WILLINGNESS TO COMMUNICATE IN LEARNING MANDARIN
AS A FOREIGN AND HERITAGE LANGUAGE

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)
MAY 2012

By
Ying Zhou

Dissertation Committee:
Tao-Chung Yao, Chairperson
Kimi Kondo-Brown
Ying-Che Li
Cynthia Ning
Lourdes Ortega
ACKNOWLEDGEMENTS

I would like to thank my dissertation chair, Dr. Tao-Chung Yao, and committee members, Dr. Lourdes Ortega, Dr. Ying-che Li, Dr. Kimi Kondo-Brown, and Dr. Cynthia Ning for their constant support during the process of this dissertation. I especially wish to thank Dr. Yao and Dr. Li for their intellectual guidance and believing in me and encouraging me throughout my graduate studies. I would also like to thank Dr. Kondo-Brown, who gave me guidance on the methodology of the dissertation. I am also deeply indebted to Dr. Ortega, who guided me through the dissertation, offered invaluable feedback and suggestions, and showed most patience and encouragement during my dissertation.

I thank my classmate, Shu-Ling Wu, for working with me to develop the Mandarin elicited imitation task. I am also very grateful to my friends and classmates, Changqin Geng, Qiaona Yu, Shuhui Su, Yang Xiao, and Christopher Magriney, who helped me to collect data from their students. I also would like to thank my friend, Tanny Tang, for her encouragement and assistance in data collection of the dissertation.

My gratitude is also due to Richard Medina, the IT specialist at the University of Hawai‘i at Mānoa, who helped to develop a web-based Mandarin elicited imitation Task. I also thank Language Learning for the grant to develop a web-based questionnaire and purchase the software AMOS to analyze the data.

Finally and foremost, I thank my husband, Enning Zhang, for undertaking all the family responsibilities to allow me to spend my time on the dissertation and encouraging
me whenever I was in need of it. Without him, my dissertation would have been impossible.
ABSTRACT

The research investigated willingness to communicate (WTC) and its related antecedents and outcomes in learning Mandarin as a heritage and foreign language. The study examined the relationship between L2 WTC and two antecedents: the perceived competence and the communication anxiety. The outcomes of WTC investigated are the frequency of language use in Mandarin as an L2 and proficiency in the L2. The study also compared the differences between heritage and non-heritage learners of Mandarin on their L2 WTC and its antecedents and outcomes. One hundred and seventy-nine Mandarin learners at five universities in the United States participated in the study. They are English native speakers and learners of Mandarin at various proficiency levels. Among them, 78 are heritage learners of Mandarin and the remaining 101 participants do not have any heritage background of Mandarin. The data were collected via a questionnaire to measure their language background, L2 WTC, and the two antecedents and one outcome (frequency of language use) for L2 WTC. In addition, a Mandarin elicited imitation task developed by the researcher and a colleague was used to measure the participants’ Mandarin proficiency levels. The resulting evidence was analyzed and inspected via AMOS, the structural equation modeling (SEM), to test causal relationships among the variables. ANOVA was conducted to compare the differences between heritage and non-heritage learners on the variables. The result shows that L2 communication anxiety predicts learners’ perceived competence. When learners have higher perceived competence, they are more willing to communicate in Mandarin in classroom which leads to higher WTC outside classroom. Learners with higher WTC outside classroom actually use Mandarin more frequently and are able to achieve higher
proficiency levels. L2 perceived competence is also important in directly predicting learners’ frequency of use and proficiency levels of Mandarin. In addition, the evidence suggests a tendency for heritage language learners towards higher communication anxiety and lower WTC and frequency of use in Mandarin than non-heritage language learners. Suggestions are provided to reduce communication anxiety and enhance perceived competence, L2 WTC and eventually improve proficiency levels in classrooms of Mandarin learning.
# TABLE OF CONTENTS

Acknowledgements.................................................................................................................. ii

Abstract ........................................................................................................................................ iv

List of tables.................................................................................................................................... xi

List of figures................................................................................................................................... xiv

List of abbreviations ....................................................................................................................... xv

Chapter 1 Introduction ................................................................................................................. 1

1.1 Statement of the research problem ............................................................................................. 2

1.2 The purpose of the present study ............................................................................................... 5

1.3 Outline of the dissertation ........................................................................................................ 8

Chapter 2 Willingness to communicate construct ......................................................................... 11

2.1 Willingness to communicate ...................................................................................................... 11

2.1.1 Willingness to communicate in L1 ...................................................................................... 11

2.1.2 Willingness to communicate in L2 ...................................................................................... 15

2.2 Perceived competence .............................................................................................................. 18

2.3 Communication anxiety ........................................................................................................... 20

2.4 Synthesis study of L2 WTC ...................................................................................................... 24

2.4.1 Problem specification with the primary studies on L2 WTC .............................................. 26

2.4.2 Purpose of present research synthesis ............................................................................... 29

2.4.3 Method ................................................................................................................................. 29

2.4.3.1 Study identification and retrieval ................................................................................... 29
2.4.3.2 The inclusion and exclusion criteria ........................................31
2.4.3.3 Coding of the primary studies ..............................................32
2.4.4 Methodological findings .........................................................34
  2.4.4.1 Participants’ characteristics ..............................................34
  2.4.4.2 Research design ..............................................................37
  2.4.4.3 Instruments ..................................................................39
2.4.5 Substantive Findings ...............................................................41
  2.4.5.1 The roles of perceived competence and communication anxiety on L2 WTC ...........................................41
  2.4.5.2 The relationship between L2 WTC and frequency of L2 use ....44
2.4.6 Discussion and conclusion .......................................................46
Chapter 3 Heritage language learners ..............................................51
  3.1 Definition of heritage language learners ..................................51
  3.2 The primary studies on affective factors of heritage language learners ....56
  3.3 Summary ............................................................................60
Chapter 4 Elicited imitation as a measure of global proficiency of Mandarin as a second/foreign language ........................................63
  4.1 Problem specification with the primary studies on L2 EI .............65
  4.2 Purpose of present research synthesis ......................................69
  4.3 Method ...............................................................................70
    4.3.1 Study identification and retrieval ......................................70
    4.3.2 The inclusion and exclusion criteria ..................................72
    4.3.3 Coding of the primary studies ..........................................73
  4.4 Results ...............................................................................76
6.1 Analysis..........................................................................................................................................127
  6.1.1 Descriptive findings of participants ........................................................................................127
  6.1.2 Analysis on instruments ........................................................................................................135
    6.1.2.1 Elicited imitation task ........................................................................................................135
    6.1.2.2 L2 WTC ..........................................................................................................................137
  6.1.3 Findings of L2 WTC construct ...............................................................................................139
    6.1.3.1 Model 1 ..........................................................................................................................139
      6.1.3.1.1 Model identification ....................................................................................................140
      6.1.3.1.2 Goodness-of-fit indices of Model 1 ...........................................................................142
      6.1.3.1.3 Model misspecification ..............................................................................................144
    6.1.3.2 Modified models ..............................................................................................................145
  6.1.4 Findings of the affective factors in individual l2 use contexts ..............................................150
  6.1.5 Findings on heritage language learners of Mandarin ..............................................................151
  6.1.6 Summary of the findings ......................................................................................................153
6.2 Discussion........................................................................................................................................155
  6.2.1 L2 WTC construct .................................................................................................................155
  6.2.2 Heritage and non-heritage language learners of Chinese .....................................................161
  6.3 Summary .....................................................................................................................................165
Chapter 7 Conclusion ......................................................................................................................166
  7.1 Significance of the present study ...............................................................................................166
  7.2 Pedagogical implication ............................................................................................................169
    7.2.1 Identifying sources of language anxiety ............................................................................169
    7.2.2 Alleviating language anxiety in Mandarin classrooms ......................................................173
7.2.2.1 Alleviating language anxiety on proposed sources ..................173
7.2.2.2 Alleviating language anxiety on grammar and vocabulary ..........176
7.2.2.3 Alleviating language anxiety on four skills ..............................178
7.3 Limitations of the study .........................................................180
7.4 Suggestions for future studies ................................................182

Appendix A Coding variables for L2 WTC ..................................184
Appendix B Coding variables for EI task .........................................186
Appendix C Mandarin EI task .......................................................188
Appendix D EI task scoring rubric .................................................190
Appendix E Survey ......................................................................193
References ...............................................................................198
LIST OF TABLES

2.1 Online studies retrieval through Cambridge Scientific Abstracts (CSA) ..................30
2.2 Features of participants in the primary studies ..................................................36
2.3 Features of research design in the primary studies ............................................38
2.4 Correlations between WTC and perceived competence and communication anxiety ..........................................................43
2.5 Correlation between frequency of L2 use and WTC ...........................................45
4.1 Online studies retrieval through online database .................................................71
4.2 Descriptive data of participants’ n-size ..................................................................77
4.3 Participants’ characteristics in 20 independent samples of the primary studies .......79
4.4 Language features of the primary studies ............................................................81
4.5 Design features of EI tasks I .................................................................................83
4.6 Design features of EI tasks II .................................................................................86
4.7 Reliability of EI task .............................................................................................90
4.8 Correlation between EI and pairing measure(s) ......................................................91
5.1 Excluded participants in data analysis ..................................................................107
5.2 Item discrimination analysis of Mandarin EI task .................................................113
5.3 Descriptive statistics of EI scores ..........................................................................114
5.4 Descriptive statistics of L2 contact ......................................................................116
5.5 High-low proficiency group comparison of EI and L2 contact ...............................117
6.1 Biographic data of the participants ......................................................................127
6.2 Participants’ strongest languages ..........................................................................128
6.3 Participants who have visited the regions of the target language .........................131
6.4 Length of visits to the regions of the target language (month) .................................................131
6.5 Length of study in the regions of the target language (year) .....................................................131
6.6 Individual Mandarin use contexts .................................................................................................132
6.7 Individual Mandarin use contexts in other occasions .................................................................133
6.8 Differences of heritage and non-heritage learners of Mandarin in individual L2 use contexts..........................................................................................................................134
6.9 Descriptive statistics of EI and L2 contact scores ........................................................................136
6.10 Reliability of the instruments ........................................................................................................137
6.11 Descriptive data of L2 WTC ..........................................................................................................138
6.12 Pearson correlation between L2 WTC ..........................................................................................138
6.13 Effect size of differences in each skill inside and outside classroom ..........................................139
6.14 Assessment of normality of Model 1 ............................................................................................141
6.15 AMOS output for parameter estimates of Model 1 .....................................................................142
6.16 Goodness-of-fit indices of Model 1 ...............................................................................................143
6.17 Standardized residual covariance of Model 1 ..............................................................................144
6.18 Modification Indices of Model 1 ....................................................................................................145
6.19 Goodness-of-fit indices of Model 2 ...............................................................................................145
6.20 Modification indices of Model 2 ...................................................................................................145
6.21 Goodness-of-fit indices of Model 4 ...............................................................................................146
6.22 AMOS output for parameter estimates of Model 4 .....................................................................147
6.23 AMOS output for parameter estimates of the final model ..........................................................148
6.24 Goodness-of-fit indices of the final model ..................................................................................149
6.25 Correlation between L2 PC, L2 CA and L2 WTC in individual L2 use context ............................151
6.26 Descriptive data of heritage and non-heritage language learners .......................152
6.27 ANOVA results of heritage and non-heritage language learners .........................153
LIST OF FIGURES

2.1 Hypothesized causal sequences for predicting WTC using personality-based variables .......................................................... 14

2.2 MacIntyre’s heuristic model of variables influencing WTC ................................................................. 17

6.1 Scatter plot of inter rater correlation ......................................................................................... 136

6.2 Path diagram of Model 1 ........................................................................................................ 140

6.3 Path diagram of Model 5 (Final model) ........................................................................ 148

6.4 L2 WTC construct in the final model ..................................................................................... 149
LIST OF ABBREVIATIONS

ACTFL: The American Council on the Teaching of Foreign Languages
AGFI: The Adjusted Goodness-of-Fit Index
AMOS: Analysis of Moment Structures
AP: Chinese Advanced Placement test
CA: Communication anxiety
CFI: The Comparative Fit Index
CLT: Communicative Language Teaching
CSA: Cambridge Scientific Abstracts
EI: Elicited imitation
ERIC: Educational Resources Information Center
FLCAS: Foreign Language Classroom Anxiety Scale
GFI: Goodness-of-Fit Index
HSK: Hanyu Shuiping Kaoshi
ID: Item Discrimination
IELTS: International English Language Testing System
L1: First languages
L2: Second (or foreign; also third, fourth, and so on) language
LLBA: Linguistics and Language Behavior Abstracts
LLL: The College of Language, Linguistics, and Literature
MI: Modification Indices
PC: Perceived competence
PRCA: The Personal Report of Communication Apprehension
RMSEA: The Root-Mean-Square Error of Approximation
<table>
<thead>
<tr>
<th>Acronym</th>
<th>Full Form</th>
</tr>
</thead>
<tbody>
<tr>
<td>SEM</td>
<td>Structural Equation Model</td>
</tr>
<tr>
<td>SOPI</td>
<td>Simulated Oral Proficiency Interviews</td>
</tr>
<tr>
<td>SPSS</td>
<td>Statistical Package for the Social Sciences</td>
</tr>
<tr>
<td>UCS</td>
<td>Unwillingness to Communicate Scale</td>
</tr>
<tr>
<td>UG</td>
<td>Universal Grammar</td>
</tr>
<tr>
<td>WAU</td>
<td>Web Audio Utility</td>
</tr>
<tr>
<td>WTC</td>
<td>Willingness to communicate</td>
</tr>
</tbody>
</table>
CHAPTER 1
INTRODUCTION

The importance of communicating in a second language (L2) has been recognized by researchers in the field of second language acquisition and pedagogy since the 1960s. Hymes in 1966 first discussed publicly the notion of communicative competence in a conference on language development among disadvantaged children (Hymes, 1966). Communicative Language Teaching (CLT) started to become the trend of language teaching in the 1970s (Richards, 2006). Since then, researchers and teachers have attempted to seek a way to teach a second language in a communicative approach and help language learners to produce meaningful output in the language. Savignon (1972) did the first study to compare the effects of various practices on communicative skills development. Richards and Rogers (1986) posited three principles of CLT: the communication principle, the task principle, and the meaningfulness principle (p. 72). Lee (2000), Rulon and McCreary (1986), and Thompson (1996) suggested pair and group work could facilitate CLT in a second language classroom. Yang and Cheung (2003) emphasized the importance of meaningful and authentic tasks in CLT. Richards (2006) highlighted activities focusing on fluency as well as the role of a teacher as a facilitator and monitor in CLT. Furthermore, researchers (e.g., De Bot, 1996; Larsen-Freeman, 1986) also investigated types of feedback that would enhance CLT and encourage language learners to communicate in target languages.

However, previous studies on L2 communication largely emphasized the effects of an approach on a group of students and ignored the role of differences existing in learners as individual units on communication in a target language (Dörnyei & Skehan,
It was not until the 1990s that researchers started to consider the possible effects of individual differences on language learners’ willingness to communicate in a second language (e.g., MacIntyre & Charos, 1996). The concept of willingness to communicate (WTC) was introduced into the field of second language acquisition by MacIntyre and Charos in 1996. Willingness to communicate in L2 is defined as “a readiness to enter into discourse at a particular time with a specific person or persons, using a L2” (MacIntyre, Clément, Dörnyei, & Noels, 1998, p. 547). The importance of L2 WTC was stated by MacIntyre and his colleagues (1998) as follows:

The ultimate goal of the learning process should be to engender in language students the willingness to seek out communication opportunities and the willingness actually to communicate in them. That is, a proper objective for L2 education is to create WTC. A program that fails to produce students who are willing to use the language is simply a failed program. (p. 547)

1.1 Statement of the Research Problem

During the last decade, researchers conducted studies to investigate the role of WTC on second language communication behavior (e.g., Clément, Baker, & MacIntyre, 2003; Yashima, Zenuk-Nishide, & Shimizu, 2004). They also sought to find the antecedents of WTC, which are the factors predicting WTC, in a second language and attempted to shed light on the ways that language teachers can increase learners’ WTC and output in a target language (e.g., Baker & MacIntyre, 2000; MacIntyre, Baker, Clément, & Donovan, 2003). The antecedents that each study investigated were diverse, including perceived competence (PC), communication anxiety (CA), motivation, attitude,
international posture, and so on (e.g., Baker & MacIntyre, 2000; Liu & Jackson, 2008; Yashima, 2002). Nevertheless, the ultimate outcomes of WTC is a topic that relatively fewer studies have visited. Therefore, if the L2 WTC research is to be really convincing to teachers, then researchers need to explore more systematically and extensively how the construct of WTC leads to tangible outcomes and substantially affects the levels of proficiency attained in an L2.

