领域比较容易区分。“个人”、“公共”领域需要再进一步地阐明 。“个人”领域是一个很宽泛的领域，泛指与家人、朋友，甚至是陌生人等的交流、出行等。“公共”领域主要是指在公共场所与特定的服务人员进行交易、处理事务等，比如：在餐馆点餐。本问卷调查主要是关于留学生在“个人领域”的中文使用需求的调查。

4.背景信息

性别：_____

年龄：_____

教中文的年数：_____

教学城市：_____

教过的课程级别：_____

5.开放式问题

（1）如前言所述，“个人领域”是一个很宽泛的概念。您觉得应该包括哪些次级领域？（例如：娱乐休闲、家庭聚会等）

（2）您觉得留学生在“个人领域”中使用中文最常遇到的情景有哪些？（例如：下课后，留学生A邀请留学生B一起吃晚饭）

（3）您觉得留学生在“个人领域”中使用中文遇到的最大障碍是什么？（例如：不知道如何跟陌生人挑起话题）

（4）您觉得留学生在“个人领域”中使用中文最常使用的“言语行为”有哪些？（主要包括：请求、道歉、拒绝、邀请、同意/不同意、抱怨、询问、回答、给与、建议、争论、开玩笑、承诺，等等）

（5）您觉得留学生在和刚认识的人用中文聊天时最常聊的话题有哪些？（例如：个人基本信息）

（6）您觉得留学生在和朋友用中文聊天时最常聊的话题有哪些？（例如：旅行计划）

Appendix B (continued): Survey of International Students' Needs for Chinese Use in the Personal Domain, from Their Teachers' Perspectives (in English)

1. Foreword

My name is Xue Xia. I am a Ph.D. student at the University of Hawaii at Manoa in the Department of East Asian Languages and Literatures. As part of the requirements for earning my Ph.D. degree, I am doing a research project. The purpose of my research is to design Chinese oral testing to assess Chinese-language learners' interactional competence across different proficiency levels. To optimize the design of the testing tasks, I need to know about international students' Chinese use in the personal domain from your perspective. Thus, your participation is valuable to me.

2. Consent Form

Procedures: You will be asked questions regarding your background; and frequently used situations, topics, etc. in the personal domain.

Duration: It will take approximately 15-20 minutes to complete the survey.

Benefits and risks: Participating in this survey may not result in any direct benefit to you. However, its findings can help researchers to understand conversations in the personal domain better. I believe there is little risk to you in participating in this survey.

Privacy and Confidentiality: Your responses will not be associated with individually identifiable information at any point. Your answers will be combined with the responses of others for purposes of analysis, and your name will be kept anonymous. Only the research team will have access to the survey data.

Voluntary Participation: Your participation in this project is entirely voluntary. You may stop participating at any time. If you withdraw from the project, there will be no penalty or loss to you.

Compensation: There is no compensation for completing this survey.

Questions: If you have any questions about this survey, please email me at xuexia@hawaii.edu. You may also contact my adviser, Dr. Haidan Wang, at haidan@hawaii.edu. If you have questions about your rights as a research participant, you may communicate with the UH Human Studies Program at uhirb@hawaii.edu.

***I have read and understood the above information. I agree to participate in this survey and permit the researcher to use the data as described above.**

Yes_____
No _____

3. Introduction of "personal domain"

According to the *Common European Framework of Reference for Languages: Learning, teaching, assessment* (CEFR), language learners' foreign-language use can be divided into four domains: personal, public, educational, and occupational. The educational and occupational domains are relatively easy to distinguish, whereas the personal and public

domains need to be further clarified. The personal domain is broad, and generally refers to communications between family members, friends, and even strangers. The public domain refers to transactions with service workers in public places: for instance, ordering food in a restaurant. This survey is mainly about international students' needs to use Chinese in the personal domain.

4. Background information
Gender: _____
Age: _____
Year(s) of teaching Chinese: _____
Cities where you have taught Chinese: _____
Levels of Chinese courses you have taught: _____

5. Open-ended questions
(1) As described in the foreword, the "personal domain" is broad. Which sub-domains do you think should be included in it? (leisure activity, family gatherings, etc.)

(2) What situations do you think international students most frequently encounter in the personal domain? (e.g., after class, international student A invites international student B for dinner)

(3) What are the biggest obstacles do you think international students face in the personal domain? (e.g., not knowing how to initiate conversational topics with strangers)

(4) What actions do you think international students most frequently use in the personal domain? (e.g., request, apology, refusal, invitation, agreement/disagreement, complaint, inquiry, response, offering, suggestion, argument, joking, commitment)

(5) What topics do you think international students use most frequently when they are talking with strangers? (e.g., personal information)

(6) What topics do you think international students use most frequently when they are talking with friends? (e.g., travel plans)

Appendix B (Continued): Survey about Your Needs of Chinese Use in the Personal Domain in Chinese

yǒuguānnín zài gè rénlǐng yù zhōng shǐyòngzhōngwén de qíngkuàngdiàochá
有关您在“个人领域”中使用中文的情况调查

yǐnyán
1.引言：

wǒ shì xià xuě xiànwéixiàwēi yí dàxué dōngyà xì de yī míng zàidú bóshìshēng wèile huò dé wǒ de bóshì
我是夏雪，现为夏威夷大学东亚系的一名在读博士生。为了获得我的博士
xuéwèi wǒ xūyào wánchéng y í g è yánjiūxiàngmù wǒ d e yánjiūm ù d ìshìyàoc èsh ìbùtóngshuǐpínghànyǔxu éxí
学位，我需要完成一个研究项目。我的研究目的是要测试不同水平汉语学习
zhě de hùdòngnénglì wèile gènghǎo di shèjì cèshì wǒ xūyàocóngnín de jiǎodù liǎojiě liúxuéshēngzài gèrén
者的互动能力。为了更好地设计测试，我需要从您的角度了解留学生在个人
lǐngyùzhōngshǐyòngzhōngwén de qíngkuàng suǒyǐnínd ecānyùduìwǒláishuōfēichángzhòngyào
领域中使用中文的情况。所以您的参与对我来说非常重要。

yì xiàngshū
2.意向书

bùzhòu nǐ xūyàojiǎndān di tiánxiě gèrén bèijǐngxìn xī ránhòuxūyào huídá jǐ gè kāifàngshì de wèntí zhèxiē
步骤: 你需要简单地填写个人背景信息，然后需要回答几个开放式的问题。这些
kāifàngshì de wèn tí zhǔyàoshìguān yú zài gè rénlǐng yù zhōnghuìchángyòngdào de qíngjǐng jí huà tí děng
开放式的问题主要是关于在个人领域中会常用到的情景及话题等。

shí zhǎng wánchéngzhěng gè wènjuàn dà gài xū yào dào fēnzhōng
时长: 完成整个问卷大概需要15到20分钟。

lì yì jí fēngxiǎn cānjiāzhègewènjuàndiàochá duìnínláishuōkěnéngméiyǒuzhíjiēd e lì yì dànshì
利益及风险: 参加这个问卷调查，对您来说可能没有直接的利益。但是，
diàochá d ejiéguǒhuìbāngzhùyánjiūrényuángènghǎo di liǎojiěgèrénlǐng yù de duìhuàqíngkuàng duìnínláishuō
调查的结果会帮助研究人员更好地了解个人领域的对话情况。对您来说，
cānjiāzhègewènjuàndiàochá jī hū méiyǒurènhéfēngxiǎn
参加这个问卷调查几乎没有任何风险。

yǐn sī jí bǎo mì xìng zài rènhé shíhòu nín de huídá dōu búhuì yǔ gèrén shíbié xìnxī xiāng liánxì nín de
隐私及保密性: 在任何时候，您的回答都不会与个人识别信息相联系。您的
huídá huì hé qí tārén de huídáfàngzài yì qǐ zuòf ēnx īh ébàogào zhěnggèchéngxùdōushìnìmíng de zhǐyǒu
回答会和其他人的回答放在一起做分析和报告。整个程序都是匿名的。只有
yánjiūxiǎozǔ kě yǐ kàndàoxiāngguānzīliào
研究小组可以看到相关资料。

zìyuàn cānjiā shìfǒu cānjiā běnyánjiūxiàngmùwánquánshì zìyuàn de nín kěyǐ zài rènhé shíhòutuìchū gāixiàngmù
自愿参加: 是否参加本研究项目完全是自愿的。您可以在任何时候退出该项目。
rúguǒ níntíngzhǐ cānyù gāixiàngmù yě búhuì yǒuchéng fá huòzhěsǔnshī
如果您停止参与该项目，也不会有惩罚或者损失。

bàochou méiyǒubàochou
报酬: 没有报酬。

wèntí rúguǒ nínduìběnyánjiūyǒu rènhé wèntí qǐngyòngyóuxiāng gěiwǒ fā yóujiàn
问题: 如果您对本研究有任何问题，请用邮箱 xuexia@hawaii.edu给我发邮件。
nín yě kě yǐ gēnwǒ de dǎoshī wánghǎidān bó shìlián xì tā de yóuxiāngshì rú
您也可以跟我的导师，王海丹博士联系，她的邮箱是haidan@hawaii.edu。如

guǒnínduìzuòwéi yígè yánjiū cānyù rénsuǒyōng yǒude quán lì yǒu wèntí qǐnglián xì xiàwēi yí dàxué rénlèi yánjiū
果您对作为一个研究参与人所拥有的权利有问题，请联系夏威夷大学人类研究
shěnchá zǔ yóuxiāngwéi
审查组，邮箱为：uhirb@hawaii.edu。