In order to examine the effects of L2 WTC on language learning, it is necessary to measure language learners’ competence and test whether L2 WTC leads to high language proficiency. Hanyu Shuiping Kaoshi (HSK) and Chinese Advanced Placement test (AP) are the two biggest standardized exams of Mandarin in the United States. HSK was developed by Beijing Language Institute and is administered through the Confucius Institutes. The first Confucius Institute in the United States was founded on November 17, 2004 at the University of Maryland (Hanban, 2011). AP was developed by the College Board and provides secondary school students a chance to advance to college language courses. The first Chinese AP test was administered in 2007 and 2,694 students from public high schools in the United States participated in the test in 2009 (The College Board, 2010). Apparently, the two tests are still relatively new in the United States and at the current stage it is difficult for researchers who would like to measure proficiency in their studies to recruit a large number of participants with existing scores of HSK or AP. It is also not practical for researchers to request participants to take HSK or Chinese AP test for the purpose of the research only, particularly because of costs and time. The registration fee for HSK costs from $30 to $80 and takes 135 to 155 minutes to complete the test depending on the levels of the test. It takes $87 and approximately 135 minutes to
complete a Chinese AP test (The College Board, 2011). Therefore, a convenient and quick measure is indeed in need for the purpose of measuring proficiency in research in second language acquisition.

The research on L2 WTC mainly focused on commonly-taught languages, such as English, French, and Spanish. Mandarin, on the other hand, has gained increased weight in curriculums of foreign languages in K-12 in the United States. According to the report by the American Council on the Teaching of Foreign Languages (The ACTFL, 2011), in 2004 and 2005, there were 20,292 enrollment in Chinese language and in 2007-2008, the enrollment increased to 59,860 (p. 5). However, Chinese language still has not received sufficient attention in the research field of second language acquisition. Therefore, it is not surprising to find out that to this date only one study has targeted WTC in learning Mandarin as a second language. Mandarin carries its unique features in linguistics. It is an isolating language which contains few inflections or case markers for learners to figure out sets of rules to follow (Norman, 1988). Mandarin also has a tone system that language learners usually find difficult to acquire. Chinese orthography increases difficulty in learning Mandarin, especially for students without an East Asian background. The linguistic challenges of Mandarin may influence learners’ affective factors. At the same time, culture of a target language is also brought into language classrooms through the language and language teachers and has an additional impact on learners’ affective factors of language learning, including their L2 WTC. Traditional Chinese culture and value system influence language teachers’ approach to language teaching, for example, the Mandarin learning classroom is likely to be more teacher-centered, instead of learner-centered. These factors may also cause difficulties in enhancing WTC in Mandarin
classrooms. Therefore, the results in previous studies cannot be generalized in classrooms where Mandarin is taught as a second, heritage, or foreign language. The need for studies on WTC in Mandarin is obvious.

The issue of heritage language learners in Mandarin learning is ubiquitous in all classrooms and should also be considered in research, due to large number of ethnic Chinese in the United States. The Migration Policy Institute (2010) reported 3.2 million Chinese ethnics resided in the United States in 2008, including 1.2 million native-born US citizens with Chinese ancestry. According to the Migration Policy Institute (2010), there were 1.0 million native-born US citizens with Chinese ancestry at age 5 and older, and 39.4 percent of those reported speaking Chinese, including Chinese dialects, at home. In addition, about 420,000 children under age 18 resided in a household with a Chinese immigrant parent. These statistics support the claims by researchers (e.g., Chao, 1997; Kondo-Brown, 2003; Kondo-Brown & Brown, 2008; Shen, 2003) that more attention needs to be drawn towards heritage language learners in a second language classroom. Heritage language learners may differ greatly from non-heritage language learners in linguistic competence, cultural awareness, and affective factors in learning a language. However, previous studies on heritage language learners did not examine the effects of L2 WTC on heritage language learners.

1.2 The Purpose of the Present Study

The present study intends to investigate the WTC in learning Mandarin as a second language. Based on the L2 WTC heuristic model established by MacIntyre et al. (1998) and the results retrieved from previous studies, the present study will examine the
relationship between L2 WTC and two antecedents: L2 perceived competence and L2 communication anxiety. By investigating the causal relation among these two antecedent factors and WTC, the study aims to provide theoretical evidence about L2 WTC in Mandarin as well as pedagogical suggestions for teachers to increase their students’ willingness to pursue active use of the L2 in their language classrooms. The outcomes of WTC investigated in the study, in turn, are the frequency of language use in Mandarin as an L2 and actual proficiency in the L2. The hypothesis is that the higher the WTC is, the more frequently a learner tends to use Mandarin, which leads to a higher proficiency level. By investigating the outcomes of WTC, the study hopes to provide theoretical support for researchers and teachers to attach importance to enhancing learners’ willingness to use the L2 in language classrooms. In order to compare with previous research, the present study investigates the relationship of L2 perceived competence and L2 communication anxiety to L2 WTC in different individual L2 use contexts at home, at school outside the Mandarin classroom, in the community, and other. The current study also examines heritage language learners of Mandarin and attempts to find out differences of L2 PC and L2 CA for heritage versus non-heritage language learners. It also hopes to provide pedagogical suggestions for teachers of heritage language learners on how to improve these students’ active strategies for using the L2 maximally in the classroom and in the community. In order to address the need for proficiency measures that are feasible for research use, a Mandarin elicited imitation (EI) task developed and piloted by the researcher and a colleague (Zhou & Wu, 2009) is used to measure language learners’ proficiency levels. The Mandarin EI task is a parallel task of the ones originally developed by Ortega and her associates (in preparation) for German, English,
Spanish, and Japanese as second languages. The present study will test the reliability and validity of the Mandarin EI task and attempts to suggest a quick and reliable measure for researchers in future studies on learning Mandarin as a second language.

The study recruited 208 participants from Mandarin classrooms of five colleges in the United States. All the data were collected through the years of 2009 and 2010. Part of the data was collected through the administration of the researcher herself and part was collected through the Mandarin instructors of the colleges. The participants were required to complete an EI task that measured their proficiency in Mandarin as the first investigated outcome of WTC and a survey which provided the data on affective factors of Mandarin learning, including the main construct of L2 WTC, its two antecedents of L2 PC and L2 CA, as well as the second outcome of L2 WTC investigated in the study, which is frequency of L2 use. The survey took approximately thirty minutes. The entire data collection procedure, including the Mandarin EI task, took approximately forty minutes. The number of participants who completed all parts of the data collection is 179. Therefore, 179 participants were included in the data analysis and discussion.

Analysis of Moment Structures (AMOS, Arbuckle, 2011), one of Structural Equation Model (SEM) (Byrne, 2010; Schumacker & Lomax, 2004), was employed to examine the causal relationship within the affective factors (i.e., the two antecedents of L2 PC and L2 CA and the main construct of WTC) and between these affective factors and the outcomes of WTC: frequency of L2 use and proficiency. Statistical Package for the Social Sciences (SPSS) was used to investigate the differences of L2 PC and L2 CA in L2 WTC in different individual L2 use contexts, as well as all factors between heritage and non-heritage language learners. The reliability and validity of the Mandarin EI task
were also calculated through SPSS. The goals of the present study were to (a) establish a model of L2 WTC with its antecedents and outcomes, (b) confirm the centrality of L2 WTC in successful Mandarin L2 learning and provide suggestions for enhancing L2 WTC, (c) investigate differences between heritage and non-heritage language learners and provide pedagogical suggestions, and (d) examine the reliability and validity of the Mandarin EI task in hopes to provide researchers with a short and convenient measure of proficiency to use in future studies of L2 Mandarin as well.

1.3 Outline of the Dissertation

The present dissertation is organized into seven chapters. Chapter 1 highlights the importance of communicative competence in language learning and points out the limitation of the research on communication in a second language. The chapter introduces the concept of L2 WTC and its structure of antecedents and outcomes. Further, the purposes of the present study are explained and the data collection procedure is briefly introduced. In the end, the chapter presents the outlines of the dissertation to guide readers through the study.

Chapter 2 reviews the literature of the affective factors, which include the main construct and its two posited antecedents: L2 WTC, L2 PC, and L2 CA. It introduces the developmental history and provides theoretical background of the three affective factors. It also discusses the relationship between the three factors in the literature. The chapter presents the commonly-used survey scales for the three factors and discusses strengths and limitations of each scale. In order to better explore L2 WTC in previous studies, the chapter reports on a synthesis study on the previous primary studies of L2 WTC. It compares and analyzes the studies and identifies the gaps existing in the primary studies.
Chapter 3 reviews the literature of heritage language learners. It goes over the definition of heritage language learners and discusses the problems in defining and identifying heritage language learners of Mandarin. In addition, this chapter reviews previous literature on affective factors of heritage language learners.

Chapter 4 explores the literature on elicited imitation tasks. This chapter introduces controversial opinions on using EI as a measure and presents three major criticisms brought up by Bley-Vroman and Chaudron (1994). The chapter then conducts a synthesis involving EI as a measure of proficiency. It compares and analyzes the designs and grading procedures of EI tasks in the primary studies and discusses the problems that might lead to the criticisms. At the same time, this chapter proposes possible ways to minimize the problems and enhance the validity and reliability of EI tasks.

Chapter 5 and Chapter 6 discuss the current study. Chapter 5 states the research questions of the study and it describes the participants, instruments, and data collection procedures. This chapter also presents the pilot study on Mandarin EI task (Zhou & Wu, 2009) and discusses the ways the current EI task was used to minimize the problems mentioned in Chapter 4. Chapter 6 starts with the descriptive statistics of the participants, including gender, current age, length of formal instruction of Mandarin, family background of Chinese languages, and experiences of visiting a region speaking Chinese. It then calculates the reliability of the scores resulting from all instruments used to collect the data, including EI, L2 WTC, L2 PC, L2 CA, and L2 frequency of use. SPSS is used to investigate the differences across affective factors between heritage and non-heritage language learners, and then conducts AMOS to examine the relationship between the
variables in the study. Based on the analysis, the chapter continues to discuss the results of the study. The last chapter, Chapter 7, concludes the study, proposes pedagogical suggestions, points out limitations, and provides suggestions for future studies on L2 WTC of learning Mandarin.
CHAPTER 2
WILLINGNESS TO COMMUNICATE CONSTRUCT

Willingness to communicate (WTC) is originated from the concept of unwillingness to communicate in first languages (L1) (Burgoon, 1976). WTC was introduced into the field of second language acquisition by MacIntyre and Charos (1996). The WTC construct includes the antecedents of perceived competence (PC) and communication anxiety (CA), and the outcome of frequency of language use. This chapter will review the theoretical background and measuring instruments of WTC and its two antecedents. The chapter also synthesizes the previous primary studies on L2 WTC.

2.1 Willingness to Communicate

2.1.1 Willingness to Communicate in L1

Burgoon (1976) defined the unwillingness to communicate as the predisposition which "represents a chronic tendency to avoid and/or devalue oral communication" (p. 60). She believed that people with the predisposition of anomie, alienation, introversion, low self-esteem, and high communication apprehension tend to be unwilling to communicate. In order to investigate the personality antecedents of unwillingness to communicate, Burgoon (1976) developed the Unwillingness to Communicate Scale (UCS), which contains twenty Likert-scale items, and showed that unwillingness to communicate was significantly correlated with the trait-like factors of communication apprehension and the approach-avoidance. Another study by Burgoon (1977) investigated
the consequences of unwillingness to communicate and showed that participants with less unwillingness to communicate provided and sought more information in small group discussions, which indicated that unwillingness to communicate predicts communication behavior in L1.

The notion of unwillingness to communicate was rephrased into its positive term, willingness to communicate, in McCroskey and Baer’s study (1985). L1 WTC is defined as the probability that an individual will choose to communicate when free to do so (McCroskey & Baer, 1985, p. 7). WTC in L1 is also believed to be a trait-like factor, which is enduring and stable over times and across situations. As pointed out by McCroskey and Richmond (1987), a person’s degree of willingness to communicate may vary according to situations, for example a person may be more willing to communicate when free chatting with close friends, but less willing when talking to strangers. However, McCroskey and Baer (1985) persisted that to a major extent WTC in L1 still tends to be consistent across situations. They believed that a person who is less willing to communicate than other people in certain situations will be consistently less willing in other situations when using the first language.

McCroskey and Baer (1985) developed an L1 WTC scale, which contains four communication contexts: public speaking, talking in meetings, talking in small groups, and talking in dyads; and three types of receivers: strangers, acquaintances, and friends. The scale includes 20 items consisting of 12 scored and 8 filler items. The scale was proved to be valid in its content when the responses collected in the scale were highly correlated with each other, which means the scale is uni-dimensional (p. 21). McCroskey and Baer (1985) also confirmed the constructional validity of the scale in which the L1
WTC scale measures distinctive features other than communication apprehension and talking behavior (p. 22).

Using the L1 WTC scale, a number of studies (e.g., Barraclough, Christophel, & McCroskey, 1988; McCroskey & Richmond, 1990; MacIntyre, 1994) were conducted to investigate its underlying antecedents, that is, factors that would predict WTC. The variables investigated included self-esteem, extraversion, perceived communication competence, and communication apprehension. The results of some studies revealed that the two major factors affecting and predicting L1 WTC across cultures are communication apprehension and self-perceived competence (Barraclough, Christophel, & McCroskey, 1988; Burroughs, Marie, & McCroskey, 2003; Donovan & MacIntyre, 2005; MacIntyre, 1994; MacIntyre, Babin, & Clément, 1999). MacIntyre (1994) postulated the L1 WTC model shown in Figure 2.1 and concluded that L1 WTC consists of two most immediate antecedents, L1 communication apprehension (CA) and L1 perceived competence (PC) of communication. People who are less anxious in communication and consider themselves capable in communication skills are more willing to communicate in their L1, and vice versa. At the same time, L1 CA is caused by a combination of introversion and low self-esteem, and L1 PC is caused by a combination of L1 CA and introversion (p. 139). Therefore, MacIntyre (1994) further concluded that L1 PC is the strongest indicator of L1 WTC and a change in L1 CA will cause changes in L1 WTC directly and indirectly through L1 PC (p. 139).
As shown above, L1 CA and L1 PC are proposed as predictors (i.e., antecedents) of L1 WTC. In turn, the question regarding what L1 WTC would predict was raised by researchers. Following the belief that L1 unwillingness to communicate predicts L1 communication behavior, McCroskey and Baer (1985) posited that a person’s L1 communication behavior is also rooted in the predisposition of L1 willingness to communicate. In order to verify the assumption, Zakahi and McCroskey (1989) investigated participants of high and low L1 WTC in their behavior of agreeing to participate in research involving communication and actually appearing in the research. They found that there was a significant difference in the behavior between people with low and high L1 WTC and that L1 WTC was positively correlated with the behavior. Chan and McCroskey (1987) also observed three classrooms of the students’ actual L1 behavior of communication in class. They discovered that students with high L1 WTC participated more in class. As a result, Chan and McCroskey (1987) drew the conclusion that L1 WTC can predict the actual behavior of communication. MacIntyre, Babin, and Clément (1999) did a t-test study which revealed that the group who attended the
laboratory reported significantly higher WTC than the group who did not. They also found out that for difficult speaking tasks, L1 WTC was the only predictor for the participants who had attempted to do the task when given free choices. The above studies proved the predictive validity of the L1 WTC scale created by McCroskey and Baer (1985) (McCroskey, 1992).

In summary, L1 WTC is, to a major extent, a trait-like factor, which is directly influenced by communication anxiety and self-perceived competence in communication skills. People with lower communication anxiety and higher self-perceived competence would be more willing to communicate with others when given a free choice and across situations and interlocutors. L1 WTC also plays a crucial role in predicting people’s behavior in communication. Those with higher L1 WTC would actually talk more frequently than those with lower L1 WTC.