wǒyǐjīngyuèdúbìnglǐjiěshàngshùxìnxī wǒtóngyìcānyùzhègewènjuàndiàochá bìngqiěyǔnxǔyánjiū
***我已经阅读并理解上述信息。我同意参与这个问卷调查，并且允许研究**
rényuánshǐyòngguānshàngshùnèiróngyǒuguāndexìnxī
人员使用跟上述内容有关的信息。

tóng yì
同意 _____
bù tóng yì
不同意_____

guānyú gèrénlǐngyù dejièshào
3.关于"个人领域"的介绍

gēnjù ōuzhōuxiàndàiyǔyánjiàoxuédàgāng wàiyǔshǐyònglǐngyùzhǔyàofēnwéi gèrén gōnggòng
根据《欧洲现代语言教学大纲》，外语使用领域主要分为：个人、公共、
jiàoyù zhíyèděngsìgèlǐngyù qízhōng jiàoyù zhíyè lǐngyùbǐjiàoróngyìqūfēn gèrén
教育、职业等四个领域。其中"教育"、"职业"领域比较容易区分。"个人"、
gōnggòng lǐngyùxūyàozàijìn yí bù dì chǎnmíng gèrén lǐngyùshìyígèhěnkuānfàndelǐngyù fànzhǐyǔ
"公共"领域需要再进一步地阐明。"个人"领域是一个很宽泛的领域，泛指与
jiārén péngyǒu shènzhìshìmòshēngréndeng de jiāoliú chūxíngděng gōnggòng lǐngyùzhǔyàoshìzhǐzài
家人、朋友，甚至是陌生人等的交流、出行等。"公共"领域主要是指在
gōnggòngchǎngsuǒ yǔ tèdìng de fú wùrényuánjìnxíngjiāoyì chǔ lǐ shìwùděng bǐ rú zàicānguǎndiǎncān
公共场所与特定的服务人员进行交易、处理事务等，比如：在餐馆点餐。
běnwènjuàndiàocházhǔyàoshìguānyúliúxuéshēngzài gèrénlǐngyù de zhōngwénshǐyòngxūqiú de diàochá
本问卷调查主要是关于留学生在"个人领域"的中文使用需求的调查。

bèijǐngdiàochá
4.背景调查

xìngbié
性别：_____
niánlíng
年龄：_____
mǔ yǔ
母语：_____
guó jí
国籍：_____
xuézhōngwén de niánshù
学中文的年数：_____
xuéxíguò de chéngshì
学习过的城市：_____
xuéguò de zhōngwénkèchéng jí bié
学过的中文课程级别：_____
rúguǒcānjiāguò kǎoshì shuǐpíngwéi
如果参加过HSK考试，水平为：_____

kāifàng shì wèn tí
5.开放式问题

rú qiányán suǒshù gèrén lǐngyù shì yígè hěnkuānfàn de gàiniàn nín juéde yīnggāibāokuò nǎ jǐ lèi
（1）如前言所述，“个人领域”是一个很宽泛的概念。您觉得应该包括哪几类？

lì rú yú lè xiūxián jiā tíng jù huìděng
（例如：娱乐休闲、家庭聚会等）

nín juéde zài gèrén lǐngyù zhōngshǐyòngzhōngwén zuìcháng yù dào de qíngjǐngyǒu nǎxiē lìrú
（2）您觉得在“个人领域”中使用中文最常遇到的情景有哪些？（例如：

xiàkè hòu yāoqǐngtóngxué yì qǐ chīwǎnfàn
下课后，邀请同学一起吃晚饭）

nín juéde zài gèrén lǐngyù zhōngshǐyòngzhōngwén yùdào zhàngài de qíngjǐngyǒu nǎxiē lìrú
（3）您觉得在“个人领域”中使用中文遇到障碍的情景有哪些？（例如：

xiǎng hé mòshēngrén jiāoliú què bù zhīdào r ú h é kāishǐshuōhuà
想和陌生人交流，却不知道如何开始说话）

nín juéde zài gèrénlǐngyù zhōngshǐyòngzhōngwénzuìchángshǐyòngd e yányǔxíngwéi yǒunǎxiē
（4）您觉得在“个人领域”中使用中文最常使用的“言语行为”有哪些？

zhǔyàobāokuò qǐngqiú dàoqiàn jùjué yāoqǐng tóngyì bùtóngy ì bàoyuàn xúnwèn huídá
（主要包括：请求、道歉、拒绝、邀请、同意/不同意、抱怨、询问、回答、

gěiyǔ zhēnglùn kāiwánxiào chéngnuò děngděng
给与、争论、开玩笑、承诺，等等）

nín juéde zài hé gāngrènshi de rényòngzhōngwénliáotiānshí zuìchángliáo dehuà tí yǒu nǎxiē lìrú
（5）您觉得在和刚认识的人用中文聊天时最常聊的话题有哪些？（例如：

gèrén jī běnxìn x ī
个人基本信息）

nín juéde zài hé péngyǒuyòngzhōngwénliáotiānshí zuìchángliáo dehuà tí yǒu nǎxiē lìrú lǚxíng
（6）您觉得在和朋友用中文聊天时最常聊的话题有哪些？（例如：旅行

jìhuà
计划）

Appendix B (Continued)：Survey about Your Needs of Chinese Use in the Personal Domain in English

1. Foreword

My name is Xue Xia. I am a Ph.D. student at the University of Hawaii at Manoa in the Department of East Asian Languages and Literatures. As part of the requirements for earning my Ph.D. degree, I am doing a research project. The purpose of my research is to design Chinese oral testing to assess Chinese-language learners' interactional competence across different proficiency levels. To optimize the design of the testing tasks, I need to know about international students' Chinese use in the personal domain from your perspective. Thus, your participation is valuable to me.

2. Consent Form

Procedures: You will be asked questions regarding your background; and frequently used situations, topics, etc. in the personal domain.

Duration: It will take approximately 15-20 minutes to complete the survey.

Benefits and risks: Participating in this survey may not result in any direct benefit to you. However, its findings can help researchers to understand conversations in the personal domain better. I believe there is little risk to you in participating in this survey.

Privacy and Confidentiality: Your responses will not be associated with individually identifiable information at any point. Your answers will be combined with the responses of others for purposes of analysis, and your name will be kept anonymous. Only the research team will have access to the survey data.

Voluntary Participation: Your participation in this project is entirely voluntary. You may stop participating at any time. If you withdraw from the project, there will be no penalty or loss to you.

Compensation: There is no compensation for completing this survey.

Questions: If you have any questions about this survey, please email me at xuexia@hawaii.edu. You may also contact my adviser, Dr. Haidan Wang, at haidan@hawaii.edu. If you have questions about your rights as a research participant, you may communicate with the UH Human Studies Program at uhirb@hawaii.edu.

***I have read and understood the above information. I agree to participate in this survey and permit the researcher to use the data as described above.**

Yes_____
No _____

3. Introduction of "personal domain"

According to the *Common European Framework of Reference for Languages: Learning, teaching, assessment* (CEFR), language learners' foreign-language use can be divided into four domains: personal, public, educational, and occupational. The educational and

occupational domains are relatively easy to distinguish, whereas the personal and public domains need to be further clarified. The personal domain is broad, and generally refers to communications between family members, friends, and even strangers. The public domain refers to transactions with service workers in public places: for instance, ordering food in a restaurant. This survey is mainly about international students' needs to use Chinese in the personal domain.

4. Background information

Gender: _____
Age: _____
Native language(s): _____
Nationality: _____
Years of learning Chinese: _____
Cities where you have studied Chinese：_____
Levels of Chinese courses you have taken：_____
If you took HKS before, the highest level：_____

5. Open-ended questions

(1) As described in the foreword, the "personal domain" is broad. Which sub-domains do you think should be included in it? (leisure activity, family gatherings, etc.)

(2) What situations do you think international students most frequently encounter in the personal domain? (e.g., after class, international student A invites international student B for dinner)

(3) What are the biggest obstacles do you think international students face in the personal domain? (e.g., not knowing how to initiate conversational topics with strangers)

(4) What actions do you think international students most frequently use in the personal domain? (e.g., request, apology, refusal, invitation, agreement/disagreement, complaint, inquiry, response, offering, suggestion, argument, joking, commitment)

(5) What topics do you think international students use most frequently when they are talking with strangers? (e.g., personal information)

(6) What topics do you think international students use most frequently when they are talking with friends? (e.g., travel plans)