2.1.2 Willingness to Communicate in L2

MacIntyre and his associates (1998) defined L2 WTC as “a readiness to enter into discourse at a particular time with a specific person or persons, using a L2” (p. 547). The very definition reflects the major difference between L1 and L2 WTC: L1 WTC is a personality construct, whereas L2 WTC is considered to be more directly influenced by situational and state factors of particular time and specific persons. One fact that may lead to the difference is the wider range of possibilities in the antecedents of L2 WTC. For example, the perceived competence of a second language can vary from 0% to 100% and an L1 perceived competence would probably never reach 0% (MacIntyre et al., 1998, p. 546). To further elaborate the distinction, MacIntyre (2007) clarified that affective
factors consist of three levels: trait, situational, and state levels (p. 565). The trait level refers to a long-term, stable, and typical pattern of a person’s predisposition; at the situational level, the focus is on a specific pattern defined under a certain situation; and at the state level, it is mainly related to a pattern at a specific moment (p. 565).

MacIntyre and his associates (1998) built a heuristic model of L2 WTC consisting of trait, situational, and state levels of antecedents and its outcome. The model is shown in Figure 2.2 The pyramid model includes six layers: Layer VI to III are the antecedents, Layer II is the L2 WTC, and Layer I is the outcome of L2 WTC. The lowest three layers of the model contain enduring factors, which represent stable and long-term, but distant and indirect influences on L2 WTC. For example, intergroup climate at Layer VI indicates the socioeconomic power of a language community. A language with higher socioeconomic power contains higher ethnolinguistic vitality and would attract more speakers, which indirectly increase the WTC of the language (p. 555). The factor of intergroup climate exists in the social economic context and is stable and consistent over a long period. The third layer consists of situational and state factors, which may change at a given moment or occasion, but closely and directly affect L2 WTC (p. 547). As the immediate and direct influences of L2 WTC, the third layer consists of desire to communicate with a specific person and state communicative self-confidence. State communicative self-confidence further includes two constructs: (a) state perceived competence, and (b) state anxiety (p. 549). These two state constructs are different from the L2 self-confidence in Layer IV in that the former ones are changing over time and across situations, while the latter is relatively stable. According to Eagly and Chaiken
intention is a strong predictor of actual behavior. Willingness to communicate in L2, therefore, is presupposed to be able to predict the L2 use in Layer I.

*Figure 2.2* MacIntyre’s heuristic model of variables influencing WTC (reproduced from MacIntyre et al., 1998, p. 547)

MacIntyre, Baker, Clément, and Conrod (2001) adopted a notion of communication with a broader connotation which includes the productive ends in communication, speaking and writing, and its receptive ends, reading and comprehension. They hypothesized that the L2 WTC of the four skills would not be correlated with each other due to the situational features of factors affecting L2 WTC (p. 372). MacIntyre et al. (2001) also made the distinction of L2 WTC inside and outside classrooms. Different from L1, a large part of L2 learning takes place inside a language classroom. The WTC of L2 is very likely different inside and outside a classroom. Therefore, the inclusion of these features in L2 WTC reflects the characteristics of L2 learning. Based on the above rationale, MacIntyre et al. (2001) proposed the L2 WTC scale with two parts: inside and
outside the classroom. Each part contains 27 items referring to L2 WTC in four skill areas: speaking, comprehension, reading, and writing. Students are asked to rate the scale from 1, almost never willing, to 5, almost always willing, showing how willing they would be to communicate in a second language in proposed situations. Examples in the L2 WTC scale inside a classroom contain likely classroom activities such as speaking in a group about your summer vacation, reading reviews for popular movies, writing a newspaper article, listening to instructions and completing a task, and so on. The L2 WTC outside classroom includes the same statements, but the scenarios are moved to the real world outside of language classrooms. MacIntyre and his associates (2001) carried out a study to test the L2 WTC scale, and the reliability of the scale was confirmed (p. 375).

In summary, L2 WTC is a situational construct and receives direct and close influences from situational and state factors. The heuristic model of L2 WTC by MacIntyre and his associates (1998) shows L2 state communicative apprehension and perceived competence are the two direct antecedents of L2 WTC, and L2 WTC predicts the L2 communicative behavior. The L2 WTC model by MacIntyre et al. (2001) incorporates the four language skills and scenarios inside and outside language classrooms.

2.2 Perceived Competence

Like most affective factors in L2 learning, communication competence also stems from the field of first language communication. Various definitions have been proposed by researchers in the field during the 1970s and 1980s. Larson, Backlund, Redmond, and Barbour (1978) defined communication competence as the ability to demonstrate the
knowledge of certain communication behavior in certain situations. Allen and Brown (1976) even more explicitly stated that competence “is tied to actual performance of the language in social situations” (p. 248, cited in McCroskey, 1982). Both definitions more or less equate competence with performance. However, McCroskey (1982) pointed out that competence and performance are not the same thing. He claimed that knowing does not equal doing. As an example put forth by McCroskey (1982), a professor in the field of communication possesses abundant knowledge of communication but may be a poor communicator who can barely teach a class effectively. McCroskey (1984) defined communication competence as “adequate ability to make ideas known to others by talking or writing” (p. 263). In order to operationalize the concept of communication competence, McCroskey and McCroskey (1988) synthesized four possible ways to measure communication competence through performance: “(a) objective observation, (b) subjective observation, (c) self-report, and (d) receiver-report” (p. 109). The first two measures intend to measure a speaker’s actual performance by observation. The last measure, receiver-report, uses listeners’ or readers’ perception to judge the ability of the speaker. The measure of self-report, on the other hand, indicates the speaker’s self-assessment of his or her competence. The self-assessment is affected by other affective factors, such as communication anxiety and introversion (MacIntyre, 1994), and therefore is usually different from the actual competence. McCroskey and McCroskey (1988) stated that many decisions made by people in communication are actually decided by their self-perceived competence, instead of their actual one (p. 110). As a result, people decide whether or not to communicate based on whether they think they have the competence to do so or not, instead of whether they actually have the competence.
Mc Croskey and McCroskey (1988) created a scale of L1 Self-Perceived Communication to allow speakers to rate on the probability of self-assessed communication competence from 0, completely incompetent, to 100, competent. The scale includes 12 statements describing four basic communication contexts—public speaking, talking in a large meeting, talking in a small group, and talking in a dyad, as well as three common types of receivers—strangers, acquaintances, and friends. The response format of 0-100 probability, explained by McCroskey (1992), is adopted to be easily understood by lay people, as the same format is usually used in many instructional systems and weather patterns in news reports (p. 21). MacIntyre and Charos (1996) rephrased the 12 items in the L1 Self-Perceived Communication scale to highlight L2 features in the scale. For example, they changed the “Present a talk to a group of acquaintances” to “Give a presentation in English (or any other L2) to a group of acquaintances”.

In summary, the conceptual construct of L2 WTC is affected by L2 self-perceived competence. McCroskey and McCroskey (1988) created the Self-Perceived Communication scale to measure L1 perceived competence, and MacIntyre and Charos (1996) adopted the scale and rephrased it into a scale highlighting L2 features.

2.3 Communication Anxiety

L2 communication anxiety has been claimed by researchers to be one of the major factors that affect learners’ second language learning (e.g., Chen & Chang, 2004; Gardner, 1985; Gardner & MacIntyre, 1993; Zhang & Wang, 2002). L2 communication anxiety originated from the construct of communication apprehension in the field of
communication, which was defined as “an individual's level of fear or anxiety associated with either real or anticipated communication with another person or persons” (McCroskey, 1977, p. 78). People with high L1 communication apprehension levels often avoid or withdraw from conversations (Barraclough, Christophel, & McCroskey, 1988). McCroskey and Richmond (1987) categorized four types of communication apprehension: (a) trait-like, (b) context-based, (c) receiver-based, and (d) situational (p. 142). The first three types of communication apprehension are relatively enduring orientations toward communication. Trait-like communication apprehension is consistent across a wide variety of contexts. Context-based apprehension is oriented more towards a given type of context, whereas receiver-based apprehension is geared towards a given type of person or persons. Situational communication apprehension is considered as “a transitory orientation toward communication with a given person or group of people” (McCroskey & Richmond, 1987, p. 144).

McCroskey (1997) proposed a continuum view of communication apprehension. He suggested viewing communication apprehension as ranging from the extreme trait pole to the extreme state pole (p. 84). In contrast with the trait-like property of communication apprehension in L1, communication apprehension in L2 is more at the situational and state end of the pole. When it is defined in a classroom of second language learning, communication anxiety refers to “a distinct complex of self-perceptions, beliefs, feelings, and behaviors related to classroom language learning arising from the uniqueness of the language-learning process” (Horwitz, Horwitz, & Cope, 1986, p. 128). Facilitating and debilitating anxieties are the two types of language anxiety (Scovel, 1978). The former encourages learners to take the challenge of a task,
and the latter sets obstacles in learning a second language. Nevertheless, MacIntyre and Gardner (1994) pointed out that facilitating anxiety was less influential and the extra effort brought by it was not necessary rewarded with an increased achievement. Therefore, only the issue of debilitating anxiety is addressed in L2 WTC research, and anxiety in the dissertation always refers to debilitating anxiety.

Generally speaking, language anxiety is negatively correlated with learners’ language performance. Tobias (1980) proposed a model of three stages where second language anxiety took effect: input, processing, and output. Language anxiety acts like a filter at the input stage. It blocks and reduces the amount of input information going into learners’ cognitive processing system. Therefore, for a debilitatingly anxious learner, language anxiety will cause delay or less information going into the processing stage from listening or reading. At the processing stage, “anxiety can influence both the speed and accuracy of learning” (MacIntyre, 1999, p. 35). Language anxiety hinders the information which has entered the processing stage from being processed. Therefore, even less is learned by anxious learners at the processing stage. At the output stage learners need to produce output based on what they have learned. However, language anxiety again influences the quality of their performance. They may stumble and give broken sentences or produce very short sentences in the performance of productive skills. Language anxiety causes discrepancy in learners’ actual competence and performance. Therefore, anxious learners may be underestimated on their language competence. In a second language learning context, communication anxiety and learners’ perceived competence may interact and affect learners’ willingness to communicate (e.g., MacIntyre & Charos, 1996; Peng 2007; Yashima, 2002).
A universal characteristic of communication anxiety is the internal feeling of discomfort when affected by communication anxiety (McCroskey, 1984). Therefore, a self-report scale of communication anxiety may well reflect the internal feeling of a language learner. There are three major scales of communication anxiety: the Personal Report of Communication Apprehension (PRCA) by McCroskey (1982), the Communication Anxiety scale by MacIntyre and Charos (1996), and the Foreign Language Classroom Anxiety Scale (FLCAS) by Horwitz et al. (1986).

McCroskey’s PRCA is a five-point Likert scale which consists of 24 items. The participants are asked to rate from 1, strongly disagree, to 5, strongly agree on the contexts of public speaking, dyadic interaction, small groups, and large groups. Examples of the PRCA include “Generally, I am nervous when I have to participate in a meeting”, “I have no fear of giving a speech”, and so on. The total possible score ranges from a minimum of 24 to a maximum of 120. According to McCroskey’s criteria, people with scores below 51 have very low CA, and those with scores between 51 and 80 usually have average CA. Scores above 80 represent people who have high levels of trait CA.

The Communication Anxiety scale by MacIntyre and Charos (1996) was developed on the basis of McCroskey’s Self-Perceived Communication scale (McCroskey & McCroskey, 1988). The participants are required to rate on the scale of 0-100 to indicate the degree of anxiety they may feel under the situations stated in 12 items. The 12 items include 4 communication contexts: public speaking, talking in a large meeting, talking in a small group, and talking in a dyad; and 3 types of receivers: strangers, acquaintances, and friends. Examples of the scale include “talk in the L2 to
acquaintances”, “give a presentation in L2 to a group of friends”, “have a small-group conversation in L2 with strangers”, and so on.

The FLCAS by Horwitz and his associates (1986) was designed to measure language anxiety inside a language classroom specifically. It comprises 31 items following a five-point Likert scale format, ranging from “strongly disagree” to “strongly agree”. The participants are requested to mark the degree of agreement of the statements listed in the 31 items, such as “I don’t worry about making mistakes in the foreign language class”, “I keep thinking that others are better at the language than I am”, and so on. The FLCAS has been commonly used in the studies of L2 language anxiety, but it only tests the anxiety inside a second language classroom. It is not widely employed in L2 WTC studies.

In summary, higher language anxiety is thought to lead to poorer performance in L2 use and lower self-perceived competence. The two factors work together and may reduce a learner’s willingness to communicate in the L2. The three major scales used in the field of second language learning to measure anxiety are the scales by McCroskey (1982), MacIntyre and Charos (1996), and the FLACS by Horwitz and his associates (1986).

2.4 Synthesis Study of L2 WTC

The importance of communicating in a second language has been recognized by researchers in the field of second language pedagogy. Researchers and teachers attempt to teach a language in a communicative way to help learners to produce meaningful output in a target language. However, previous studies on L2 communication largely emphasized the effects of this teaching approach on groups of students and ignored the
role of differences existing in learners as individual units (Dörnyei & Skehan, 2003). The concept of willingness to communicate was introduced into the field of second language acquisition by MacIntyre and Charos (1996). Since then, researchers have been conducting studies to investigate the role of WTC on second language communication behavior, as well as the antecedents of WTC in a second language. The overall conclusion claims that WTC directly predicts learners’ frequency of use of a second language (Clément et al., 2003) and in turn, learners’ perceived competence and communication anxiety predict learners’ WTC in L2. In other words, frequency of use of a second language is an outcome of WTC and perceived competence and communication anxiety are antecedents of WTC. However, the studies on L2 WTC have different foci. Some focused on the relation between affective factors and WTC (e.g., MacIntyre & Charos, 1996; Peng 2007; Yashima, 2002) and/or the relation between WTC and L2 communication behavior (e.g., Cao & Philip, 2006; Yashima et al., 2004), some investigated the differences of WTC in L1 and L2 (e.g., Baker & MacIntyre, 2000; MacIntyre, Baker, Clément, & Donovan, 2002), and still others examined the role of gender or age in L2 WTC (e.g., Baker & MacIntyre, 2000; Weaver, 2005).

In order to systematically review the primary studies, a systematic research synthesis was conducted on the L2 WTC (for an overview of this research methodology, see Norris & Ortega, 2006; Ortega, 2010). This section will first present the purpose of the synthesis of the WTC construct, then review and analyze the characteristics of the participants and research designs across L2 WTC studies, and finally compare the findings across these accumulated studies. The goal is to provide a better understanding and generalization of the research results in the domain, identify gaps among the primary
studies, and provide designing justification for the current dissertation study, as well as suggestions for future researchers.

2.4.1 Problem Specification with the Primary Studies on L2 WTC

The studies on L2 WTC vary on the problems investigated in the research. Most focused on the relation between affective factors and L2 WTC, but with a great variety of other foci. The problems investigated in the primary studies can be categorized into the five main types:

1. To examine the relation between L2 WTC and its antecedents;
2. To examine the relation between self-reported WTC and self-reported or observed communication behavior in L2;
3. To examine the differences of WTC in L2 low and high use context;
4. To examine the effects of sex, age, and proficiency levels on L2 WTC;
5. To examine the differences of WTC in L1 and L2.

The main affective factors related to L2 WTC examined in the primary studies are perceived competence, communication anxiety, introversion, integrativeness, motivational intensity, contact with the L2 group, attitude, frequency of communication, and international posture. Most studies agreed that perceived competence was the strongest predictor of learners’ L2 WTC (e.g., Clément et al., 2003; MacIntyre et al., 2002; Yashima et al., 2004). Some studies claimed that communication anxiety also predicted L2 WTC (e.g., Baker & MacIntyre, 2000; Liu & Jackson, 2008; Yashima, 2002). Besides the above two affective factors, other factors, such as L2 contact (Liu & Jackson, 2008), introversion (Matsuoka, 2006), international posture (Yashima, 2002;
Yashima et al., 2004), and attitude toward the target language (Tannenbaum & Tahar, 2008) were also reported to be strongly correlated with L2 WTC. Researchers (e.g., Yashima et al., 2004) have also investigated the relation of L2 WTC and frequency of L2 use, concluding that high L2 WTC results in high self-reported frequency of L2 communication.

Baker and MacIntyre (2000), MacIntyre et al. (2003), and Yashima et al. (2004) investigated differences of WTC in different L2 use contexts. MacIntyre and Baker (Baker & MacIntyre, 2000; MacIntyre et al., 2003) recruited participants from an immersion and a non-immersion program. The results of both studies indicated in the immersion programs, where L2 use is relatively higher, communication anxiety was strongly correlated with L2 WTC, but perceived competence was less so. On the other hand, perceived competence of participants in the non-immersion group was more strongly correlated with L2 WTC than communication anxiety. Yashima et al. (2004) conducted two investigations, one with Japanese participants learning English in Japan, that is, in a low L2 use context, and the other with Japanese participants in an English program in the United States, where L2 English is highly used. The study showed in both contexts, perceived competence strongly correlated to L2 WTC.