Appendix C: The Speaking Test in Chinese

kǒu yǔ kǎo shì
口语考试

yī yí gè rén de kǎoshì
一、一个人的考试

qǐng yòng zhōngwén huídá wèntí nǐ yào shuō dào wèntí xiàmiàn de měi yìdiǎn nǐ yǒu
请 用 中文 回答 问题，你 要 说 到 问题 下 面 的 每 一点 。你 有
miǎo de shíjiān zhǔnbèi rán hòu yǒu fēn zhōng de shíjiān huídá jiù xiàng píngcháng
30 秒 的 时间 准备 ， 然 后 有 1 分 钟 的 时间 回答 。就 像 平常
shuōhuà nàyàng shuō qǐngjǐn kě néng di duō shuō shuō rúguǒ nǐ yǒu bùmíngbai de dì fāng
说话 那样 说 ， 请 尽 可 能 地 多 说 说 。 如果 你 有 不明 白 的 地 方 ，
qǐngwèn wèn tí
请 问 问 题。

qǐng shuō chū yígè nǐ lǚyóuguò de dìfāng bǐrú chéngshìhuò yígè zhùmíng de
1. 请 说 出 一个 你 旅游 过 的 地方 （比如： 城市 或 一个 著名 的
jǐngdiǎn tā gěi nǐ liúxià le hěnshēn de yìnxiàng
景点 ），它 给 你 留下 了 很 深 的 印象 。

nǐ yīnggāi shuō yī shuō
你 应 该 说 一 说 ：

- tā zài nǎli bǐrú guójiāhuò dìqū
  它 在 哪里（比如： 国家 或 地区）
- yǒu shénme tèbié de fāngmiàn bǐrú
  有 什么 特别的 方面 （比如：
  míngshènggǔjì tèsè xiǎochī jiànzhù
  名胜古迹 ， 特色 小吃 ， 建 筑 ，
  děngděng
  等 等 ）
- wèishénme gěi nǐ liúxià le hěnshēn de yìnxiàng
  为 什 么 给 你 留 下 了 很 深 的 印 象

qǐng nǐ bǐ yíxià bùtóng dìfāng de yǐnshí xíguàn bǐrú zhōngguó de nánfāng hé běifāng
2. 请 你比一下 不同 地方 的 饮食 习惯 （比如： 中国 的 南方 和 北方 ，
huò nǐ de guójiā hé zhōngguó
或 你的 国家 和 中国 ）。

nǐ yīnggāi shuō yī shuō
你 应 该 说 一 说 ：

- tā menfēnbié shì nǎliǎng gè dì fāng
  它 们 分 别 是哪 两 个地 方
- tā menyǒu shénme xiāngtóng hébùtóng de fāngmiàn
  它 们 有 什 么 相 同 和不 同 的 方 面

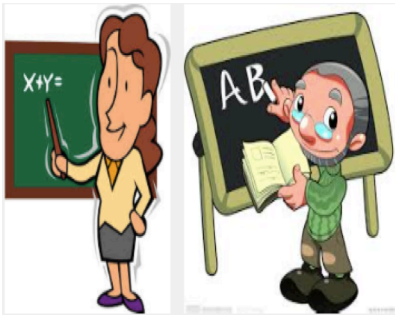
bǐ rú shí wù de tè sè cài de kǒuwèi yòng
（比如： 食 物 的特色， 菜 的 口 味， 用
shénme zuò de chī qǐ lái zěnmeyàng děngděng
什 么 做 的， 吃 起来 怎 么 样 ， 等 等 ）

- nǐ gèng xǐ huan nǎzhǒng wèishénme
  你 更 喜 欢 哪 种 ， 为 什 么

qǐngxiǎng yī xiǎng rú guǒ yǒu yī tiān nǐ dāng le lǎoshī
3. 请 想 一 想 如 果 有 一 天 你 当 了 老 师 。

nǐ yīnggāi shuō yī shuō
你 应 该 说 一 说 ：

- nǐ huì jiāo shénme
  你 会 教 什么
- nǐ shì yígè shénmeyàng de lǎoshī bǐrú zěnme
  你 是 一个 什么样 的 老师 （比如： 怎么
  jiāo xuéshēng duì xuéshēng yǒuhǎo ma děngděng
  教 学生 ， 对 学生 友好 吗， 等 等 ）
- rú guǒ nǐ yǒubùhǎo hāoxué xí de xuéshēng nǐ gāizěn
  如 果 你 有 不 好 好 学 习的 学 生 ， 你 该 怎
  meban
  么 办



èr liǎnggèréndekǎoshì gěi de
二、 两个人的考试(给A的)

yī hé juésèbànyǎn
（一）A和B角色扮演

zàixiàmiàndeliǎnggèkǎoshìzhōng nǐmenkěnéngshì huòshì yòngnǐmennádào
在下面的两个考试中，你们可能是A或是B，用你们拿到
de kǎoshìzhǐshuōhuà qǐngbúyàokànbiérendekǎoshìzhǐ jiùxiàngpíngchángshuōhuà
的考试纸说话，请不要看别人的考试纸。就像平常说话
nàyàngshuō qǐngjǐnkěnéngdiduōshuōshuō rúguǒnǐmenyǒubùmíngbáidedìfāng qǐng
那样说，请尽可能地多说说。如果你们有不明白的地方，请
wènwèn tí
问问题。

zàizhōngwénjiǎohébùrènshide liáotiānér
4. 在中文角和不认识的B聊天儿

nǐshì nǐhé búrènshi
你是A，你和B不认识。

měigèxīngqīsānwǎnshang diǎn zàixuéxiàoyǒugè zhōngwénjiǎo zhōngwén
每个星期三晚上7点，在学校有个“中文角”。中文
jiǎoshìxuéxízhōngwéndexuéshēngliànxíshuōzhōngwéndedìfāng zàinàli yǒuxiērén
角是学习中文的学生练习说中文的地方。在那里，有些人
gēntāmenrènshiderénshuōhuà yǒuxiērénqùrènshixīnpéngyǒu gēnxīnpéngyǒuliáotiān
跟他们认识的人说话，有些人去认识新朋友，跟新朋友聊天
ér nǐshìdìyīcìqùzhègezhōngwénjiǎo zhèlǐméiyǒunǐrènshiderén nǐkànjiànle
儿。你是第一次去这个中文角，这里没有你认识的人。你看见了B，
jiùqùzhǎotā tāliáotiānérle nǐmenliáoleyìhuìer jiùrènshile hòulái nǐjiànyì nǐ
就去找他/她聊天儿了，你们聊了一会儿，就认识了。后来，你**建议你**
menliǎnggèzàiqùgēnbiérénshuōshuōhuàér duōliànxíliànxízhōngwén
们两个再去跟别人说说话儿，多练习练习中文。

rènshixīnpéngyǒu
认 识 新 朋 友 B

xiāndǎzhāohu
- 先 打招呼

xiānxiàng shuōshuō nǐ zìjǐ yěwèn yìxiēwèntí bǐrú xìngmíng cóng
- 先 向 B 说 说 你自己，也 问 B一些问题（比如： 姓名 ， 从

nǎgeguójiāláide láizhōngguózuòshénme děngděng
哪个 国家 来的，来 中国 做 什么 ， 等 等 ）

wènwèn duìzhōngguódeyìnxiàngzěnmeyàng bǐrú xǐhuanhuòbùxǐhuan
- 问 问 B对 中国 的 印象 怎么样 （比如： 喜欢 或 不 喜欢

zhōngguódenǎxiēfāngmiàn bìngshuō yī shuōshìbùshìtóng yì tā tā dehuà wèi
中国 的哪些 方面 ）， 并 说 一 说 是不是 同 意他/她的 话 ， 为

shénme
什 么 。

jiànyìzàigēnbiérénliáoliáotiānér
建议 再 跟 别人 聊 聊天儿

shuō yī shuō zìjǐ hěngāoxìnghé liáotiānér jiànyìzàiqùhébiérénshuōshuōhuà
- 说 一 说 自己 很 高兴 和B 聊天儿 ， **建议** 再 去和 别人 说 说话

èr liǎng gè rén de kǎoshì gěi de
二、 两个人的考试(给B的)

yī hé jué sè bànyǎn
（一）A和B角色扮演

zàixiàmiàn de liǎng gè kǎoshì zhōng nǐmen kěnéng shì huòshì yòng nǐmen ná dào
在下面的两个考试中，你们可能是A或是B，用你们拿到
de kǎoshì zhǐ shuōhuà qǐng búyào kàn bié rén de kǎoshì zhǐ jiù xiàng píngcháng shuōhuà
的考试纸说话，请不要看别人的考试纸。就像平常说话
nàyàng shuō qǐng jǐn kě néng di duō shuō shuō rúguǒ nǐmen yǒu bùmíng bái de dìfāng
那样说，请尽可能地多说说。如果你们有不明白的地方，
qǐngwènwèntí
请问问题。

zàizhōngwénjiǎohébùrènshide liáotiān ér
4. 在中文角和不认识的A聊天儿

nǐ shì nǐ hé bùrènshi
你是B，你和A不认识。

měigè xīngqīsān wǎnshang diǎn zàixué xiào yǒu gè zhōng wén jiǎo zhōng
每个星期三晚上7点，在学校有个“中文角”。中
wénjiǎoshìxué xí zhōngwén de xuéshēng liàn xí shuōzhōngwén de dì fāng zàinà li yǒuxiē
文角是学习中文的学生练习说中文的地方。在那里，有些
rén gēn tāmen rènshi de rén shuōhuà yǒuxiē rén qù rènshi xīn péngyǒu gēn xīn péngyǒu
人跟他们认识的人说话，有些人去认识新朋友，跟新朋友
liáotiānér nǐ shì dìyī cì qùzhège zhōngwénjiǎo zhèlǐ méiyǒu nǐ rènshi de rén xiāngēn
聊天儿。你是第一次去这个中文角，这里没有你认识的人。A先跟
nǐ shuōhuà nǐ menliáo le yì huì er jiùrènshi le
你说话，你们聊了一会儿，就认识了。

rènshixīnpéngyǒu
认识 新 朋友 A

- wènwèn nǐxiǎngzhīdàodehé tā tā yǒuguāndewèntí bǐrú wèishénmexué
  问 问 A你 想 知道 的和他/她 有关 的问题 （比如： 为什么 学
  zhōngwén yǐhòuxiǎngzuòshénme děngděng
  中 文， 以后 想 做 什么， 等 等）

zàigēnbiérénqùliáoliáotiānér
再 跟 别 人 去 聊 聊天儿

- wèn kěbùkěyǐ liúxiàwēixìnhàoděng
  问 A可不可以 留下 **微 信号** 等
- shuōyíxià zìjǐ xiǎng yǐhòuzàihé liáotiān
  说 一下自己 想 以 后 再和A 聊天