Other researchers have pointed out that females tended to have a higher L2 WTC than males (Peng, 2007) and that age might cause significant differences in WTC at a certain grade, for example from grade seven to eight (MacIntyre et al., 2002), but that higher proficiency is not a significantly correlated antecedent to L2 competence and also not significantly correlated with L2 WTC directly (D’Amico, 2010; Yashima, 2002; Yashima et al., 2004).
There have also been several qualitative studies (Cao 2009; Freiermuth & Jarrell, 2006; Kang, 2005; MacIntyre & Legatto, 2010) which have focused more on situational factors in L2 WTC. Kang (2005) investigated the relationship between L2 WTC and psychological antecedents, such as security, excitement, and responsibility, and the influence of situational factors, such as topics, interlocutors, and conversational contexts. Freiermuth and Jarrell (2006) compared L2 WTC in online chat versus face to face discussions, and concluded that online chat stimulated more willingness to communicate in L2. MacIntyre and Legatto (2010) developed the idiodynamic method, which allowed them to record the fluctuation of WTC from moment-to-moment of six French language learners. In their discussion of reasons for the WTC fluctuation across their participants, MacIntyre and Legatto concluded that language anxiety and L2 WTC do not always appear negatively correlated (p. 164). They also noticed that the ease of flow of vocabulary into mind and familiarity of the topics were postively correlated with learners’ L2 WTC (p. 165). Cao (2009) used an ecological frame to investigate the situated nature of the L2 WTC and concluded that topic, interlocutor, teacher, opportunity for talking, and self-competence were all factors affecting L2 WTC (p. 207).

The primary studies also carry different features, such as participants’ backgrounds, venues of studies, the target languages, and L2 use contexts. The instruments employed in the studies are far from consistent as well. Some used the measurement scales developed by McCroskey and his associates (1985, 1987, 1990), some used the scales by MacIntyre et al. (2001), while others developed their own scales to measure WTC and related factors (Weaver, 2005; Fushino, 2010). The varied foci,
features, and scales of the primary studies make it difficult for readers to generalize the results from this accumulated research.

2.4.2 Purpose of Present Research Synthesis

The inconsistent results and varied foci and features of the primary studies failed to provide readers with a standpoint to generalize the research results. A synthesis study which starts from investigating and comparing the features helps to better understand and generalize the results of the primary studies (Norris & Ortega, 2006; Ortega, 2010). Since the concept of willingness to communicate was introduced into the field of second language acquisition in the 1990s, no synthesis studies have been conducted on the topic. Therefore, a systematic synthesis study will be employed here to examine the features of L2 WTC studies, antecedents of L2 WTC, and evidence about the role of WTC in second language learning, as well as to identify the gaps in WTC studies investigated in the SLA literature to date. This synthesis attempts to address the following research questions:

1. How has L2 WTC been investigated across studies?
2. What is the relationship between L2 WTC and the main posited outcome, that is, frequency of second language use?
3. What roles do the two main posited antecedents of WTC, perceived competence and communication anxiety, play on WTC in a second language?

2.4.3 Method

2.4.3.1 Study identification and retrieval. An extensive and inclusive search was performed to locate studies for the present synthesis. In order to retrieve all the
possible studies on WTC in second language learning, I first did an extensive online
database search, then a reference tracing via footnote chasing.

I conducted an extensive online search in the following four databases through
Cambridge Scientific Abstracts (CSA): Educational Resources Information Center
(ERIC), Linguistics and Language Behavior Abstracts (LLBA), MLA International
Bibliography, and PsycINFO. To increase the ability for researchers to replicate the
synthesis study, I limited the search to “English only”. The first key term used to retrieve
data was “KW= (Willingness to Communicate) and ((Language learning) or (Language
teaching))”. Nineteen relevant studies were identified. As the concept of WTC originated
in the concept of unwillingness to communicate, I tried the the key terms
“KW=(unwillingness to communicate)” to locate possible studies. One additional study
was located. All together, there were 20 relevant studies retrieved through the online key
terms search (Table 2.1).

Table 2.1
| Online studies retrieval through Cambridge Scientific Abstracts (CSA) |
|-----------------------------|-----------------------------|
| Data base | Key terms | Studies |
| PsycINFO | KW=(Willingness to Communicate) and ((Language learning) or (Language teaching)) | 8 |
| MLA | KW=(Willingness to Communicate) and ((Language learning) or (Language teaching)) | 1 |
| CSALLBA | KW=(Willingness to Communicate) and ((Language learning) or (Language teaching)) | 7 |
| ERIC | KW=(Willingness to Communicate) and ((Language learning) or (Language teaching)) | 3 |
| PsycINFO | KW=(unwillingness to communicate) | 0 |
| CSALLBA | KW=(unwillingness to communicate) | 0 |
| ERIC | KW=(unwillingness to communicate) | 1 |
In order to minimize publication bias and assure that I had retrieved all the
relevant doctoral dissertations, I explored the database of Dissertations and Theses
through ProQuest by the following key term: “ABS(Willingness to Communicate) AND
ABS(Language)”. Five more studies were located. I then searched the online homepage
of McCroskey and MacIntyre, and one more study was identified from McCroskey’s
homepage. Finally, I conducted a reference search of the studies cited by six most recent
studies on L2 WTC: Fushino (2010), MacIntyre and Doucette (2010), Peng and
Woodrow (2010), Xie (2011), Yu (2010), and Zeng (2010). One more primary study was
retrieved during this process. As a result of this process, a total of 27 studies were
retrieved as eligible for the current research synthesis. All the studies included are
marked with an asterisk in the references.

2.4.3.2 The inclusion and exclusion criteria. The studies in the present synthesis
were selected based on the following inclusion and exclusion criteria:

1. They were (quasi-)experimental in design. Qualitative Studies were excluded,
   but will be discussed briefly on their methodological features (Cao, 2009;
   Freiermuth & Jarrell, 2006; Kang, 2005; MacIntyre & Legatto, 2010). These
   qualitative studies are marked with two asterisks in the references.

2. They involved participants who were learning the languages as second
   languages. Studies which involved first language learning were excluded (e.g.,
   Donovan & MacIntyre, 2005; MacIntyre, Babin, & Clément, 1999).

3. They collected data on WTC either through questionnaires or interviews.
   Studies which, though discussing WTC, did not include WTC as a part of data
   collection were excluded (e.g., Léger & Storch, 2009). Studies which claimed
to examine L2 WTC but measured L2 communication behavior instead were also excluded (e.g., Kissau, McCullough, & Pyke, 2010).

4. They were conducted in formal school settings or adults second language courses. Studies conducted in working training courses were excluded (e.g., Derwing, Munro, & Thomson, 2008; Ekkens & Winke, 2009).

5. They were published articles in journals or books, and unpublished doctoral dissertations. Working papers, unpublished manuscripts, and Master’s theses (e.g., Fricka, 1997; Hashimoto, 2002; Lu, 2007; MacDonald, Clément, & MacIntyre, 2003) were excluded to avoid the potential bias caused by their partial availability, as well as to ensure the quality of the included studies and make it easier for readers to replicate the synthesis study.

6. They were written in English and published after 1996, since the first primary study on L2 WTC was conducted in 1996.

7. If the same sample was reported in several studies, only one sample was included in the current synthesis study. For example, when a dissertation and a published article were based on the same sample, the published article was included in the current synthesis study and the dissertation was only used to retrieve necessary data for the synthesis study (e.g., Fushino, 2008, 2010).

2.4.3.3 Coding of the primary studies. A coding book was developed in order to collect sufficient information on the primary studies and to compare the characteristics of the studies and find underlying relationships between the characteristics and their conclusions. Four main categories, which are believed to relate to WTC, were coded: (a) participants’ background, (b) research design, (c) instruments, and (d) research purposes.
The coding of participants’ background includes (a) the n size of participants, (b) school status of participants, (c) proficiency levels, and (d) proficiency assessment types. The relevant sections in the studies were read carefully to locate the necessary information. The school status were categorized based on whether the participants were in universities, secondary schools, or primary schools. The categories proposed by Thomas (2006) were used in coding the variables related to proficiency: (a) impressionistic judgment, (b) institutional status, (c) in-house assessment, and (d) standardized test (pp. 282-283). Few studies explicitly stated the proficiency levels of participants. I looked for words such as “beginning”, “intermediate”, or “advanced” in the relevant sections. If such impressionistic descriptions and words were not found in the study, I looked for the description of their institutional status. I coded participants as “beginning learners” if they were described as “enrolled in a first or second semester language course”, and “intermediate learners” if they were at the second year level of a language course. They were coded as “advanced learners” if they were enrolled in the third year or higher of the language course. Correspondingly, if learners’ proficiency levels were defined based on the language courses they had enrolled in, the assessment type was coded as “institutional status”.

The second coding category “research design” includes (a) the first languages, (b) the target language, (c) the experiment venue, and (d) L2 use context. The first two coding items aimed to find out whether there were any preferences on investigating certain first and target languages. The third coding item, experiment venues, refers to the location where the research was carried out to examine whether previous studies were conducted in a foreign or second language learning environment. In order to code the
fourth item, L2 use context, I looked for words such as “unilingual community” and “immersion program”. I coded “immersion program” and “bilingual community” as high L2 use context, and “nonimmersion program” and “unilingual community” as low L2 use context. If no related descriptions were identified, I conducted the coding based on whether the venue of the experiment was on a foreign (low L2 use) or second (high L2 use) language learning context. For example, if the study was carried out in Japan and investigated English learning, I coded the context as low L2 use.

Instruments were also coded to examine which instruments the researchers had used to measure L2 WTC, L2 perceived competence, L2 communication anxiety, and frequency of L2 use, if applicable. Finally, the coding of research purposes was essential to find out what research problems are most widely investigated and what problems need more examination (See the complete coding book at Appendix A).

2.4.4 Methodological Findings

2.4.4.1 Participants’ characteristics. Table 2.2 shows the participants’ characteristics of each of the 27 unique samples involved in the 27 primary studies. A total of 6600 language learners participated in the studies. The maximum n size of the studies is 772 and the minimum is 8. The mean n size of the 27 studies is 244.44 with a standard deviation of 180.34. Eighteen (66.67%) out of 27 studies were carried out in a university setting, seven (25.93%) were conducted in secondary schools, one (3.7%) in a primary school, and one (3.7%) in an adult evening school.

Only sixteen studies (59.26%) provided information about participants’ proficiency levels. Three studies (11.11%), by MacIntyre and his colleagues (1996, 2003)
and Xie (2011), involved beginning learners as participants. Three studies (11.11\%) involved participants at the intermediate level. Eight studies (29.63\%) claimed that they had participants of mixed proficiency levels, but only two (Yashima, 2002; Yashima et al., 2004) of them investigated the differences of L2 WTC at different proficiency levels. The other two studies (7.4\%) reported that their participants were “above average” and “at higher average”, but failed to clarify the corresponding proficiency levels of these descriptions.

Among the sixteen studies which provided description of participants’ proficiency levels, institutional status was the most frequently used assessment ($n = 7$, 25.93\%), followed by in-house assessment ($n = 4$, 14.81\%), standardized tests ($n = 4$, 14.81\%), and self-assessment ($n = 1$, 3.7\%).
<table>
<thead>
<tr>
<th>Studies</th>
<th>N</th>
<th>School status</th>
<th>Proficiency level</th>
<th>Proficiency assessment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baker &amp; MacIntyre (2000)</td>
<td>174</td>
<td>secondary</td>
<td>missing</td>
<td>missing</td>
</tr>
<tr>
<td>Burroughs et al. (2003)</td>
<td>131</td>
<td>university</td>
<td>missing</td>
<td>missing</td>
</tr>
<tr>
<td>Cao &amp; Philip (2006)</td>
<td>8</td>
<td>university</td>
<td>intermediate</td>
<td>institutional</td>
</tr>
<tr>
<td>Cetinkaya (2006)</td>
<td>304</td>
<td>university</td>
<td>mixed</td>
<td>in-house</td>
</tr>
<tr>
<td>Chu (2008)</td>
<td>364</td>
<td>university</td>
<td>mixed</td>
<td>in-house</td>
</tr>
<tr>
<td>Clément et al. (2003)</td>
<td>378</td>
<td>university</td>
<td>missing</td>
<td>missing</td>
</tr>
<tr>
<td>Fushino (2010)</td>
<td>772</td>
<td>university</td>
<td>higher range</td>
<td>in-house</td>
</tr>
<tr>
<td>Hsu (2006)</td>
<td>384</td>
<td>university</td>
<td>missing</td>
<td>missing</td>
</tr>
<tr>
<td>Kim (2005)</td>
<td>191</td>
<td>university</td>
<td>missing</td>
<td>missing</td>
</tr>
<tr>
<td>Liu &amp; Jackson (2008)</td>
<td>547</td>
<td>secondary</td>
<td>missing</td>
<td>missing</td>
</tr>
<tr>
<td>MacIntyre (1996)</td>
<td></td>
<td>adult evening</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MacIntyre &amp; Charos (2010)</td>
<td>92</td>
<td>adult evening</td>
<td>missing</td>
<td>missing</td>
</tr>
<tr>
<td>MacIntyre &amp; Doucette (2010)</td>
<td>238</td>
<td>secondary</td>
<td>intermediate</td>
<td>institutional</td>
</tr>
<tr>
<td>MacIntyre et al. (2001)</td>
<td>79</td>
<td>secondary</td>
<td>missing</td>
<td>missing</td>
</tr>
<tr>
<td>MacIntyre et al. (2002)</td>
<td>268</td>
<td>secondary</td>
<td>missing</td>
<td>missing</td>
</tr>
<tr>
<td>MacIntyre et al. (2003)</td>
<td>59</td>
<td>university</td>
<td>beginning</td>
<td>institutional</td>
</tr>
<tr>
<td>Matsuoka (2006)</td>
<td>180</td>
<td>university</td>
<td>average</td>
<td>standardized</td>
</tr>
<tr>
<td>Peng (2007)</td>
<td>118</td>
<td>university</td>
<td>mixed</td>
<td>institutional</td>
</tr>
<tr>
<td>Peng &amp; Woodrow (2010)</td>
<td>503</td>
<td>university</td>
<td>missing</td>
<td>missing</td>
</tr>
<tr>
<td>Tannenbaum &amp; Tahar (2008)</td>
<td>143</td>
<td>primary</td>
<td>missing</td>
<td>missing</td>
</tr>
<tr>
<td>Weaver (2005)</td>
<td>490</td>
<td>university</td>
<td>missing</td>
<td>missing</td>
</tr>
<tr>
<td>Xie (2011)</td>
<td>197</td>
<td>university</td>
<td>beginning</td>
<td>institutional</td>
</tr>
<tr>
<td>Yashima (2002)</td>
<td>297</td>
<td>university</td>
<td>mixed</td>
<td>standardized</td>
</tr>
<tr>
<td>Yashima et al. (2004)</td>
<td>214</td>
<td>secondary</td>
<td>mixed</td>
<td>standardized</td>
</tr>
<tr>
<td>Yashima &amp; Zenuk-Nishide (2008)</td>
<td>165</td>
<td>secondary</td>
<td>mixed</td>
<td>standardized</td>
</tr>
<tr>
<td>Zeng (2010)</td>
<td>46</td>
<td>university</td>
<td>mixed</td>
<td>in-house</td>
</tr>
</tbody>
</table>
2.4.4.2 Research design. Table 2.3 summarizes the findings yielded by the coding of research design variables. The first languages the participants spoke in the 27 study samples included English \((n = 8 \text{ or } 29.63\%)\), Mandarin \((n = 7 \text{ or } 25.93\%)\), Japanese \((n = 5 \text{ or } 18.52\%)\), Korean \((n = 1 \text{ or } 3.7\%)\), and Turkish \((n = 1 \text{ or } 3.7\%)\). Five studies \((18.52\%)\) employed participants speaking various languages, including one study (Clément et al., 2003) where the participants spoke English and French and one (Tannenbaum & Tahar 2008) where the participants spoke Hebrew and Arabic. Most studies \((n = 8 \text{ or } 29.63\%)\) were carried out in Canada with the target language of French \((n = 6 \text{ or } 22.22\%)\), English \((n = 1 \text{ or } 3.7\%)\), and both French and English \((n = 1 \text{ or } 3.7\%)\). Six studies \((22.22\%)\) were conducted in Japan, investigating English as a foreign language. Six studies \((22.22\%)\) examined WTC in English learning in China. The remaining seven studies were conducted respectively in North America \((n = 1 \text{ or } 3.7\%)\), Korea \((n = 1 \text{ or } 3.7\%)\), New Zealand \((n = 1 \text{ or } 3.7\%)\), Israel \((n = 1 \text{ or } 3.7\%)\), Micronesia \((n = 1 \text{ or } 3.7\%)\), the United States \((n = 1 \text{ or } 3.7\%)\), and Turkey \((n = 1 \text{ or } 3.7\%)\). English was the most commonly investigated target language. Seventeen studies \((62.96\%)\) investigated English learning as a second or foreign language, six \((22.22\%)\) investigated French as a second or foreign language, and one study \((3.7\%)\) investigated both English and French language learning. Only three studies \((11.11\%)\) investigated languages other than English or French: Hebrew and Arabic in Tannenbaum and Tahar (2008), Spanish in D’Amico (2010), and Mandarin in Xie (2011).