gěi de
给A的

qǐngpéngyǒu cānjiāwǎnhuì jí jièkāfēi jī
5. 请 朋 友B参加晚 会及借咖啡机

nǐshì nǐhé shìlǎopéngyǒu
你是A，你和B是老 朋 友

kuàiguòshèngdànjié le nǐzhèngzàizhǔnbèi yígè shèngdànjiédewǎnhuì shàngwán
快 过 圣诞节 了，你 正在 准备 一个 圣诞节 的 晚会 。 上 完
zhōngwénkèhòu nǐqǐng cānjiāzhègewǎnhuì rìqī héshíjiānshì yuè rì zhōusìxiàwǔ
中文 课后，你 请B参加 这个 晚会 。日期和 时间 是12月24日 周四 下午
huìzàiwǎnshang qiánwán dìdiǎnshì yī hàolóu jiàoshì érqiěyīnwéi nǐ
4:30，会在 晚上 10:00前 完， 地点 是一 号 楼101 教室 。而且 因 为你
méiyǒukāfēi jī nǐyòuzhīdào yǒu yí gè suǒyǐ nǐxiǎngxiàng tā tājiè
没 有 咖啡机,你又 知 道B有 一个， 所以你 想 向 他/她借。

qǐng cānjiāwǎnhuì
请 B参加 晚 会

- qǐng cānjiānǐdeshèngdànjiéwǎnhuì
  请 B参加你的 圣 诞 节 晚 会
- shuōyīshuōgēnwǎnhuì yǒuguāndewèntí bǐrú rìqī shíjiān dìdiǎnděng
  说 一 说 跟 晚 会 有 关 的问题（比如：日期， 时间 ， 地点 等 ）
- rúguǒ bùzhīdàoyàobùyàolái xiǎngyīxiǎngzěnmenéngràng yídìngláicānjiā
  如果B不知 道 要 不要 来， 想 一 想 怎么 能 让 B一定 来参加
  bǐrú kěyǐjiāohěnduōxīnpéngyǒu děngděng
  （比如：可以交 很 多 新 朋 友， 等 等 ）

jièkāfēi jī
借咖啡机

- wèn kěbùkěyǐjiènǐkāfēi jī
  问 B可不可以借你咖啡机
- rúguǒbùnéngjiè wènwèn yígèkāfēi jī yàoduōshǎoqián
  如果 不能 借， 问 问 B一个咖啡机要 多 少 钱

gěi de
给 B的

bèiyāoqǐngqùcānjiāwǎnhuì jí bèijièkāfēi jī
5. 被 邀 请 去 参 加 晚 会 及 被 借 咖 啡 机

nǐ shì nǐ hé shìlǎopéngyǒu
你 是 B，你 和 A 是 老 朋 友

kuài guò shèngdànjié le zhèngzài zhǔnbèi yígè shèngdànjié de wǎnhuì shàng wán
快 过 圣 诞 节 了，A 正 在 准 备 一 个 圣 诞 节 的 晚 会 。 上 完
zhōngwén kè hòu qǐng nǐ cānjiā zhège wǎnhuì érqiě yīnwéi tā tā méi yǒu kāfēi jī
中 文 课 后 ，A 请 你 参 加 这 个 晚 会 。 而 且，因 为 他/她 没 有 咖 啡 机，
xiǎngxiàng nǐ jiè dànshì nǐ de kāfēi jī huài le
想 向 你 借， 但 是 你 的 咖 啡 机 坏 了 。

bèiyāoqǐngqùcānjiāwǎnhuì
被 邀 请 去 参 加 晚 会

wèngēnwǎnhuìyǒuguāndewèntí bǐrú yǒushuícānjiā yǒushénmehuódòng
- 问 跟 晚 会 有 关 的 问 题（比 如：有 谁 参 加，有 什 么 活 动 ，
děngděng
等 等 ）

shuō zì jǐ hěnxiǎngcānjiāwǎnhuì dànshìnàtiānzàiwǎnshang diǎndeshíhòuhái
- 说 自 己 很 想 参 加 晚 会，但 是 那 天 在 晚 上 7 点 的 时 候 还
yǒu yí gèyuēhuì
有 一 个 约 会

bèijièkāfēi jī
被 借 咖 啡 机

gàosù zìjǐ dekāfēi jī huài le
- 告 诉 A 自 己 的 咖 啡 机 坏 了

bāng xiǎngxiǎngzěnmenéngnádào yí gèkāfēi jī bǐ rú kě yǐ bāngzhewènwèn
- 帮 A 想 想 怎 么 能 拿 到 一 个 咖 啡 机 （比 如：可 以 帮 着 问 问
biédepéngyǒu
别 的 朋 友 ）

èr qíngjǐnghuà tí tǎolùn
（二） 情 景 话 题讨 论

ài hǎo
6.1 爱 好 ：

nǐ menliǎnggèrénzàizhōngwénjiǎorènshi le yǐ hòu jiùchángcháng yì qǐ chūqù
你 们 两 个人 在 中文 角 认识 了以 后， 就 常 常 一起 出 去。
yǒuyītiān nǐ menyuēzhe yìqǐ qùmǎidōng xī děngchēdeshíhòuliáo qǐ le bùtóngdexìngqù ài
有一天 你 们 约 着 一起去 买 东 西， 等 车 的 时候 聊 起了不 同 的 兴趣 爱
hǎo nǐmen kěyǐ zìjǐ xuǎnzéshì háishì
好 。 你们 可以自己 选择 是A还 是B。

yàoliáo
要 聊 ：

zìjǐ yǒunǎxiē ài hǎo
- 自己 有 哪些爱 好 ？

bùtóngde ài hǎoyǒushénmehǎochu wèishénme
- 不同 的爱 好 有 什么 好处 ， 为什么 ？

bùtóngde ài hǎoyòuyǒushénmehuàichu wèishénme
- 不同 的爱 好 又 有 什么 坏处 ， 为什么 ？

nǐmen kěyǐ xuǎnwénzì jí zhàopiànzhōngde ài hǎo yě kěyǐ xuǎn zìjǐ xiǎngyàoshuōde
你们 可以 选 文字 及 照片 中 的爱 好， 也可以 选 自己 想 要 说 的
bié de ài hǎo jiùxiàngpíngchángshuōhuànàyàngshuō qǐngjǐnkěnéng di duōshuōshuō
别的爱 好。就 像 平常 说话 那样 说， 请 尽可 能 地 多 说 说 。
rú guǒ nǐ menyǒubùmíngbáide dì fāng qǐngwènwèn tí
如 果 你 们 有 不 明 白的地 方， 请 问 问 题。

dúshū
读 书

kàndiànyǐng
看 电 影

dǎdiàn zǐ yóu xì
打 电 子 游 戏

zuòyùndòng
做 运 动

lǚxíng
旅 行

èr qíngjǐnghuà tí tǎolùn
（二） 情 景 话 题讨 论

guójiā
6.2 国 家：

nǐ menliǎnggèrénzàizhōngwénjiǎorènshi le yǐhòu jiùchángcháng yì qǐchūqù
你 们 两 个人在 中文 角 认识了以后， 就 常 常 一起出去。
yǒuyītiān nǐmenyuēzhe yìqǐ qùmǎidōng xī děngchēdeshíhòuliáo qǐ le bùtóngdeguójiā
有一天 你 们 约 着一起去 买 东 西， 等 车的时 候 聊 起了不 同 的国 家。
nǐmenkě yǐ zì jǐ xuǎnzéshì háishì
你 们 可以自己 选 择是A还 是B。

nèiróngbāokuò
内 容 包 括：

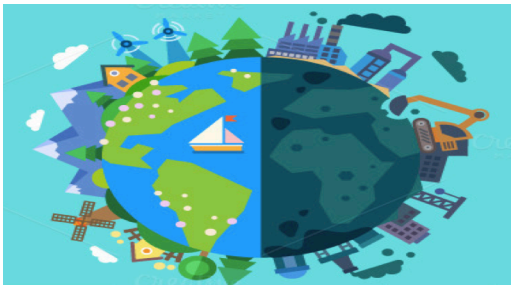
- zìjǐ deguójiāhézhōngguóyǒunǎxiēxiāngtóngdefāngmiàn
  自己的 国家 和 中国 有 哪些 相同 的 方面 ？
- zìjǐ deguójiāhézhōngguóyǒunǎxiēbùtóngdefāngmiàn
  自己的 国家 和 中国 有 哪些 不同 的 方面 ？
- guānyúzhèxiēbùtóngdefāngmiàn nǐgèngxǐhuannǎxiē
  关于 这些 不同 的 方 面 ，你 更 喜 欢 哪些？

nǐmen kěyǐ xuǎnzé wénzì jí zhàopiànzhōng de fāngmiàn láishuō yě kěyǐ xuǎnzé zìjǐ
你们 可以 选择 文字 及 照片 中 的 方面 来说 ， 也可以 选择 自己
xiǎngyàoshuō debié defāngmiàn qǐngzìrán dì duìhuà rú yǒubùmíngbái de dì fāng qǐng
想 要 说 的别的 方面 。 请 自然地 对话 ， 如 有 不 明 白的地 方 ， 请
suíshí tí wèn
随时提 问 。