Different venues of the primary studies lead to different levels of L2 use contexts. Most studies \((n = 17 \text{ or } 62.96\%)\) chose a context of low L2 use, where participants had few opportunities to use the target language outside school. By contrast, four studies
(14.81%) were conducted in a context of high L2 use: New Zealand and bilingual communities in Canada. Another six studies (22.22%) included both contexts of low and high L2 use.

Table 2.3
Features of research design in the primary studies

<table>
<thead>
<tr>
<th>Studies</th>
<th>L1</th>
<th>L2</th>
<th>Venue</th>
<th>L2 use context</th>
</tr>
</thead>
<tbody>
<tr>
<td>Burroughs et al. (2003)</td>
<td>Mixed</td>
<td>English</td>
<td>Micronesia</td>
<td>L</td>
</tr>
<tr>
<td>Cetinkaya (2006)</td>
<td>Turkish</td>
<td>English</td>
<td>Izmir</td>
<td>L</td>
</tr>
<tr>
<td>Chu (2008)</td>
<td>Mandarin</td>
<td>English</td>
<td>China</td>
<td>L</td>
</tr>
<tr>
<td>Clément et al. (2003)</td>
<td>English /</td>
<td>French /</td>
<td>Canada</td>
<td>L &amp; H</td>
</tr>
<tr>
<td></td>
<td>English</td>
<td>English</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fushino (2008)</td>
<td>Mixed</td>
<td>English</td>
<td>Japan</td>
<td>L</td>
</tr>
<tr>
<td>Hsu (2006)</td>
<td>Mandarin</td>
<td>English</td>
<td>China</td>
<td>L</td>
</tr>
<tr>
<td>Kim (2005)</td>
<td>Korean</td>
<td>English</td>
<td>Korea</td>
<td>L</td>
</tr>
<tr>
<td>Liu &amp; Jackson (2008)</td>
<td>Mandarin</td>
<td>English</td>
<td>China</td>
<td>L</td>
</tr>
<tr>
<td>MacIntyre &amp; Charos (1996)</td>
<td>English</td>
<td>French</td>
<td>Canada</td>
<td>H</td>
</tr>
<tr>
<td>MacIntyre and Doucette (2010)</td>
<td>English</td>
<td>French</td>
<td>Canada</td>
<td>L</td>
</tr>
<tr>
<td>MacIntyre et al. (2001)</td>
<td>English</td>
<td>French</td>
<td>Canada</td>
<td>H</td>
</tr>
<tr>
<td>MacIntyre et al. (2002)</td>
<td>English</td>
<td>French</td>
<td>Canada</td>
<td>L</td>
</tr>
<tr>
<td>MacIntyre et al. (2003)</td>
<td>English</td>
<td>French</td>
<td>Canada</td>
<td>L &amp; H</td>
</tr>
<tr>
<td>Peng (2007)</td>
<td>Mandarin</td>
<td>English</td>
<td>China</td>
<td>L</td>
</tr>
<tr>
<td>Peng &amp; Woodrow (2010)</td>
<td>Mandarin</td>
<td>English</td>
<td>China</td>
<td>L</td>
</tr>
<tr>
<td>Tannenbaum &amp; Tahar (2008)</td>
<td>Hebrew</td>
<td>Arabic /</td>
<td>Israel</td>
<td>L &amp; H</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Hebrew</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Weaver (2005)</td>
<td>Japanese</td>
<td>English</td>
<td>Japan</td>
<td>L</td>
</tr>
<tr>
<td>Xie (2011)</td>
<td>English</td>
<td>Mandarin</td>
<td>United States</td>
<td>L</td>
</tr>
<tr>
<td>Yashima et al. (2004)</td>
<td>Japanese</td>
<td>English</td>
<td>Japan</td>
<td>L</td>
</tr>
<tr>
<td>Yu (2010)</td>
<td>Mandarin</td>
<td>English</td>
<td>China</td>
<td>L</td>
</tr>
<tr>
<td>Zeng (2010)</td>
<td>Mandarin</td>
<td>English</td>
<td>Canada</td>
<td>H</td>
</tr>
</tbody>
</table>

*Note. L = low, H = high.*
The qualitative studies by Kang (2005) and Freiemuth and Jarrell (2006) both targeted at English as an L2, with Korean as the first language in the former study, and Japanese as the first language in the latter study. Kang (2005) carried out the study in the United States, where English is highly used, and Freiemuth and Jarrell (2006) did the study in Japan, where English is learned at a low L2 use context. MacIntyre and Legatto (2010) targeted English-speaking language learners of French in Canada in an immersion program, and thus contributed qualitative insights into WTC in high L2 use contexts. Cao (2009) chose to conduct the study in New Zealand and the participants with various first languages were learning English, which is highly used in New Zealand.

2.4.4.3 Instruments. Even though all of the 27 studies investigated WTC on L2, 14 studies (51.85%) used McCroskey’s L1 WTC scale and only four studies (14.81%) used MacIntyre’s L2 WTC scale, including one study that used both the MacIntyre’s and McCroskey’s WTC scales. Four studies (14.81%) used modified items based on McCroskey or MacIntyre’s studies. One study (3.7%) employed Burgoon’s (1976) unwillingness to communicate scale (USC). One study (3.7%) adopted a scale from Sick and Nagasaka (2000, cited in Matsuoka, 2006). Two (7.4%) wrote their own items to measure L2 WTC. One (3.7%) adopted the scale by Weaver (2005), which is one of the two primary studies employing a self-developed L2 WTC scale.

Fourteen studies (51.85%) investigated the relationship between L2 WTC and perceived communication competence. Among the 14 studies, seven used the perceived competence scale by McCroskey, which contains 12 items measuring speakers perceived communication competence in a variety of oral communication contexts and with a variety of types of receivers. The PC scale developed by MacIntyre and adopted by five
studies consists of similar 12 items as McCroskey’s scale, but with more L2 learning features. One study contained both the MacIntyre’s and McCroskey’s PC scale (MacIntyre and Doucette, 2010). One researcher (Fushino, 2010) wrote his own items to measure perceived communication competence.

Sixteen studies (59.26%) investigated the communication anxiety (CA). Eight among the 16 studies used McCroskey’s CA scale, five used MacIntyre’s scale, and three used scale from Horwitz et al. (1986), a scale developed to measure foreign language classroom anxiety. Only ten studies (37.04%) investigated frequency of L2 use of communication. Nine adopted self-reported frequency by using a questionnaire designed by each researcher specifically for each study. Only Cao and Philip (2006) observed the participants’ actual frequency of L2 use by observing classes.

The qualitative study conducted by MacIntyre and Legatto (2010) included McCroskey’s scales to measure L2 WTC and L2 Communication anxiety. At the same time, they also used the idiodynamic method to investigate the changes in WTC. The participants were asked to complete eight language tasks and were video recorded. They were then required to rate their moment-to-moment L2 WTC from -5 to 5 while watching the playback of the videos and discuss their affective factors from moment to moment. Their communication anxiety was also evaluated by a research assistant based on their discussions. The frequency of L2 use was measured through both self-reported questionnaire and observation of the tasks. The remaining three qualitative studies did not use any of the instruments employed in the experimental studies. Kang (2005) identified WTC and its antecedents from interviews and stimulated recalls. Freiemuth and Jarrell (2006) searched learners’ WTC and its related factors from a questionnaire assessing the
students’ experiences of working in the mode of online chat and face to face discussions. Both studies employed observed actual frequency of L2 use of communication from the conversational tasks. Cao (2009) obtained data of L2 WTC, L2 perceived competence, and L2 communication anxiety through participants’ interviews and journals. She also evaluated the participants’ frequency of L2 use by a self-reported questionnaire and observation.

2.4.5 Substantive Findings

2.4.5.1 The roles of perceived competence and communication anxiety on L2 WTC. L2 perceived competence and L2 communication anxiety are the two major predictors of L2 WTC. Eleven studies investigated the relation of both L2 perceived competence and L2 communication anxiety with L2 WTC, whereas two studies examined only the relation between L2 communication anxiety and L2 WTC. Three further studies combined L2 perceived competence and L2 communication anxiety together as a single variable of L2 confidence. Table 2.4 summarizes the obtained correlations between WTC and perceived competence and communication anxiety. Among the studies discussing the correlations in high L2 use context, Baker and MacIntyre (2000) and MacIntyre et al. (2003) indicated that communication anxiety was more correlated with WTC than perceived competence, MacIntyre and Charos (1996) reported the correlation was the same for the two antecedents, and Yashima et al. (2004) reported perceived competence was more strongly correlated with WTC in high L2 use context. Japanese culture values more on self-criticism and sees less need for a positive self-regard (Heine, Lehman, Markus, & Kitayama, 1999), which may explain why
perceived competence is a strong predictor in Japanese students’ L2 WTC. All the studies in low L2 use contexts reported higher correlation in perceived competence with WTC than communication anxiety. More studies need to be conducted in the context of high L2 use to test the whether and how the result can be generalized.
Table 2.4  
*Correlations between WTC and perceived competence and communication anxiety*

<table>
<thead>
<tr>
<th>Studies</th>
<th>L2 use context</th>
<th>n</th>
<th>PC</th>
<th>CA confidence</th>
<th>L2 use confidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baker &amp; MacIntyre (2000)</td>
<td>L2 High</td>
<td>60</td>
<td>.17</td>
<td>-.44*</td>
<td></td>
</tr>
<tr>
<td></td>
<td>L2 Low</td>
<td>114</td>
<td>.72*</td>
<td>-.29*</td>
<td></td>
</tr>
<tr>
<td>Burroughs et al. (2003)</td>
<td>L2 Low</td>
<td>131</td>
<td>.8*</td>
<td>-.52</td>
<td></td>
</tr>
<tr>
<td>Cetinkaya (2006)</td>
<td>L2 Low</td>
<td>304</td>
<td>.57*</td>
<td>.03</td>
<td></td>
</tr>
<tr>
<td>Chu (2008)</td>
<td>L2 Low</td>
<td>364</td>
<td></td>
<td>-.416*</td>
<td></td>
</tr>
<tr>
<td>Clément et al. (2003)</td>
<td>L2 High</td>
<td>248</td>
<td></td>
<td>.7*</td>
<td></td>
</tr>
<tr>
<td></td>
<td>L2 Low</td>
<td>130</td>
<td></td>
<td>.87*</td>
<td></td>
</tr>
<tr>
<td>Fushino (2008)</td>
<td>L2 Low</td>
<td>772</td>
<td>.71*</td>
<td>-.22*</td>
<td></td>
</tr>
<tr>
<td>Kim (2005)</td>
<td>L2 Low</td>
<td>191</td>
<td>.44*</td>
<td>-.2*</td>
<td></td>
</tr>
<tr>
<td>Liu &amp; Jackson (2008)</td>
<td>L2 Low</td>
<td>547</td>
<td></td>
<td>-.535*</td>
<td></td>
</tr>
<tr>
<td>MacIntyre &amp; Charos (1996)</td>
<td>L2 High</td>
<td>92</td>
<td>.24*</td>
<td>-.24*</td>
<td></td>
</tr>
<tr>
<td>MacIntyre &amp; Doucette (2010)</td>
<td>L2 Low</td>
<td>238</td>
<td>.72*a</td>
<td>-.55*a</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>.88*b</td>
<td>-.54*b</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>.83*c</td>
<td>-.46*c</td>
<td></td>
</tr>
<tr>
<td>MacIntyre et al. (2002)</td>
<td>L2 Low</td>
<td>268</td>
<td>.474*</td>
<td>-.344*</td>
<td></td>
</tr>
<tr>
<td>MacIntyre et al. (2003)</td>
<td>L2 High</td>
<td>27</td>
<td>.4*</td>
<td>-.62*</td>
<td></td>
</tr>
<tr>
<td></td>
<td>L2 Low</td>
<td>32</td>
<td>.53*</td>
<td>-.18</td>
<td></td>
</tr>
<tr>
<td>Peng and Woodrow (2010)</td>
<td>L2 Low</td>
<td>503</td>
<td></td>
<td>.69*</td>
<td></td>
</tr>
<tr>
<td>Yashima (2002)</td>
<td>L2 Low</td>
<td>297</td>
<td></td>
<td>.68*</td>
<td></td>
</tr>
<tr>
<td>Yashima et al. (2004)</td>
<td>L2 High</td>
<td>60</td>
<td>.46*</td>
<td>-.15</td>
<td></td>
</tr>
<tr>
<td></td>
<td>L2 Low</td>
<td>154</td>
<td>.53*</td>
<td>-.25*</td>
<td></td>
</tr>
<tr>
<td>Yu (2010)</td>
<td>L2 Low</td>
<td>235</td>
<td>.504*</td>
<td>-.367*</td>
<td></td>
</tr>
</tbody>
</table>

*Note. a = McCroskey’s WTC Scale; b = MacIntyre’s WTC scale of inside classroom; c = MacIntyre’s WTC scale of outside classroom.*
2.4.5.2 The relationship between L2 WTC and frequency of L2 use.

MacIntyre’s (MacIntyre et al., 1998) heuristic model indicates that L2 use is the ultimate goal of language learning and it is directly predicted by Layer II of the model, willingness to communicate (cf. Figure 2.2). He claimed that “the ultimate goal of the learning process should be to engender in language students the willingness to seek out communication opportunities and the willingness actually to communicate in them” and “a program that fails to produce students who are willing to use the language is simply a failed program” (MacIntyre et al., 1998, p. 547). However, up to this date, only eight studies have examined the relation between WTC and observed or self-reported L2 communication behavior.

Among the nine studies which used a questionnaire to elicit self-reports of frequency of L2 use by asking participants how often he or she had engaged in various speaking activities, only four of them did report the correlation between L2 WTC and frequency of L2 use. The data are summarized in Table 2.4 Baker and MacIntyre (2000) reported that the correlation between WTC and frequency of L2 use was high in both immersion ($r = .45, n = 61$) and non-immersion program ($r = .68, n = 113$). (The original $n$ size for each group was 71 and 124, but 21 data were excluded due to the incompleteness of the questionnaires. However, the researchers did not specify how many data were excluded from each group, so I deducted 10 from the immersion group and 11 from the non-immersion group.) Clément et al. (2003) reported a correlation between the two variables of $r = .27$ among Anglophones ($n = 130$), who used L2 French in a low L2 use context, and $r = .28$ among Francophones ($n = 248$), who used L2 English in a high L2 use context. Yashima et al. (2004) also conducted the study in both
low and high L2 use contexts and reported $r = .42$ in the low L2 use context ($n = 154$) and $r = .27, .20, .28, .05, .07$ in the high L2 use context ($n = 60$), with the multiple correlation coefficients reported corresponding to five different frequency of L2 use items and WTC. Zeng (2010) reported a correlation of $r = .658$ between L2 WTC and L2 frequency of use in a high L2 use context. The study by Cao and Philip (2006) was unique in that they obtained the data of observed frequency of L2 use by observing classrooms using a scheme. However, in the end, they gave the conflicting conclusion that no clear correlations were found between learners’ self-reported WTC and observed oral behavior in classrooms, but they did not provide the statistics of any correlation coefficient.