gòu wù xí guàn
购 物 习 惯

jiāotōnggōng jù
交 通 工 具

yú lè ài hào
娱 乐 爱 好



huánjìngwèn tí
环 境 问 题

jiàoyùfāngshì
教 育 方 式

èr qíngjǐnghuà tí tǎolùn
（二） 情 景 话 题讨 论

chéngshì
6.3 城 市：

nǐ menliǎnggèrénzàizhōngwénjiǎorènshi le yǐhòu jiùchángcháng yì qǐchūqù
你 们 两 个人在 中文 角 认识了以后， 就 常 常 一起出去。
yǒuyītiān nǐmenyuēzhe yìqǐ qùmǎidōng xī děngchēdeshíhòuliáo qǐ le nǐmenzàixuǎnzé
有一天 你 们 约 着一起去 买 东 西， 等 车的时候 聊 起了你 们 在 选 择
wèiláishēnghuódechéngshìshíhuìkǎo lǜ defāngmiàn nǐmen kěyǐ zìjǐ xuǎnzéshì háishì
未来 生 活的 城 市时会 考虑的 方 面 。 你们可以自己 选 择是A还 是B。

nèiróngbāokuò
内容 包括 ：

- nǐzàixuǎnzéwèiláishēnghuódechéngshìshí huìkǎo lǜ nǎxiēfāngmiàn
  你在 选 择未来 生 活的 城 市时，会 考虑哪些 方 面 ？
- nǐxiànzàijūzhùdechéngshìcúnzàishénmewèntí
  你 现在 居住的 城市 存在 什么 问题？
- duìyú nǐxiànzàijūzhùchéngshìcúnzàidewèntí nǐjué deyǒushénmejiějuédebànfǎ
  对于你 现在 居住 城市 存在 的问题，你觉 得有 什么 解决的办法？

nǐmen kěyǐ xuǎnzé wénzì jí zhàopiànzhōng de fāngmiàn láishuō yě kěyǐ xuǎnzé zìjǐ
你们 可以 选择 文字 及 照片 中 的 方面 来说 ， 也可以 选择 自己
xiǎngyàoshuō de biéde fāngmiàn qǐng zì rán di duìhuà rú yǒubùmíngbáide dì fāng qǐng
想 要 说 的别的 方 面 。 请 自 然 地 对 话 ， 如 有 不 明 白的地 方 ， 请
suíshí tí wèn
随 时 提 问 。

huánjìngwūrǎn
环 境 污 染

chéngshìānquán
城 市安 全

shēnghuózhìliàng
生 活 质 量

jīng jì fā zhǎn
经 济发 展

jūmínyǒushàndù
居 民 友 善 度

Appendix C (continued)： The Speaking Test in English

Part one: Solo tasks

Instruction: Answer the questions in Chinese according to the prompts. You will have 30 seconds to prepare and 2 minutes to respond. Speak as much as possible. If you have any problems, please ask the administrator.

Task 1. Describe a place where you have traveled that impressed you greatly (e.g., a city or a famous tourist attraction).

You should talk about:

- Where it is (e.g., country or region)
- Unique aspects of it (e.g., scenic spot, historical site, food, architecture, etc.)
- Why it impressed you

Task 2. Compare the eating habits of two different places (e.g., northern and southern China, or your country and China).

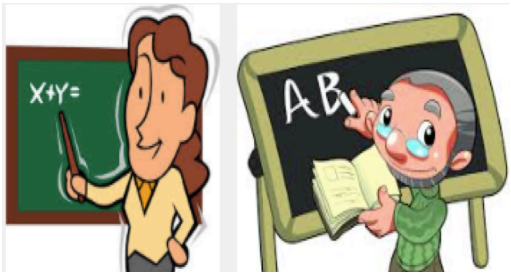
You should talk about:

- Where the two places are
- What the similarities and differences between them are (e.g., features of food, taste, ingredients, etc.)
- Which you prefer, and why

Task 3. Imagine one day you will become a teacher.

You should talk about:

- What you will teach
- What kind of teacher you will be (e.g., how to teach students, whether you are friendly, etc.)
- If you have students who do not study hard, how you will deal with it



Part two: Paired interactive tasks (for student "A")

One. Open role-play tasks

In the following two tasks, you may choose to be A or B. Fulfill the requirements talking as naturally as possible. Do not look at each other's prompts. If you have any problems, please ask the administrator.

Task 4. Chatting with a stranger at "Chinese corner"
You are A and don't know B.
"Chinese corner" is held on campus every Wednesday at 7:00 p.m. International students can practice Chinese there, either with people they are already familiar with, or with new friends. It is your first time going to this event, and you don't know anybody. You see B and begin to talk with him/her, and then you two chat for a while and get to know each other. Eventually, you suggest that you both should talk with others to gain more practice.

Getting to know new friend B

- Greet first
- Make self-introduction first, and also ask some questions about B (e.g., name, where he/she came from, why he/she came to China)
- Ask for B's impression of China (e.g., aspects he/she likes or dislikes about it)
- Agree or disagree with B, and explain why

Making a suggestion to talk with others

- Express your pleasure at communicating with B, and suggest that both of you talk with others

Part two: Interactive tasks (for student "B")

One. Open role-play tasks

In the following two tasks, you may choose to be A or B. Fulfill the requirements talking as naturally as possible. If you have any problems, please ask the administrator.

Task 4. Chatting with a stranger at "Chinese corner"
You are B and don't know A.
"Chinese corner" is held on campus every Wednesday at 7:00 p.m. International students can practice Chinese there, either with people they are already familiar with, or with new friends. It is your first time going to this event, and you don't know anybody. A begins to talk with you first, then the two of you chat for a while and get to know each other.

Getting to know new friend A

- Ask questions about A (e.g., why he/she studies Chinese, what he/she will do in the future)
- Respond to A naturally

Before ending the conversation

- Ask if A will exchange WeChat details with you
- Express that you would like to talk with A again in the future

For A

Task 5. Invite B to a party, and ask him/her if you can borrow a coffee machine
You are A. B is your old friend.
Christmas is coming soon, and you are preparing for an evening Christmas party. After Chinese class, you invite B to this party, which will take place from 4:30 p.m. to about 10:00 p.m. on Thursday, December 24th in Room 101, No. 1 Building. In addition, you don't have a coffee machine, and you know B has one. Thus you want to borrow it.

Inviting B to the party

- Invite B to your evening Christmas party
- Talk about details of the party (e.g., date, time, location)
- Try to persuade B to come if B is hesitant (e.g., B can make more new friends)

Borrowing the coffee machine

- Ask whether B can lend you his/her coffee machine
- If B cannot lend it to you, ask how much a coffee machine costs

For B

Task 5. Being invited to a party and asked to lend someone a coffee machine
You are B. A is your old friend.
Christmas is coming soon, and A is preparing for an evening Christmas party. After Chinese class, A invites you to this party. In addition, since he/she does not have a coffee machine, and knows you have one, A wants to borrow yours. However, your coffee machine is broken.

Being invited to the party

- Ask for more details about the party (e.g., who will join in, what activities it will have)
- Tell A that you do want to attend, but you have another appointment at 7:00 p.m. on that day

Being asked to lend the coffee machine

- Tell A your coffee machine is broken
- Suggest another way for A to obtain a coffee machine (e.g., offer to ask other friends)

Two. Situational topic discussion

Task 6.1: Hobby

You can be either A or B. Since getting to know each other at "Chinese corner", you have always taken part in activities together. One day, you are going grocery shopping together, and while at the bus stop, you begin to talk about different hobbies.

You should talk about:

- What your hobbies are
- What the advantages of different hobbies are, and why
- What the disadvantages of different hobbies are, and why

You may talk about the hobbies prompted by the following pictures, or choose any others. Speak as naturally as possible. If you have any problems, please ask the administrator.

Reading

Watching movies

Playing videogames

Doing exercise

Traveling

Two. Situational topic discussion (continued)

Task 6.2: Country

You can be either A or B. Since getting to know each other at "Chinese corner", you have always taken part in activities together. One day, you are going grocery shopping together and while waiting at the bus stop, you begin to talk about different countries.

You should talk about:

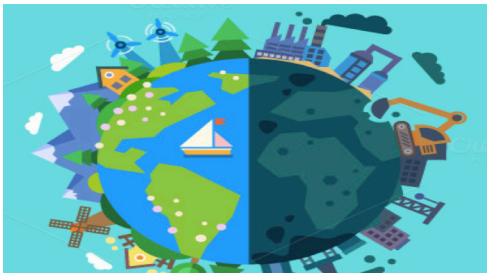
- What the similarities between your country and China are
- What the differences between your country and China are
- For each difference, state which country you prefer

You may compare the aspects prompted by the following pictures, or choose any others. Speak as naturally as possible. If you have any problems, please ask the administrator.

Shopping habits

Means of transportation

Recreations and entertainment



Environmental issues

Educational modes

Two. Situational topic discussion (continued)

Task 6.3: Urban livability

You can be either A or B. Since getting to know each other at "Chinese corner", you have always taken part in activities together. One day, you are going grocery shopping together and while waiting at the bus stop, you begin to talk about urban livability.

You should talk about:

- What factors determine whether a city is livable
- Problems with the city you live in
- Solutions to those problems

You may talk about the aspects prompted by the following pictures, or choose any others. Speak as naturally as possible. If you have any problems, please ask the administrator.

Environmental pollution

Social security

Life's quality

Economic development

Residents' friendliness

Appendix D: Rating Criteria for the Solo Proficiency Tasks

	Range	Accuracy	Fluency	Coherence
3	Has a good command of a broad range of language, allowing him/her to select a formulation to express him/herself clearly in an appropriate style on a wide range of general or leisure topics without having to restrict what is said.	Consistently maintains a high degree of grammatical accuracy; errors are rare, difficult to spot and generally corrected when they do occur.	Can express him/herself fluently and spontaneously, almost effortlessly. Only a conceptually difficult subject can hinder a natural, smooth flow of language.	Can produce clear, smoothly flowing, well-structured speech, showing controlled use of organizational patterns, connectors and cohesive devices.
2	Has enough language to get by, with sufficient vocabulary to express him/herself with some hesitation and circumlocutions, on topics such as family, hobbies and interests, work, travel, and current events.	Uses, reasonably accurately, a repertoire of common "routines" and patterns associated with more predictable situations.	Can keep going comprehensibly, though pauses for grammatical and lexical planning and repair are very evident, especially in longer stretches of free production.	Can link a series of shorter, discrete simple elements into a connected, linear sequence of points.
1	Has a very basic repertoire of words and simple phrases related to personal details and particular concrete situations.	Shows only limited control of a few simple grammatical structures and sentence patterns from a memorized repertoire.	Can manage very short, isolated, mainly pre-packaged utterances, with much pausing to search for expressions, to articulate less familiar words, and to repair communication.	Can link words or groups of words with very basic linear connectors like "and" or "then".

Appendix D (continued): Rating Criteria for the Paired Speaking Tasks

Score	Language Use	Situation	Turn-taking Organization	Sequence Organization	Topic Management
3	• Has a good command of a broad range of language • Consistently maintains a high degree of grammatical accuracy; errors are rare	• Consistently demonstrates full awareness of the situation • Reacts appropriately in line with the situation at all times	• Fluently interacts without awkward pauses or abrupt overlaps and interruptions • Frequently shows moderate turn length	• Throughout the interaction, next turns show full understanding of and correct response to the previous turns • Employs abundant and diverse response tokens • Conducts dispreferred actions in a way that minimizes "face" threats	• Always initiates and terminates topics naturally and smoothly • Fully develops not only his/her own but also his/her partner's topics • Shifts topic naturally, using cohesive devices

2	• Has enough language to get by, with sufficient vocabulary to express him/herself with some hesitation and circumlocutions • Uses, reasonably accurately, a repertoire of common "routines" and patterns	• Sometimes cannot demonstrate full understanding of the situation • Occasionally reacts inappropriately to the situation	• Sometimes uses short pauses, or abruptly overlaps or interrupts • Sometimes exhibits unusual turn length: too long or too short	• Sometimes, a next turn does not show a full understanding of or correct response to a previous turn • Employs a limited number of different response tokens • Does not always conduct dispreferred actions in a way that minimizes "face" threats	• Sometimes abruptly initiates or terminates a topic • Sometimes develops his/her own or the other party's topics in a simple way • Does not utilize clear transitional cues between topics
1	• Has a basic repertoire of words and simple phrases • Shows limited control of a few simple grammatical structures and sentence patterns	• Only demonstrates limited awareness of the situation • Reactions may not match the situation at all	• Shows noticeably long pauses between or within turns, or very abrupt overlaps or interrupts • Most or all turns are short	• Next turns often show a misunderstanding of and incorrect response to the previous turn • Employs few response tokens • Always conducts dispreferred actions in a straightforward way	• Always abruptly initiates or ends a topic • Rarely develops topics • Does not use transitional cues between topics

Appendix E：DA Transcription Conventions Adapted from Atkinson and Heritage (1984)

-	Abrupt cutoff
(.)	Pause shorter than 0.2 seconds
(n)	Long pause, with the length given in seconds
[	Starting point of overlap
=	A turn latched immediately onto the previous turn
~	Laughter
:	Extending the preceding sound
*	Unclear talk
Word	Louder sound
Word	Softer sound
Word	Changing tone
‘ ’	Wrong sound
“ ”	Wrong word
< >	Wrong grammar pattern

REFERENCES

Al-Gahtani, S., & Roever, C. (2012). Proficiency and sequential organization of L2 requests. *Applied Linguistics*, *33*(1), 42–65.

Antaki, C. (2002). "Lovely": Turn-initial high-grade assessments in telephone closings, *Discourse Studies*, *4*(1), 5–23.

Atkinson, J., & Heritage, J. (Eds.). (1984). *Structures of social action.* Cambridge: Cambridge University Press.

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

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

Bachman, L. F., & Palmer, A. S. (2010). *Language assessment in practice.* Oxford: Oxford University Press.

Baker, R. (1997*). Classical test theory and item response theory in test analysis.* Lancaster: Center for Research in Language Education, Lancaster University.

Bardovi-Harlig, K. (1999). Exploring the interlanguage of interlanguage pragmatics: A research agenda for acquisitional pragmatics. *Language Learning, 49*(4), 677–713.

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

Brown, A. (2005). *Interviewer variability in oral proficiency interviews.* Frankfurt am Main: Peter Lang.

Brown, A. (2006). Candidate discourse in the revised IELTS Speaking Test. In P. McGovern & S. Walsh (Eds.), *IELTS research reports 2006* (pp. 71–89). Canberra

& Manchester: IELTS Australia and British Council.

Brown, J. D. (1997). Statistics Corner: Questions and answers about language testing statistics: Skewness and kurtosis. *Shiken: JALT Testing & Evaluation SIG Newsletter, 1*(1), 16-18.

Brown, J. D. (2001). Pragmatics tests. In K. R. Rose & G. Kasper (Eds.), *Pragmatics in language teaching* (pp. 301–325). Cambridge: Cambridge University Press.

Brown, J. D. (2007). Statistics Corner. Questions and answers about language testing statistics: Sample size and power. *Shiken: JALT Testing & Evaluation SIG Newsletter, 11*(1), 31-35.

Brown, J. D. (2008). Statistics Corner. Questions and answers about language testing statistics: Effect size and eta squared. *Shiken: JALT Testing & Evaluation SIG Newsletter, 12*(2), 36-41.

Brown, J. D. (2012). Introduction to rubric-based assessment. In J. D. Brown (Ed.), *Rubrics in language assessment with case studies in Asian and Pacific languages* (pp.1–9). Honolulu: University of Hawai‘i, National Foreign Language Resource Center.

Brown, J. D. (2013). *Research for TESOL: Quantitative, qualitative, and mixed method*. Endinburgh: Edinburgh University Press.

Brown, J. D. (2014). *Mixed methods research for TESOL*. Edinburgh: Edinburgh University Press.

Brown, P., & Levinson, S. C. (1987). *Politeness: Some universals in language usage.* Cambridge: Cambridge University Press.

Brown, G., & Yule, G. (1983). *Discourse analysis*. Cambridge: Cambridge University

Press.

Button, G. (2018). Conversation-in-a-series. In D. Boden & D. H. Zimmerman (Eds.), *Talk and social structure: Studies in ethnomethodology and conversation analysis* (pp.251–277). Cambridge: Polity.

Button, G. & Casey, N. (1984). Generating the topic: The use of topic initial elicitors. In J. M. Atkinson & J. Heritage (Eds.), *Structures of social action: Studies in conversation analysis* (pp. 167–190). Cambridge: Cambridge University Press.

Button, G., & Casey, N. (1985). Topic nomination and topic pursuit. *Human Studies, 8*(1), 3–55.

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

Celce-Murcia, M. (2007). Rethinking the role of communicative competence in language teaching. In E.A. Soler, & M.P.S. Jorda (Eds.), *Intercultural language use and language learning* (pp. 41–57). Dordrecht: Springer.

Celce-Murcia, M., Dörnyei, Z., & Thurrell, S. (1995). A pedagogical framework for communicative competence: A pedagogically motivated model with content specifications. *Applied Linguistics*, *6*(2), 5–35.

Chalhoub-Deville, M. (2003). Second language interaction: current perspectives and future trends. *Language Testing*, *20*(4), 369–383.

Chapelle, C. A. (1998). Construct definition and validity inquiry in SLA research. In L. F. Bachman & A. D. Cohen (Eds.), *Interfaces between second language acquisition and language testing research* (pp. 32–70). Cambridge: Cambridge University Press.

Chapelle, C. A. (1999). Validity in language assessment. *Annual Review of Applied*

Linguistics, 19, 254–272.

Chapelle, C. A. (2008). The TOEFL validity argument. In C. A. Chapelle, M. K. Enright & J. M. Jamieson (Eds.), *Building a validity argument for the test of English as a foreign language* (pp. 319–352). New York: Routledge.

Cooper, D. & Schindler, P. (2014). *Business research methods*. Boston: McGraw-Hill/Irwin.

Coulmas, F. (1981). *Conversational routines: Exploration in standardized communication situations and prepatter end speech.* New York: Mouton Publishers.

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

Coupland, N. (1983). Patterns of encounter management: Further arguments for discourse variables. *Language in Society, 12,* 459–476.

Davidson, J. (1985). Subsequent versions of invitations, offers, requests, and proposals dealing with potential or actual rejection. In J. Atkinson (Ed.), *Structures of social action* (pp. 102–128). Cambridge: Cambridge University Press.

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

Davis, L., & Kondo-Brown, K. (2012). Assessing student language performance: Type and uses of rubrics. In J. D. Brown (Ed.), *Rubrics in language assessment with case studies in Asian and Pacific languages* (pp.33–55). Honolulu: University of Hawai‘i, National Foreign Language Resource Center.

D’Hondt, S. (2009). The pragmatics of interaction: A survey. In S. D’Hondt, J.-O.

Östman & J. Verschueren (Eds.), *The pragmatics of interaction* (pp. 1–19). Amsterdam: John Bejamins.

Dörnyei, Z. (2007). *Research methods in applied linguistics: Quantitative, qualitative, and mixed methodologies*. Oxford: Oxford University Press.