Table 2.5

<table>
<thead>
<tr>
<th>Studies</th>
<th>High L2 use</th>
<th>Low L2 use</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>n size</td>
<td>Correlation</td>
</tr>
<tr>
<td>Baker and MacIntyre (2002)</td>
<td>61</td>
<td>.45*</td>
</tr>
<tr>
<td>Clément et al. (2003)</td>
<td>248</td>
<td>.28*</td>
</tr>
<tr>
<td></td>
<td></td>
<td>.27*</td>
</tr>
<tr>
<td></td>
<td></td>
<td>.20</td>
</tr>
<tr>
<td>Yashima et al. (2004)</td>
<td>60</td>
<td>.28*</td>
</tr>
<tr>
<td></td>
<td></td>
<td>.05</td>
</tr>
<tr>
<td></td>
<td></td>
<td>.07</td>
</tr>
<tr>
<td>Zeng (2010)</td>
<td>46</td>
<td>.658*</td>
</tr>
</tbody>
</table>

Table 2.5 shows that the correlation between WTC and self-reported frequency of L2 use is around medium. The correlation of the two variables is generally higher in low than high L2 use contexts. It was expected that there would be a higher correlation between the two variables in high L2 use context, since participants had more opportunities to actually use the L2. The results might be caused by the imbalanced number of participants in each context in the studies. A larger n size is more likely to lead
to a better result in inferential statistics. However, as the number of the studies which actually examined the question is quite limited, no generalization can be applied based on the data. Therefore, more studies are needed to investigate the relation between WTC and self-reported frequency of L2 use. Meanwhile, observed frequency of L2 use also needs to be examined, even though it may place more challenges on the operationalization of the research data collection process. The high or low context of L2 use in the previous studies was determined exclusively on the group of learners’ physical learning venues, including where the studies were conducted and whether the group of students was in an immersion or non-immersion program. However, the contexts of L2 use may vary on individuals. Some individual learners may use L2 more outside language classrooms, such as at home, at work, or with friends. No studies on L2 WTC so far have attempted to measure the context of L2 use of individual learners. Yeni-Komshian, Flege, and Liu (2000) developed a scale to measure language use in a study on pronunciation in L1 and L2. In future studies, a scale as Yeni-Komshian et al.’s can be used to record the language use contexts of individual learners in school, at home, and in the community.

2.4.6 Discussion and Conclusion

The discussion about the relation of L2 perceived competence and L2 communication anxiety with L2 WTC demonstrated that they might weigh differently in WTC under different L2 use contexts. Clearly, their relation under high L2 use context needs more exploration. The role of L2 use context has started to gain attention in the field and researchers have attempted to determine the L2 use context by self-reported questionnaires (e.g., Yeni-Komshian et al., 2000). The questionnaire exploring individual
L2 use context can surely help to investigate the impact of L2 use contexts on L2 WTC in future studies.

The primary studies that have investigated WTC in a quantitative design to date present a variety of characteristics in participants, research design, and instruments. Many studies failed to provide enough information of participants’ proficiency levels which may be one of the factors to cause differences in L2 WTC, even though Yashima (2002; Yashima et al., 2004) and D’Amico (2010) found no significant correlation between the two variables. On the other hand, it is also important to investigate the causal relation from L2 WTC to learners’ proficiency levels. The ultimate goal of language teaching is for language learners to use the language and achieve higher proficiency levels. The pedagogical significance of L2 WTC is confirmed if higher L2 WTC can be shown in empirical studies to lead to higher proficiency levels of language learners.

Though the primary studies were targeted at WTC in second language learning, most studies adopted the WTC scale by McCroskey, which was originally designed for L1 and does not provide items containing characteristics of second language learning. The fact that only four studies (MacIntyre et al., 2001; MacIntyre & Doucette 2010, Peng, 2007, Tannenbaum & Tahar, 2008) used MacIntyre’s L2 WTC scale results in the inconclusive examination on the variables within L2 WTC model. MacIntyre et al. (2001) reported that there was a substantial overlap between L2 WTC inside and outside the classroom for all the four skills: speaking, reading, writing, and comprehension. However, Peng (2007) reported that the differences in the WTC subscores of the four skills were statistically significant, with higher scores in speaking and writing, and low score in comprehension. On the other hand, Tannenbaum and Tahar (2008) concluded that L2
WTC outside the classroom was significantly lower than inside the classroom. Consequently, more studies are definitely needed to investigate the variables within the L2 WTC scale of MacIntyre. Only the study by MacIntyre and Doucette (2010) included both L2 WTC scales and the result shows the correlations obtained from each scale on L2 WTC with L2 PC and L2 CA do not contain big gaps. Nevertheless, more studies employing the two WTC scales and communication anxiety scales may help to show future researchers the potential differences caused by the different scales, and provide a platform for them to choose the appropriate scales for their studies. The frequency of L2 communication behavior was measured mostly through a self-reported questionnaire in the 27 experimental study samples, and through observations of actual conversational tasks in the four qualitative studies. Future research may attempt to combine the two types of measures of L2 use to examine whether there are any gaps between the self-reported and observed frequency of L2 communication behavior.

Even though the primary studies examined WTC in second language learning, the foci and features were quite different at different aspects. Readers are expected to be aware of the features and consider possible contributions they may make to L2 WTC and its related variables. Based on the current synthesis study, we can hypothesize that perceived competence contributes more in low L2 use contexts, and communication anxiety weighs more in L2 WTC in high L2 use context. More studies are needed to find out the generalized pattern underlying the WTC construct. Previous studies distinguished their learners’ L2 use contexts as a whole based on the language program or the overall second/foreign language learning environment. However, L2 use contexts may vary on individuals. Some learners may report higher individual L2 use contexts at home, at
school outside classroom, or in the community, and some may report lower individual L2 use contexts. Future studies employing self-reported questionnaires or actual observations to determine individual L2 use context may lead to more accurate and meaningful results on the question of the relation between individual L2 use contexts and L2 WTC.

In addition, researchers have tended to conduct their studies in Canada or Japan and have mostly investigated English as a second language. More studies examining L2 WTC in languages other than English are needed in the future to test whether L2 WTC and its related antecedents perform consistently across different languages and whether typologically different languages and their distinct social cultures will impose a different pattern on L2 WTC, as well as its related antecedents. At the same time, Mandarin as a less-commonly taught language has started to gain more attention in foreign language teaching. Xie (2011) investigated beginning learners’ L2 motivation self-system in learning Mandarin as a foreign language in the United States. The results showed that the three factors: ideal L2 self, international posture, and L2 WTC, were highly correlated both in heritage and non-heritage language learners’ of Mandarin. The fact that only one study targeted at L2 WTC in Mandarin (Xie, 2011) calls for the need for future studies to investigate this target language.

In sum, for future studies, more studies targeting languages other than English are needed. More studies are needed to examine the relation between WTC and frequency of L2 use. Future studies are also expected to test the role of L2 WTC on proficiency levels. When reporting research features, more detailed and clear explanations should be included in future studies and more justification on measurement scales is also needed for
readers to understand the motive of choosing or designing a particular scale and what possible result may be brought by the particular scales in the study.
Data obtained through Pew Hispanic Center and American Community Survey (2009) shows that in 2009 foreign-born residents in the United States have reached over 38 million which consists of 12.5% of total population in the United States. The Migration Policy Institute (2010) reported 3.2 million Chinese ethnics resided in the United States in 2008. These numbers clearly show the necessity of research focusing on the population with a language other than English from immediate or remote ancestors. This chapter will discuss the definitions of heritage language learners and point out problems and difficulties in defining Chinese heritage language learners. The chapter also reviews previous studies on affective factors of heritage language learners.

3.1 Definition of Heritage Language Learners

Heritage language learners can be defined in a broad and a narrow way. Fishman (2001) provided a broad definition of heritage language learners by including learners of indigenous, colonial, and immigrant languages, regardless of their prior knowledge or proficiency in the language. For example, if an African American wants to learn an African language, even though he or she doesn’t have any immediate family members speaking that language, he or she can still be defined as a heritage language learner because of the remote ancestors. The Heritage Language Research Priorities Conference Report at UCLA in 2001 posited a defining distinction between heritage language and foreign language acquisition that the former begins in the home and the latter usually
begins in a classroom setting. Valdés (2001) further narrowed down the definition and defined the term from the perspective of language teaching and learning. She posited that a heritage language learner “is raised in a home where a non-English language is spoken”, “speaks or at least understands the language and […] is to some degree bilingual in that language and in English” (p. 38). Valdés’ definition set up two criteria for heritage language learners: the language is spoken at home and the learner has certain proficiency on that language. Kondo-Brown and Brown (2008) followed the definitions of Valdés (2001) and the UCLA Steering Committee (2001) and defined heritage learners as “any learners who have acquired their cultural and linguistic competence in a non-dominant language primarily through contact at home with foreign-born parents and/or other family members” (p. 3). Carreira (2004) further defined heritage learners based on three criteria: “(a) the learner’s place in the heritage language community, (b) the learner’s personal connection to the heritage language and heritage culture through his/her family background, and (c) the learner’s proficiency in the heritage language” (p. 2). These definitions highlighted the features of heritage language learners that they at least initially acquired the language through family and have at least certain knowledge of the language before formal instruction at school.

However, even this narrower definition of who is a heritage language learner may need to be critically questioned. In a study comparing different profiles among heritage speakers, Kondo-Brown (2005) cautioned that not all heritage language learners with someone in their family background who speaks the language may be the same, and that heritage learners cannot be treated as a homogeneous group. Instead, she showed that certain family background characteristics are of central importance when investigating
affective and linguistic factors in heritage language learning. Specifically, Kondo-Brown divided heritage language learners in her sample into three categories: (a) those with at least one parent speaking the heritage language, (b) those with at least one grandparent but no parent speaking the language, and (c) those with heritage ethnic background but without any grandparents or parents speaking the target language. She then compared the linguistic skills, self-assessed use of the language, and self-perceived competence of the three subgroups of the heritage language learners. The results showed the majority of heritage learners with parents speaking the language (i.e., the (a) profile) have higher self-perceived competence and believed they can perform tasks from simple impersonal communication to narrative tasks. The other two subgroups of heritage language learners did not show the same degree of confidence. The learners with parents speaking the language also actually performed significantly better on linguistic skills than the other two subgroups of heritage language learners. And the latter two subgroups’ actual linguistic performance was nearly identical to learners of foreign language. Kondo-Brown’s (2005) findings are particularly important when defining “heritage” and assigning participants in the present study to the heritage and non-heritage language groups, since the heritage and non-heritage groups will be compared for affective and linguistic variables (this issue will be discussed in more detail in section 6.1.1).

If it is difficult to define heritage language learners, it is even harder to define heritage language learners of Chinese. “Chinese” is an umbrella term. It includes various dialects spoken in China. Chinese dialects can be divided into seven broad families: Wu, Xiang, Gan, Min, Cantonese, Hakka, and Mandarin (Norman, 1988, p.181). Many of them are not mutually intelligible (DeFrancis, 1984; Norman, 1988). As early as 1976,
Chao (1976) pointed out that Chinese dialects are “practically different languages.” It is often mentioned that Cantonese speakers cannot understand the Speakers of Mandarin just as Dutch speakers cannot understand English speakers or French cannot understand Spanish speakers (Chao, 1976, p. 97; DeFrancis, 1984, p. 38). Studies on Mandarin and the dialects also show great differences between them in phonology (e.g., Ting, 1982, 1983) and syntax (e.g., Hashimoto, 2001; Lien, 1994; McCawley, 1994). However, Duanmu (2000) pointed out that Chinese dialects share some systematic rules, especially in phonology. He put forward an example that low and following tones in Beijing dialect are falling and low in Chengdu dialects (p. 2). Therefore, he concluded that such systematic rules help speakers of different dialects understand each other rather quickly.

The distinguishing differences between dialects impose a question on defining heritage language learners of Mandarin. According to the report from the Migration Policy Institute (2010), there are 940,000 Chinese immigrants with limited English proficiency. Among them, 21.2% indicated they spoke Cantonese, and 63.7% indicated that they spoke Chinese which includes all kinds of dialects, and only 11.9% specifically indicated Mandarin was their mother tongue. In order to simplify the problem, most studies on Chinese language usually take an all-inclusive approach and consider that all Mandarin and dialect speakers are heritage language learners of Mandarin (Wang & Xiao, 2010, p. 153).

Unfortunately, currently, studies investigating heritage language learners rarely if ever distinguish between those whose ancestry included speakers of Mandarin and speakers of a dialect of Chinese, and when they report such distinctions they conduct their analyses combining data from both. He (2006) defined the heritage learners of
Chinese in her study as “a language student who is raised in a home where Chinese is spoken and who speaks or at least understands the language and is to some degree bilingual in Chinese and in English” (p. 1). She conducted a qualitative study to investigate the identity issues of Chinese heritage language learners of different dialects background. Ke (1998) investigated Chinese character learning of Chinese heritage learners who were defined as bilingual speakers of English and Chinese, including Mandarin Chinese or one of the Chinese dialects, such as Cantonese, Hokkian, or Southern Min (p. 94). Xiao (2006) compared language performances of heritage learners and non-heritage learners of Chinese. She included participants with various family language backgrounds, including Mandarin, Cantonese, Hokkian, and so on. Wang and Xiao (2010) explored the identity issues of heritage language learners of Chinese by interviewing Chinese speakers of various dialects, including Mandarin, Cantonese, Shanghainese, Taiwanese, and so on. Weger-Guntharp (2008) examined the affective needs of Chinese heritage language learners of family background in Mandarin and Cantonese. Xie (2011) conducted a study on motivation of heritage language learners who were defined as “those who have one or two parents from Chinese ethnicity, and the parent(s) speak Mandarin Chinese or another Chinese dialect at home or other places” (p. 18). However, Shen (2003) did a study on Chinese characters acquisition and she defined heritage language learners of Chinese as those that “come from Chinese-speaking families; namely, at least one of their parents is native Chinese and speaks that language at home” (p. 258). She further restricted heritage language learners whose parents speak Mandarin. More such comparative studies investigating the differences between heritage learners speaking Mandarin and those speaking a dialect of Chinese may provide
suggestions on whether an all-inclusive definition of heritage language learners of Chinese is a convenient and yet scientific way to conduct research.

In the present study, the complexities of these contested definitions and issue were weighed against the need to ensure a very large sample of participants in order to perform the needed powerful statistical analyses that would address the research questions. In the end, heritage language learners of Chinese were defined as those who have at least one parent (and in 6 cases in the sample, a grandparent only) speaking Mandarin or a Chinese dialect at home and receive linguistic and/or cultural influence of Chinese at home regardless whether or not they have certain linguistic proficiency levels before formal instruction of Mandarin at school.

3.2 The Primary Studies on Affective Factors of Heritage Language Learners

The affective factors investigated in the primary studies of heritage language learners include motivation, self-perceived competence, anxiety, self-expectation, and so on. Research on heritage language learners’ self-perceived competence has yielded controversial results.

Le (2004) conducted a study investigating affective issues of American students learning Chinese. He divided the 133 participants into three groups: Chinese background, Non-Chinese Asian background, and non-Asian background. When asked the question of “How do you rate your overall proficiency in the Chinese language as compared with the proficiency of other students in your class” (p. 73), the non-Asian group showed the highest self-perceived competence. Le (2004) explained that it may be caused by cultural values. Western cultures emphasize positive self-perception, and Asian cultures value
personal modesty (p. 74). With other heritage populations of language learners who hold other cultural values, the findings may be different. For example, Noels (2005) recruited 55 non-heritage learners and 41 heritage learners of German on their motivation and its substrates. The result shows that heritage language learners indicated greater self-perceived competence than non-heritage learners. Chen (2006) recruited 30 Chinese heritage language learners and used questionnaires, interviews, and classroom observation to investigate their motivational goals and emotional responses in Chinese learning classes. One of the interesting findings in the study was that heritage language learners generally attributed their success to ability, effort, luck, and task difficulty. Ability is the most commonly reason that learners believe leads to their good performance. However, when heritage language learners failed the tasks, they tended to blame themselves for not putting enough effort or they attributed it to lack of luck, instead of their ability. The findings show that the heritage language learners generally have belief in their competence and have high self-perceived competence in language learning.

Anxiety of heritage language learners is also a topic under dispute. Some researchers conclude that being a heritage speaker is positively associated with feeling of less anxiety. For example, Alghothani (2010) compared three foreign language learners and two heritage language learners of Arabic and reported a lower anxiety level in heritage language learners. On the other hand, other researchers have noted that cultural values and various pressures can generate an underlying anxiety among heritage learners that can reach comparable levels to the anxiety of non-heritage learners. For example, Le (2004) used Horwitz’s FLCAS score to examine the anxiety levels of the three groups
mentioned earlier: Chinese background, Non-Chinese Asian background, and non-Asian background. The result shows that the three groups’ anxiety levels were almost the same. Nevertheless, Le pointed out that there seemed to be different sources of anxiety. The Non-Asian group experienced lower anxiety than expected because of their cultural values of a positive self (p. 144). The other two groups, by contrast, experienced similar anxiety levels despite of their relatively higher proficiency levels because the Asian culture of shyness and inwardness may play a role in anxiety (p. 121). Still, there are other researchers who conclude that heritage language learners experience higher anxiety than non-heritage language learners. Among them, some argue that the causes for high anxiety level may come from high expectations from teachers, classmates, parents, and themselves; and others suggest that a source of high anxiety level comes from high personal goals of heritage learners of Chinese. I will review the evidence leading to these arguments below.

High expectations from teachers, classmates, parents, and themselves can cause anxiety among heritage language learners of Mandarin. For example, Comanaru and Noels (2009) reported that heritage learners generally show more pressure than non-heritage learners because of “pressures from others or a self-imposed feeling that they ought to learn the language” (p. 151). The same observation was made in Lacorte and Canabal’s study (2003). They investigated the classroom interactions of heritage language learners and concluded that heritage language learners may have to deal with unreasonable high expectations from teachers and peers. In addition, heritage language learners may feel foreign language learners have a better grasp of grammar knowledge and specialized terminology (p. 117). Hsiao (2010), different from most studies,
investigated the experiences of Chinese dialects speakers learning Mandarin. She reported the heritage language learners with Chinese dialects indicated a feeling of higher expectation from professors and classmates. The heritage learners of Chinese agreed that they have some advantages in learning Mandarin, but at the same time they also have different interferences from their dialects when learning Mandarin. One observation by one of the instructors was that dialect speakers have more difficulties in pronunciation and tones. Those students without any ethnic background learn what teachers have taught them, but Chinese dialect speakers always have to refer back to the pronunciations of their dialects. The high expectation from others and the interferences encountered in learning increase the anxiety level of heritage learners of Chinese with dialects background. The high expectation from parents is also confirmed in the study by Li (2006). Li interviewed parents of Chinese heritage language learners and reported a common theme in the interviews with parents is setting high standard and emphasizing the value of education (p. 146). Li gave an example of a parent criticizing her child for getting 96% in a test.

High personal goals of heritage learners of Chinese can also lead to considerable debilitating anxiety. In Carreira and Kagan’s study (2011), 61.9% of Mandarin and Cantonese heritage learners stated professional goals as their primary reason for studying their Chinese. At the same time, Carreira and Kagan (2011) expressed their worry that the ambitious goal would pose a pedagogical challenge on language teachers, since the heritage learners of Chinese have considerably less exposure to their heritage language in the United States as compared, for example, to Spanish heritage learners. This line of reasoning offers an interesting connection with another variable investigated in the
present study, frequency of L2 use, which Carreira and Kagan suggests has an impact on both proficiency and anxiety among heritage language learners of Chinese. Based on Carreira and Kagan’s statistics (2011), “only 10.5% of Chinese heritage learners visited their country of origin once a year and 45.3% never read in their heritage language and 82.2% never accessed the Internet in this language or accessed it only rarely” (p. 52). Le (2004) also pointed out the same problem. Upon answering the question of whether or not the learners want to become fluent in listening, reading, writing, and speaking, the groups of Chinese and non-Chinese Asian participants reported higher self-expectations in all the four skills than the group of non-Asian background did. Interestingly, Le (2004) also examined the beliefs of language learners of Mandarin and pointed out that students with Chinese ethnic background believed more that Chinese is a difficult language than those with non-Chinese ethnic background, despite of the fact that the former usually have higher proficiency levels. Again, Le (2004) explained that it may be caused by the higher expectation and goals of Chinese-ethnic students on learning their ancestral language.

Some of the conflicting or disagreeing findings reviewed in this section may reflect the uncertainty of broad and narrow definitions of heritage language learners and the varieties of Chinese (Mandarin or a dialect) in their ancestry, already discussed in section 3.1.

3.3 Summary

Heritage language learners can be defined based on a broad or narrow category. As broadly defined, heritage language learners can be those who are learning an ancestral language regardless of close or remote ancestors. If defined narrowly, they can be the
learners who speak or listen to the heritage language at home, for example, from one or both parents and/or one or both grandparents, and have certain proficiency levels of the language. Important affective and linguistic differences among certain heritage learner profiles, even in the narrow definition, may exist, depending on whether the ancestral language is used at home by at least one parent, which is likely to result in higher self-esteem and stronger linguistic skills, versus only grandparents (Kondo-Brown, 2005). In addition, the many dialects in the Chinese language increase the difficulty of defining Chinese heritage language learners. The majority of studies in the past have taken an all-inclusive definition approach when recruiting participants and defined all language learners exposed to any Chinese dialects at home as heritage learners.

The primary studies on heritage language learners have shown conflicting results on this learner population’s self-perceived competence and anxiety levels. Some concluded that heritage language learners have higher self-perceived competence because of their typically higher actual proficiency levels, and some asserted that they show lower self-perceived competence because of their cultural values of being modest and inward. The results for anxiety of heritage language learners also failed to reach an agreement. However, there is consensus that the sources for heritage learners’ anxiety are likely to come from high expectations from teachers, peers, parents, and high self-chosen goals that they pose to themselves.

It is often assumed that heritage learners typically are or become more proficient than foreign language learners. However, Kondo-Brown’s study (2005) discovered that the heritage learners with grandparents but without parents speaking the language at home had similar proficiency levels as non-heritage language learners. Therefore, without
real good measures of proficiency to use in the studies, the assumed higher proficiency levels of all heritage learners when compared with foreign language learners remains more of an anecdote than an empirical fact. In the current study, the development and use of an EI task that measures proficiency for research purposes as discussed in later chapters can help make progress in understanding not only affective differences between heritage and foreign language learners of Chinese, but also how proficiency may differentiate between heritage and foreign language learners and how proficiency differences may be implicated in issues of affect and motivation in heritage language learning.
CHAPTER 4
ELICITED IMITATION AS A MEASURE OF GLOBAL PROFICIENCY OF MANDARIN AS A SECOND/FOREIGN LANGUAGE

Elicited imitation (EI) is a language data elicitation method that has been mainly used to investigate three areas of language development: (a) first language acquisition (e.g., Clay, 1971), (b) neuropsychological research (e.g., Kouri, 2005; Rogers et al., 2008), and (c) second/foreign language acquisition (e.g., Erlam, 2006; Hamayan, Saegert, & Laraudee, 1977; Jessop, Suzuki, & Tomita, 2007; Markman, Spilka, & Tucker, 1975). Elicited imitation requires language learners to hear and then repeat a certain number of sentences in a language. Their performances on the language are graded based on how well they have repeated the sentences. The results of EI have been used to investigate the nature of child language development in L1 (e.g., Chien & Lust, 1985; Song & Fisher, 2001) and to compare the differences of L1 development in children and adults (e.g., Dye, 2006; Valian & Aubry, 2005). In the area of neuropsychological research, EI has been a tool to study language development of special children (e.g., Ingersoll, 2010). The main purposes of EI in research of second language acquisition are varied and include testing the theory of Universal Grammar (e.g., Martohardjono & Flynn, 1995), tapping implicit knowledge (e.g., Ellis, 2005; Erlam 2006), measuring the performance of specific grammar structures (e.g., Lindsey, 2009; Maciukaite, 2004; Thomas, 1991), and gauging overall oral competence of a target language (e.g., Hameyer, 1980).

There is a rich body of methodological discussions regarding the advantages of using EI tasks in language-related research and caveats for optimal choices in the design,
best purposes for use, and appropriate interpretations. Some researchers suggested that elicited imitation offers certain distinct advantages when measuring overall learners' proficiency levels. This is because when compared with measures such as grammatical judgment task and multiple-choice test, elicited imitation requires online processing and measures language competence, instead of language prescriptive knowledge (Munnich, Flynn, & Martohardjono, 1994). In addition, elicited imitation is easier to administer and grade than some other online processing measures, such as oral interview or timed essay writing (Jessop, Suzuki, & Tomita, 2007). The former takes shorter time and has less extraneous variables in administration with recorded stimuli and less subjective in scoring with well-developed scoring criteria. In terms of the underlying theory as to why language proficiency might be optimally captured by repeating sentences one has heard, researchers believe that a well-designed elicited imitation task does not involve parroting or echoing but is reconstructive (e.g., Erlam, 2006; Munnich, Flynn, & Martohardjono, 1994) and consists of three processes: "(a) processing a stimulus sentence, (b) reconstructing it with their own grammar, and (c) reproducing it" (Jessop, Suzuki, & Tomita, 2007, p. 215). However, there are also some concerns when using elicited imitation as a measure of global language competence (as it will be used in the present study) or as a measure that taps competence in specific grammatical subsystems of the language. The major concern is that the sentence stimuli must be well designed in order to avoid undesirable effects which include: (a) rote memory effects, (b) serial position effects, and (c) ceiling and floor effects (Bley-Vroman & Chaudron, 1994). This chapter will systematically compare and review the previous primary studies involving EI in order to identify similarities and differences in these studies. Through the comparison, the
chapter attempts to locate the reasons which originate the above concerns and suggest appropriate design and use of EI to minimize the possible problems brought up by previous literature.

4.1 Problem Specification with the Primary Studies on L2 EI

The EI in the studies of second language acquisitions has been used to serve different purposes. Some used EI to measure participants’ performance on certain structures, and others employed EI to measure global competence of a target language. In summary, the EI tasks mainly serve the purpose of the following categories:

1. To test the EI as measure tool by comparison to other tools;
2. To test the theory of Universal Grammar (UG);
3. To test the effect of pedagogical interventions;
4. To test the implicit knowledge of a language learner;
5. To test the influence of other factors on participants’ performance, including age and syntactic complexity.

The studies testing the utility of EI as a measure tool are methodological in nature and compared EI with different measures, including the International English Language Testing System (IELTS) of English proficiency (Daugherity, 2008), free production (Eisenstein, Bailey, & Madden, 1982; Naiman, 1976), and grammaticality judgments (Munnich et al., 1994). Daugherity (2008) found a moderate correlation between the listening part of IELTS and EI and concluded that EI was a valid tool to measure basic language ability. Eisenstein et al. (1982) reported a low but significant correlation between cued oral production and EI tasks on certain structures. Munnich and his
colleagues (1994) also confirmed that EI and the grammatical judgment tasks provide convergent results on grammatical sentence structures. They confirmed that EI is a powerful and sensitive instrument to measure knowledge of grammar. Naiman (1976) compared EI with comprehension and oral production tasks. He found out that performance of comprehension is equal or greater than the performance of EI, and in turn the performance of EI is equal or greater than that of production. He suggested that EI was a good predictor of how learners learning the language.

The studies using EI to test the theory of UG tackle different research questions and hypotheses that stem from this formal linguistic theory. Martohardjono and Flynn (1995) and Roebuck, Martinez-Arbelaiiz, and Perez-Silva (1999) investigate the first language transfer hypothesis. Both studies recruited participants with different backgrounds of first languages and compared their performances on certain structures measured by EI tasks. Kaplan (1993), Thomas (1991) and Van Boxtel, Bongaerts, and Coppen (2005) elicited EI evidence on specific structures of interest and compared these performances of second language learners with those of native speakers in order to examine whether second language learners have the same access to UG and can achieve native level proficiency in certain areas of language that are of interest in UG theory. Batmanian, Sayehli, and Valian (2008), Hamayan, Saegert, and Larudee (1977), and Myles (1996) investigated the developmental features of certain structures in the process of second language learning by using EI to elicit language samples.

There have also been studies that used EI tasks to examine the effect of a certain pedagogical intervention. These studies usually included a pre- and a post- EI task so as to compare the results of participants’ performances. Ellis, Loewen, and Erlam (2006)
and Li (2010) investigated the effects of feedback in language learning. Fiori-Agoren (2005) checked whether the technique of consciousness-raising assisted the learning of a certain grammar structure in the target language. Thomas (2010) compared the effect of input frequency and lexical aspect in language learning measured by an EI task. Burger and Chretien (2001) and Trofimovich, Lightbown, Halter, and Song (2009) examined the effect of a long term course or practice. The former focused on the effect of a language adjunct course on learners’ oral performance over two semesters. And the latter looked at the effect of a long-term listening and reading comprehension practice on learners’ pronunciation ability. Li (2010) is the only study involving EI that targeted Mandarin as the target language. Seventy-eight L2 Mandarin learners were recruited in the study and their performances on a structure of interest measured by an EI test and a grammaticality judgment test were compared before and after implicit and explicit feedback.

Ellis (2005, 2006) and Erlam (2006) used EI to test implicit knowledge of language learners. They concluded that EI task is constructive and is a valid test to measure implicit knowledge. Reid (1981) and Ortega (2000) used EI to measure global proficiency in a study that examined the relationship between proficiency and different aspects of syntactic complexity. Reid (1981) put forth the research question of whether complexity of overall sentence would cause difficulty in acquiring a certain grammar structure in the target language. Ortega (2000) intended to find out the correlation between syntactic complexity and global proficiency as measured by EI and other instruments.

The primary studies also differ in participants’ characteristics. Many studies chose primary or second school students as participants (e.g., Naiman, 1976; Reid, 1981), and
some recruited college students as participants (e.g., Ellis, 2006; Li, 2010). The sample sizes of participants in different studies vary from a minimum of 6 (Bartning & Schlyter, 1987) to a maximum of 220 (Ellis, 2006). The studies included participants at various proficiency levels from beginning to advance. The first languages of participants, as well as the target language in the studies, fall into a big variety.

The designs of the EI tasks in the studies are different in many aspects. The total numbers of sentences in the EI tasks ranged from 6 (Thomas, 1991; Trofimovich et al., 2009) to 129 (Hameyer, 1980). Some studies (e.g., Martohardjono & Flynn, 1995) took control of the number of syllables or chunks in each sentence to avoid any effects of rote memory, and some (e.g., Ellis et al., 2006) did not take this factor into consideration. Most studies presented the stimuli sentences only once for participants to repeat, and few presented them twice. Some studies employed ungrammatical sentences in the task in order to observe whether ungrammaticalities would be corrected or repeated as heard, and some did not. Most studies required immediate repetition after each stimulus and gave unlimited time to the participants to produce the repetition. Still some studies provided contextual assistance, such as cued pictures or vocabulary lists. These differences in EI design have been discussed in terms of options that may be less or more appropriate for specific purposes for using the EI methodology in the first place (e.g., to measure global proficiency versus to measure knowledge of specific grammatical subsystems involved in UG) and the various methodological choices have also been discussed as potentially affecting the validity and reliability of EI tasks.

In addition to differences in the design of stimuli and instructions, the studies have also adopted different scoring procedures for EI tasks. Some scoring systems ignore
any repetition errors except for those relevant to the target structures under study. This scoring procedure is typical of studies that focus on measuring knowledge of specific grammatical structures of the target language. As long as the participant repeated the target structure correctly in a sentence, he or she got credit for the repetition. However, many studies intended to measure the global competence of a language learner. As a result, they graded the elements of pronunciation, syntactic and discourse features (e.g., Burger and Chretien, 2001).

All the above differences in purposes, participants’ characteristics, EI design, and EI scoring procedures make the assessment of validity and reliability of the instrument of EI a complex task. As a result, researchers have identified different advantages and cautions in the use of EI as a research tool. This chapter intends to review and compare systematically the studies involving EI and locate the main arguments for valid and reliable design of EI tasks as tools to measure global proficiency. It will discuss the rational base for the design and use of EI for measuring global proficiency, since this is the purpose of using the EI methodology in the present study.

4.2 Purpose of Present Research Synthesis

As in chapter 2, in this chapter I conducted a synthesis study (Norris and Ortega, 2006; Ortega, 2010) in order to better understand the uses of EI tasks for research purposes and to distill recommendations for optimal design of EI as a measure of global proficiency. Accordingly, this chapter will present and categorize different features of the EI studies uncovered, including characteristics of participants, features of EI design, and
scoring procedures followed in different studies. It also hopes to find the answers to the following questions and provide suggestions for future EI use in research:

1. How have EI tasks been used in research on second language acquisition?
2. What are the design differences in EI tasks in the primary studies?
3. What are the analysis procedures of EI tasks in the primary studies?

4.3 Method

4.3.1 Study Identification and Retrieval

In order to locate all the primary studies involving EI tasks in a second language, I conducted an extensive and inclusive search. I first conducted an extensive online search in the databases and then a reference tracing via foot notes.