Douglas, D., & Selinker, L. (1985). Principles for language tests within the "discourse domains" theory of interlanguage. *Language Testing, 2,* 205–226.

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

Edwards, D. (1997). *Discourse and cognition*. London: Sage.

Edwards, J. A. (2001). The transcription of discourse. In D. Schiffrin, D. Tannen & H. E. Hamilton (Eds.), *The handbook of discourse analysis* (pp. 321–348). Malden: Blackwell.

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

Fulcher, G., & Davidson, F. (2007). *Language testing and assessment*. New York: Routledge.

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

Gage, N. L. (1989). The paradigm wars and their aftermath: A "historical" sketch of research on teaching since 1989. *Educational Researcher, 18*(7), 4–10.

Galaczi, E. (2004). *Peer-peer interaction in a paired speaking test: The case of the First Certificate in English.* (Unpublished doctoral dissertation). Retrieved from ProQuest Dissertations and Theses. (UMI No. 3117838)

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

Galaczi, E. D. (2010). Face-to-face and computer-based assessment of speaking: Challenges and opportunities. In L. Araújo (Ed.), *Proceedings of the computer-based assessment (CBA) of foreign language speaking skills* (pp. 29–51). Brussels: European Union.

Galaczi, E. D. (2014). Interactional competence across proficiency levels: How do learners manage interaction in paired speaking tests? *Applied Linguistics*, *35*(5), 553–574.

Galaczi, E. D., ffrench, A., Hubbard, C. and Green, A. (2011). Developing assessment scales for large-scale speaking tests: A multiple-method approach, *Assessment in Education, 18*(3), 217–237.

Gan, Z. (2010). Interaction in group oral assessment: A case study of higher- and lower-scoring students. *Language Testing*, *27*(4), 585–602.

Gardner, J. (2006). Assessment and learning: An introduction. In J. Gardner (Ed.), *Assessment and learning* (pp. 1–5). London: Sage.

Garfinkel, H. (1967). *Studies in ethnomethodology.* Englewood Cliffs: Prentice-Hall.

Gill, R. (2000). Discourse analysis. In M. W. Bauer & G. Gaskell (Eds.), *Qualitative researching with image, sound and text* (pp. 172–190). London: Sage.

Glass, G. V., Peckham, P. D., & Sanders, J. R. (1972). Consequences of failure to meet assumptions underlying the fixed effects analyses of variance and covariance. *Review of Educational Research, 42*(3), 237.

Grabowski, K. (2009). *Investigating the construct validity of a test designed to measure*

grammatical and pragmatic knowledge in the context of speaking. (Unpublished doctoral dissertation). Retrieved from ProQuest Dissertations and Theses. (UMI No. 3368256)

Goffman, E. (1983). The interaction order. *American Sociological Review, 48*, 1–17.

Guba, E. G., & Lincoln, Y. S. (1989). *Fourth generation evaluation*. London: Sage.

Gumperz, J. J. (1982). *Discourse strategies* (Vol. 1). Cambridge: Cambridge University Press.

Gumperz, J. J. (1996). The linguistic and cultural relativity of conversational inference. In J. J. Gumperz & S. C. Levinson (Eds.), *Rethinking linguistic relativity* (pp. 374–406). Cambridge: Cambridge University Press.

Gumperz, J. J. & Berenz, N. (1993). Transcribing conversational exchanges. In J. Edwards & M. Lampert (Eds.), *Talking data: Transcription and coding in discourse research* (pp.91–121). Hillsdale: Lawrence Erlbaum Associates.

Hair, J. F., Black, W. C., Babin, B. J. & Anderson, R. E. (2010). *Multivariate data analysis*. New York: Pearson.

Hall, J. (1995). (Re)creating our worlds with words: A sociohistorical perspective of face-to-face interaction. *Applied Linguistics, 16*(2), 206–232.

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

Heritage, J. (1984). A change-of-state token and aspects of its sequential placement. In J. M. Atkinson & J. Heritage (Eds.), *Structures of social action: Studies in conversation analysis* (pp. 299–345). Cambridge: Cambridge University Press.

Heritage, J. (1990). Intention, meaning and strategy: Observations on constraints on interaction analysis. *Research on Language and Social Interaction, 24*(1-4), 311–332.

Hudson, T. D., Detmer, E., & Brown, J. D. (1992). *A framework for testing cross-cultural pragmatics*. Honolulu: University of Hawai‘i, Second Language Teaching and Curriculum Center.

Hudson, T. D., Detmer, E., & Brown, J. D. (1995). *Developing prototype measures of cross- cultural pragmatics*. Honolulu: University of Hawai‘i, Second Language Teaching and Curriculum Center.

Hutchby, I. and R. Wooffitt. 1998. *Conversation analysis*. Cambridge: Polity.

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

Itakura, H. (2001). Describing conversation dominance. *Journal of Pragmatics*, *33*, 1859 –1880

Iwashita, N. (1998). The validity of the paired interview format in oral performance assessment. *Melbourne Papers in Language Testing*, *5*(2), 51–66.

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

Jang, E. (2005). *A validity narrative: Effects of reading skills diagnosis on teaching and learning in the context of NG TOEFL.* (Unpublished doctoral dissertation). Retrieved from ProQuest Dissertations and Theses. (UMI No. 3182288)

Jefferson, G. (1984). On stepwise transition from talk about a trouble to inappropriately next positioned matters in J. N. Atkinson & J. Heritage (Eds.), *Structures of*

social action: Studies in conversation analysis (pp. 191–222). Cambridge: Cambridge University Press.

Jefferson, G. (1993). Caveat speaker: Preliminary notes on recipient topic-shift implicature. *Research on Language and Social Interaction, 26*(1), 1–30.

Jefferson, G. (2002). Is "no" an acknowledgment token? Comparing American and British uses of (+)/(-) tokens. *Journal of Pragmatics, 34,* 1345–1383.

Johnson, R., & Onwuegbuzie, A. (2004). Mixed methods research: A research paradigm whose time has come. *Educational Researcher, 33*(7), 14–26.

Johnson, R., Onwuegbuzie, A., & Turner, L. (2007). Toward a definition of mixed methods research. *Journal of Mixed Methods Research, 1*(2), 112–133.

Kasper, G. (2006). Speech acts in interaction: Towards discursive pragmatics. In K. Bardovi-Harlig, J. C. Félix-Brasdefer & A. S. Omar (Eds.), *Pragmatics and language learning (Vol. 11)* (pp. 281–314). Honolulu: University of Hawai'i, National Foreign Language Resource Center.

Kasper, G. (2009). L2 pragmatic development. In W. C. Ritchie & T. K. Bhatia (Eds.), *New handbook of second language acquisition*. Leeds: Emerald.

Kasper, G., & Ross, S. J. (2013). Assessing second language pragmatics: An overview and introductions. In S. Ross & G. Kasper (Eds.), *Assessing second language pragmatics* (pp. 1–40). London: Palgrave Macmillan.

Kasper, G., & Schmidt, R. (1996). Developmental issues in interlanguage pragmatics. *Studies in Second Language Acquisition, 18*(2), 149–169.

Kane, M. T. (2006). Validation. In R. Brennan (Ed.), *Educational measurement* (pp. 17–64). Westport: Greenwood Publishing.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, *50*(1), 1–73.

Kley, K. (2015). *Interactional competence in paired speaking tests: Role of paired task and test-taker speaking ability in co-constructed.* (Unpublished doctoral dissertation). Retrieved from ProQuest Dissertations and Theses. (UMI No. 3711677)

Kormos, J. (1999). Simulating conversations in oral-proficiency assessment: A conversation analysis of role plays and non-scripted interviews in language exams. *Language testing*, *16*(2), 163–188.

Kramsch, C. (1986). From Language Proficiency to Interactional Competence. *The Modern Language Journal*, *70*(4), 366–372.

Krashen, S. D. (1981). The "fundamental pedagogical principle" in second language teaching. *Studia Linguistica*, *35*(1-2), 50–70.

Lampert, M., & Ervin-Tripp, S. (1993). Structured coding for the study of language and social interaction. In J. Edwards & M. Lampert (Eds.) *Talking data: Transcription and coding in discourse research* (pp. 169–206). Hillsdale: Lawrence Erlbaum Associates.

Lazaraton, A. (2002). *A qualitative approach to the validation of oral language tests.* Cambridge: Cambridge University Press.

Lee, Y., & Greene, J. (2007). The predictive validity of an ESL placement test: A mixed methods approach. *Journal of Mixed Methods Research, 1*(4), 366–389.

Lerner, G. H. (1996). Finding "face" in the preference structures of talk-in-interaction. *Social Psychology Quarterly*, *59*(4), 303–321.

Leech, G. (1983). *Principles of pragmatics*. Harlow: Longman.

Levinson, S. (2000). *Presumptive meanings: The theory of generalized conversational implicature*. Cambridge: MIT Press.

Liao, J. (2018). Acquisition and assessment of L2 Chinese speaking. In C. Ke (Ed.), *The routledge handbook of Chinese second language acquisition* (pp. 234–260). New York: Routledge.

Linell, P., Gustavsson, L., & Juvonen, P. (1988). Interactional dominance in dyadic communication: A presentation of initiative-response analysis. *Linguistics, 26*(3), 415–442.

Lu, Y. (2017). Exploring the criterion-validity of HSK Levels 3 and 4: Are assessments and CEFR standards related? In Y. Lu (Ed.), *Teaching and learning Chinese in higher education: Theoretical and practical Issues* (pp. 35–56). New York: Routledge.

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

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

McCarthy, M. (2010). Spoken fluency revisited, *English Profile Journal*, *1*(1), 1–15.

Maynard, D. W., & Zimmerman, D. H. (1984). Topical talk, ritual and the social organization of relationships. *Social Psychology Quarterly, 47*(4), 301.

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

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: Blackwell.

Meisel, J. (1980). Linguistic simplification. In S. Felix (Ed.), *Second language development: Trends and issues* (pp. 13–40). Tübingen: Gunter Narr.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (pp. 13–103). New York: Macmillan.

Noaks, L., & Wincup, E. (2004). *Criminological research: Understanding qualitative methods*. London: Sage.

Norris, J. M. (2008). *Validity evaluation in language assessment*. New York: Peter Lang.

Ockey, G. J. (2009). The effects of group members' personalities on a test taker's L2 group oral discussion test scores. *Language Testing*, *26*(2), 161–186.

Okada, Y., & Greer, T. (2013). Pursuing a relevant response in oral proficiency interview role plays. In S. Ross & G. Kasper (Eds.), *Assessing second language pragmatics* (pp. 288–310). London: Palgrave Macmillan.

Onwuegbuzie, A. J., & Johnson, R. B. (2006). The validity issue in mixed research. *Research in the Schools*, *13*(1), 48–63.

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

O'Sullivan, B., Weir, C., & Saville, N. (2002). Using observation checklists to validate speaking-test tasks. *Language Testing*, *19*(1), 33–56.

Pomerantz, A. (1978). Compliment responses: Notes on the co-operation of multiple constraints. In J. Schenkein (Ed.), *Studies in the organization of conversational*

interaction (pp. 79–112). New York: Academic Press.

Pomerantz, A. (1984). Agreeing and disagreeing with assessments: Some features of preferred/dispreferred turn shapes. In J. M. Atkinson & J. Heritage (Eds.), *Structures of social action: Studies in conversation analysis* (pp. 57–101). Cambridge: Cambridge University Press.

Pomerantz, A. and Fehr. B. J. (1997). Conversation analysis: An approach to the study of social action as sense making practices. In T. A. van Dijk (Ed.): *Discourse as social interaction,* (pp. 64–91). London: Sage.

Pomerantz, A. & Heritage, J. (2012). Preference, In J. Sidnell & T. Stivers (Eds.), *The handbook of conversation analysis* (pp. 210–228). Oxford: Wiley-Blackwell.

Riggenbach, H. (1998). Evaluating learner interactional skills: Conversation at the macro level. In R. Young & A. W. He (Eds.), *Talking and testing: Discourse approaches to the assessment of oral proficiency* (pp. 53–67). Amsterdam: John Benjamins Publishing.

Roever, C. (2006). Validation of a web-based test of ESL pragmalinguistics. *Language Testing*, *23*(2), 229–256.

Roever, C. (2011). Testing of second language pragmatics: Past and future. *Language Testing*, *28*(4), 463–481.

Ross, S. (1992). Accommodative questions in oral proficiency interviews. *Language Testing, 9*(2), 173–185.

Ross, S., & Kasper, G. (2013). *Assessing second language pragmatics*. London: Palgrave Macmillan.

Saville, N., & Hargreaves, P. (1999). Assessing speaking in the revised FCE. *ELT*

Journal, 53(1), 42–51.

Sacks, H. (1987). On the p for agreement and contiguity in sequences in conversation. In G. Button, & J. Lee (Eds.), *Talk and social organisation* (pp. 54–69). Clevedon: Multilingual Matters.

Sacks, H. & Schegloff, E. A. (1979). Two preferences in the organization of reference to persons in conversation and their interaction. In G. Psathas (Ed.), *Everyday language: Studies in ethnomethodology* (pp. 15–21). New York: Irvington Publishers.

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

Schegloff, E. A., Jefferson, G., & Sacks, H. (1977). The preference for self-correction in the organization of repair in conversation. *Language, 53*(2), 361–382.

Schegloff, E. A. (1993), Reflections on quantification in the study of conversation. *Research on Language and Social Interaction, 26*(1), 99–128.

Schegloff, E. A. (2007). *Sequence organization in interaction: A primer in conversation analysis*. Cambridge: Cambridge University Press.

Schegloff, E. A. & Sacks, H. (1973). Opening up closings. *Semiotica, 8*(4), 69–99.

Schiffrin, D. (1994). *Approaches to discourse.* Oxford: Blackwell.

Searle, J. R. (1969). *Speech act theory*. Cambridge: Cambridge University Press.

Searle, J. R. (1975). Indirect speech acts. In P. Cole & J. Morgan (Eds.), *Syntax and semantics* (Vol. 3): *Speech acts* (pp. 59–82). New York: Academic Press.

Searle, J.R. (1992). *The rediscovery of the mind*. Cambridge: The MIT Press.

Selinker, L. (1972). Interlanguage. *International Review of Applied Linguistics*, *10*, 209–

231.

Silverman, D. (2006). *Interpreting qualitative data: Methods for analyzing talk, text and interaction*. London: Sage.

Skehan, P. (2001). Tasks and language performance. In M. Bygate, P. Skehan & M. Swain (Eds.), *Research pedagogic tasks: Second language learning, teaching, and testing* (pp. 167–185). New York: Longman.

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

Taguchi, N., & Röever, C. (2017). *Second language pragmatics*. Oxford: Oxford University Press.

Tannen, D. (1982). Introduction. In D. Tannen (Ed.), *Analyzing discourse: Text and talk–Georgetown University round table on languages and linguistics 1981* (pp. ix–xiii). Washington, DC: Georgetown University Press.

Taylor, L. (2000). Issues in speaking assessment research. *Research Notes*, *1*, 8–9.

Taylor, L. (2001). The paired speaking test format: Recent studies. *University of Cambridge ESOL examinations research notes, 6*, 15–17.

Taylor, L. (Ed.). (2011). *Examine speaking: Research and practice in assessing second language speaking*. Cambridge: Cambridge University Press.

Teddlie, C., & Tashakkori, A. (2003). Major issues and controversies in the use of mixed methods in the social and behavioral sciences. In A. Tashakkori & C. Teddlie (Eds.), *Handbook of mixed methods in social & behavioral research* (pp. 3–50). London: Sage.

Teddlie, C., & Tashakkori, A. (2006). A general typology of research designs featuring

mixed methods. *Research in the Schools*, *13*(1), 12–28.

ten Have, P. (2007). *Doing conversation analysis: A practical guide.* London: Sage.

Thomas, J. (1983). Cross-cultural pragmatic failure, *Applied Linguistics*, *4*(2), 91–112.

van Lier, L. (1989). Reeling, writhing, drawling, stretching, and fainting in coils: Oral proficiency interviews as conversation. *TESOL Quarterly*, *23*(3), 489–508.

Walters, F. S. (2007). A conversation-analytic hermeneutic rating protocol to assess L2 oral pragmatic competence. *Language Testing, 24*(2), 155–183.

Walters, F. S. (2009). A conversation analysis-informed test of L2 aural pragmatic comprehension. *TESOL Quarterly*, *43*(1), 29–54.

Wang, L. (2015). *Assessing interactional competence in second language paired speaking tasks*. (Unpublished doctoral dissertation). Retrieved from ProQuest Dissertations and Theses. (UMI No. 3713923)

Waring, H. Z. (2008). Using explicit positive assessment in the language classroom: IRF, feedback, and learning opportunities. *The Modern Language Journal*, *92*(4), 577-594.

Wong, J., & Waring, H. Z. (2010). *Conversation analysis and second language pedagogy: A guide for ESL/EFL teachers*. New York: Routledge.

Weir, C. J., Vidaković, I., & Galaczi, E. D. (2013). *Measured constructs: A history of Cambridge English examinations, 1913-2012 (Vol. 37)*. Cambridge: Cambridge University Press.

Wells, G. (1981). *Learning through interaction: The study of language development (Vol. 1)*. Cambridge: Cambridge University Press.

Westhoff, G. (2007). Challenges and opportunities of the CEFR for reimagining foreign

language pedagogy. *The Modern Language Journal, 91*(4), 676–679.

Yang, L. (2018). Pragmatics learning and teaching in L2 Chinese. In C. Ke (Ed.) *The routledge handbook of Chinese second language acquisition* (pp. 261-278). New York: Routledge.

Youn, S. (2013). *Validating task-based assessment of L2 pragmatics in interaction using mixed methods*. (Unpublished doctoral dissertation). Retrieved from ProQuest Dissertations and Theses. (UMI No. 3577270)

Youn, S. (2015). Validity argument for assessing L2 pragmatics in interaction using mixed methods. *Language Testing*, *32*(2), 199–225.

Young, R. (1995). Conversational styles in language proficiency interviews. *Language Learning, 45*(1), 3–42.

Young, R. (2000). *Interactional competence: Challenges for validity*. Paper presented at the 2000 Language Testing Research Colloquium, Vancouver, Canada.

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

Young, R. (2011). Interactional competence in language learning, teaching, and testing. In E. Hinkel (Ed.), *Handbook of research in second language teaching and learning (Vol. 2)* (pp. 426–443). New York: Routledge.

Young, R. (2012). Social dimensions of language testing. In G. Fulcher & F. Davidson (Eds.), *The Routledge handbook of language testing* (pp. 178–193). New York: Routledge.

Young, R., & He, W. (Eds.). (1998). *Talking and testing: Discourse approach to the assessment of oral proficiency.* Philadelphia: John Benjamins Publishing.

Young, R., & Milanovic, M. (1992). Discourse variation in oral proficiency interviews. *Studies in Second Language Acquisition, 14*(4)*,* 403-424.