I conducted an extensive online search in the following three databases through Cambridge Scientific Abstracts (CSA): Educational Resources Information Center (ERIC), Linguistics and Language Behavior Abstracts (LLBA), and MLA International Bibliography. To increase the ability for researchers to replicate the synthesis study, I limited the search to “English only”. The key term used to retrieve data was “KW=((elicited imitation) or (imitation test) or (repetition test)) and KW=(second language)”.

The results show there are total 22 primary studies retrieved from the above three databases. Then I carried out an online search at the database of PsychInfo using the key term “elicited imitation:Abstract OR elicitation test:Abstract OR imitation test:Abstract AND second language:Abstract”. No additional study was retrieved. In order to minimize publication bias and assure that I had retrieved all the relevant doctoral dissertations, I explored the database of Dissertations and Theses through ProQuest by
the following key term: “ABS(elicited imitation) OR ABS(imitation test) OR ABS
(repetition test) AND ABS(second language)”. One more study was located in this
database.

Table 4.1
*Online studies retrieval through online database*

<table>
<thead>
<tr>
<th>Database</th>
<th>Key terms</th>
<th>Studies</th>
</tr>
</thead>
<tbody>
<tr>
<td>CSALLBA</td>
<td>KW=((elicited imitation) or (imitation test) or (repetition test))and KW=(second language)</td>
<td>14</td>
</tr>
<tr>
<td>MLA</td>
<td>KW=((elicited imitation) or (imitation test) or (repetition test))and KW=(second language)</td>
<td>2</td>
</tr>
<tr>
<td>ERIC</td>
<td>KW=((elicited imitation) or (imitation test) or (repetition test))and KW=(second language)</td>
<td>6</td>
</tr>
<tr>
<td>PsycInfo</td>
<td>elicited imitation:Abstract OR elicitation test:Abstract OR imitation test:Abstract AND second language:Abstract</td>
<td>0</td>
</tr>
<tr>
<td>Dissertation</td>
<td>ABS(elicited imitation) OR ABS(imitation test) OR ABS (repetition test) AND ABS(second language)</td>
<td>1</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>23</td>
</tr>
</tbody>
</table>

In addition, I tracked all references cited in the study by Jessop, Suzuki, and
Tomita (2007), and one additional primary study was found. Therefore, all together 24
primary studies were retrieved. However, after close inspection, two studies conducted by
Fiori-Agoren (2004, 2005) and by Van Boxtel and his colleagues (Van Boxtel et al., 2003,
2005) employed the same participants in both studies, respectively, so each study was
counted only once. It is also turned out that the studies by Ellis (2005, 2006) and the
study by Erlam (2006) all targeted the same participants. As a result of removing
duplicated studies, a final total 20 of independent studies were eligible for the synthesis
study in this chapter. All the studies included are marked with a number sign (#) in the
references.
4.3.2 The Inclusion and Exclusion Criteria

The studies in the present synthesis were selected based on the following inclusion and exclusion criteria:

1. They were (quasi-)experimental in design. Qualitative Studies were excluded, but will be discussed briefly on their methodological features (Kaplan, 1993; Martohardjono & Flynn, 1995). These qualitative studies are marked with two number signs in the references.

2. They involved participants who were learning the languages elicited via the EI task as second or foreign languages. Studies which involved first language learning were excluded (e.g., Clay, 1971; Hirata-Edds, 2007). Studies which involved a dialect of a language were also excluded (Day, 1974; Gallimore & Tharp, 1976, 1981).

3. They investigated second or foreign language learning by adult learners. Studies which targeted on language learning by children were excluded (e.g., Arif, 1999; Hamayan, 1978; Naiman, 1976; Trofimovich et al., 2009).

4. They employed EI tasks as an instrument to measure performance of an actual language. Studies requesting participants to repeat pseudo words in an EI task were excluded (e.g., Chun & Payne, 2004; Payne & Ross, 2005).

5. They requested participants to orally repeat stimuli sentences. Studies requesting participants to write down stimuli sentences were excluded (e.g., Downs, 2009; Sauerbrey & Scheibner-Herzig, 1988).

6. They involved EI tasks with sentences repetition. Studies involving words repetition tasks were excluded (e.g., Barcroft, 2000).
7. They were published articles in journals or books, and unpublished doctoral dissertations. Working papers, unpublished manuscripts, and Master’s theses (e.g., Ghasemi, 2004; Hamayan, 1975, 1976; Naiman, 1974) were excluded to avoid the potential bias caused by their partial availability, as well as to ensure the quality of the included studies and make it easier for readers to replicate the synthesis study.

8. They were written in English to enable researchers to replicate the synthesis study.

9. If the same sample was reported in several studies, only one sample was included in the current synthesis study. The other study was referred for more information in coding.

4.3.3 Coding of the Primary Studies

A coding book was developed in order to collect sufficient information on the primary studies to compare the characteristics of the studies and find underlying relationships and gaps across the study characteristics and the designs. Five main categories related to EI tasks were coded: (a) publication characteristics, (b) participants’ background, (c) research design, (d) analysis procedure, and (e) research purposes.

The coding of publication characteristics includes (a) author(s), (b) year, and (c) type of publication. This part of coding makes it easier for researchers to locate the studies and make sure that these studies are journal articles, doctoral dissertations, or book chapters to ensure the quality of the primary studies. The coding of participants’ background includes (a) the n size of participants, (b) proficiency levels, (c) proficiency assessment types, (d) school status of participants, (e) first language of the participants,
and (f) target language of the study. The relevant sections in the studies were read carefully to locate necessary information. The school status were categorized based on whether the participants were in universities or secondary schools. As studies involving children were excluded, the school status of primary school and kindergarten are not included in the coding. The categories proposed by Thomas (2006) were used in coding the variables related to proficiency: (a) impressionistic judgment, (b) institutional status, (c) in-house assessment, and (d) standardized test (pp. 282-283). I looked for the words, such as “beginning”, “intermediate”, or “advanced” in the relevant sections. If such impressionistic descriptions and words were not found in the study, I looked for the description of their institutional status. I coded participants as “beginning learners” if they were described as “enrolled in a first or second semester language course”, and “intermediate learners” if they were at the second year of the language course. They were coded as “advanced learners” if they had enrolled in the third year or higher of the language course. Correspondingly, if learners’ proficiency levels were defined based on the language courses they were enrolled in, the assessment type was coded as “institutional status”. For a study (Li, 2010), the research defined the participants as “high” and “low” groups based on the standardized tests they took for the target language. In addition, the research also provided the participants’ institutional levels as 4th, 6th, and 8th semesters of the Chinese study. As a result I coded the proficiency levels of the participants as advanced and intermediate and the proficiency asessment type as standardized test.

The task design of elicited imitation was also carefully coded. It includes total numbers of the elicited imitation sentences and administration features of the EI tasks,
including whether the stimuli sentences were presented orally by an experimenter or through a tape recorder, and whether the sentences were presented once or more than once. Skills that were measured were also coded in this section. Worded explanations containing indications of measuring grammar structures, oral proficiency, and so on were carefully searched and recorded in the coding book accordingly. Due to the criticism that ill-designed EI stimuli can be influenced by rote memory, I coded features on efforts taken by particular researchers to control any effects of rote memory, including (a) grading stimulus sentence length so as to surpass working memory limits for echoic repetition, and (b) interventions to postpone repetition to avoid direct parroting. I looked for the descriptions of sentence length in EI tasks, which could be reported in words (or characters, for languages such as Mandarin) or in syllables. If no description was provided, I looked for appendices potentially containing the actual EI sentence stimuli used and counted the words/characters/syllables of the sentences. For the intervention to postpone repetition, if there was no clear description of such intervention and the author described the procedures as “after the sentence presentation, the participants started to repeat the sentence”, then I coded it as no intervention included. In order to find out evidence of EI is an online and constructive instrument, the following relevant features were coded, including (a) time allowed to repeat the sentences, and (b) use of ungrammatical sentences. I looked for the phrases as “unlimited time” or “as much time as they need”. If the study provided a description like “the recording was paused for participants to repeat the sentences”, unlimited time was coded. For the use of ungrammatical sentences, if the study did not mention anything about ungrammatical sentences, it was coded as no use of ungrammatical sentences. It is assumed that if the
researcher employed such technique, he or she would have definitely included this
information in the report of the study. Whether any additional support, such as contextual
or vocabulary support, was provided, as well as whether any parallel measures were used
in the study were also coded as design features of the EI study.

The EI scoring procedures were also coded. Information about score range and
scoring criteria was located and recorded. Any report of EI reliability was also searched
for, but few studies actually provided this information. For studies using parallel
measures, correlations between the measures were also recorded to compare the validity
of the EI instrument. Finally, the research purposes of each study were also recorded. The
coding of research purposes is essential to find out what research problems are most
widely investigated, how well design decisions adopted by individual researchers respond
to those purposes, and what issues and problems need more examination in future EI
designs (See the complete coding book at Appendix B).

4.4 Results

4.4.1 Participants’ Characteristics

Table 4.2 below shows the participants’ characteristics of the 20 independent
samples contained in the 24 primary studies. A total 1273 language learners participated
in the studies. The study (Ellis, 2005) with maximum n size included 220 participants and
the minimum n size was 12 participants in the study by Munnich, Flynn, and
Martohardjono (1994). The mean of the n size in the 20 independent samples is 63.56
with standard deviation of 59.89 (the samples sizes for each individual study are listed in
Table 4.3).
Table 4.2

<table>
<thead>
<tr>
<th>k</th>
<th>Min.</th>
<th>Max.</th>
<th>M</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>12</td>
<td>220</td>
<td>63.56</td>
<td>59.89</td>
</tr>
</tbody>
</table>

*Note.* k = number of studies.

The proficiency levels of the participants in the independent samples vary from beginning to advanced. The largest category of samples featured in the studies (n = 8, 40%) included participants of mixed proficiency levels. Four out of the 20 independent samples (20%) recruited participants at intermediate and advanced level separately. Three (15%) included beginning learners of the target language and only one study (5%) failed to report the proficiency levels of the participants. The proficiency description types differed in the different primary studies. Based on the four assessment categories established by Thomas (2006), institutional status is still the most employed method to define language learners’ proficiency levels in this EI research (n = 7, 35%). In-house assessments and standardized tests are the two assessment types that are equally popular in the primary studies (n = 3, 15%). Only two studies (10%) used impressionistic judgment to decide participants’ proficiency levels. Three studies out of 20 (15%) used more than one assessment type. Among the combined assessment types in these three studies, institutional status is still the most popular option, which is combined with an in-house test and a standardized test in two studies. The other study of the three used in-house assessment and standardized test to decide participants’ proficiency levels. The assessment type used in one study (5%) does not fall into one of the four categories by Thomas (2006). The researcher (see Table 4.3) collected information on the number of years of learning on the target language and decided the proficiency levels of the
participants. Unfortunately, one study (5%) failed to provide information on how the research defined the proficiency levels of the participants.

As the current synthesis study excluded the primary studies with children language learners, it is not surprising perhaps that 17 out of 20 (85%) independent primary samples recruited their participants from college. One study (5%) included participants from both secondary school and college. One study (5%) has participants from an adult language school. Again, one study (5%) did not provide information on the school status of their adult participants.
Table 4.3
*Participants’ characteristics in 20 independent samples of the primary studies*

<table>
<thead>
<tr>
<th>Studies</th>
<th>N</th>
<th>Proficiency level</th>
<th>Proficiency assess</th>
<th>School status</th>
</tr>
</thead>
<tbody>
<tr>
<td>Batmanian et al. (2008)</td>
<td>22</td>
<td>beginning</td>
<td>institutional &amp; in-house</td>
<td>college</td>
</tr>
<tr>
<td>Burger &amp; Chretien (2001)</td>
<td>31</td>
<td>advanced</td>
<td>in-house</td>
<td>college</td>
</tr>
<tr>
<td>Daugherity (2008)</td>
<td>25</td>
<td>mixed</td>
<td>Impression</td>
<td>secondary &amp; college</td>
</tr>
<tr>
<td>Eisenstein et al. (1982)</td>
<td>45</td>
<td>intermediate</td>
<td>institutional</td>
<td>college</td>
</tr>
<tr>
<td>Ellis et al. (2006)</td>
<td>34</td>
<td>intermediate</td>
<td>in-house</td>
<td>college</td>
</tr>
<tr>
<td>Fiori-Agoren (2004)</td>
<td>59</td>
<td>intermediate</td>
<td>institutional</td>
<td>college</td>
</tr>
<tr>
<td>Flynn (1986)</td>
<td>51</td>
<td>mixed</td>
<td>missing</td>
<td>college</td>
</tr>
<tr>
<td>Hameyer (1980)</td>
<td>22</td>
<td>mixed</td>
<td>years of study</td>
<td>college</td>
</tr>
<tr>
<td>Henning (1983)</td>
<td>143</td>
<td>mixed</td>
<td>institutional</td>
<td>college</td>
</tr>
<tr>
<td>Li (2010)</td>
<td>78</td>
<td>mixed</td>
<td>standard</td>
<td>college</td>
</tr>
<tr>
<td>Munnich et al. (1994)</td>
<td>12</td>
<td>advanced</td>
<td>standard</td>
<td>college</td>
</tr>
<tr>
<td>Myles (1996)</td>
<td>158</td>
<td>mixed</td>
<td>institutional</td>
<td>college</td>
</tr>
<tr>
<td>Ortega (2000)</td>
<td>16</td>
<td>intermediate</td>
<td>institutional</td>
<td>college</td>
</tr>
<tr>
<td>Perkins et al. (1986)</td>
<td>50</td>
<td>advanced</td>
<td>institutional</td>
<td>college</td>
</tr>
<tr>
<td>Roebuck et al. (1999)</td>
<td>28</td>
<td>missing</td>
<td>in-house</td>
<td>college</td>
</tr>
<tr>
<td>Thomas (1991)</td>
<td>173</td>
<td>mixed</td>
<td>in-house &amp; standard</td>
<td>college</td>
</tr>
<tr>
<td>Thomas (2010)</td>
<td>33</td>
<td>beginning</td>
<td>institutional</td>
<td>college</td>
</tr>
<tr>
<td>Van Boxtel et al. (2005)</td>
<td>43</td>
<td>advanced</td>
<td>impression</td>
<td>missing</td>
</tr>
<tr>
<td>Wang (1998)</td>
<td>30</td>
<td>beginning</td>
<td>standard</td>
<td>language school</td>
</tr>
</tbody>
</table>
The two qualitative studies uncovered both recruited their participants from college. One of them (Martohardjono & Flynn, 1995) included 77 participants at intermediate and advanced levels. The study by Kaplan (1993) is longitudinal and the participants overlapped in different EI tests over time. The total number of the participants is not provided in the research, but the study did describe the proficiency levels of the participants as mixed levels. Martohardjono and Flynn (1995) used a standardized test to decide the proficiency levels, and Kaplan (1993) recorded the length of prior studies of the target language and defined the proficiency levels in this way.

Table 4.4 provides the features of participants’ first languages and the target languages. The first languages the participants speak in the primary studies included Japanese \((n = 1, 5\%)\), Mandarin \((n = 1, 5\%)\), Swedish \((n = 1, 5\%)\), Arabic \((n = 1, 5\%)\), Spanish \((n = 1, 5\%)\), English \((n = 2, 10\%)\). The majority \((n = 12, 60\%)\) included participants speaking various first languages, including French, Turkish, Gujarati, and Korean, in addition to the above mentioned languages. For the target languages that the primary studies investigated, English \((n = 11, 55\%)\) was the most commonly examined language. Other target languages includes Spanish \((n = 2, 10\%)\), French \((n = 2, 10\%)\), Mandarin \((n = 1, 5\%)\), Dutch \((n = 1, 5\%)\), and German \((n = 1, 5\%)\). Two study \((10\%)\) investigated more than one language: Burger and Chretien (2001) studied English and French at the same time, and Thomas (1991) examined English and Japanese in her study.

The two qualitative studies examined English and Japanese as target languages, respectively. Martohardjono and Flynn (1995) included participants speaking Japanese, Chinese, and Spanish as their first languages. Kaplan (1993) has English and Chinese speakers as the participants.
Table 4.4
Language features of the primary studies

<table>
<thead>
<tr>
<th>Studies</th>
<th>1st LG</th>
<th>Target LG</th>
</tr>
</thead>
<tbody>
<tr>
<td>Batamanian et al. (2008)</td>
<td>mixed</td>
<td>English</td>
</tr>
<tr>
<td>Daugherity (2008)</td>
<td>Gujarati &amp; Spanish</td>
<td>English</td>
</tr>
<tr>
<td>Eisenstein et al. (1982)</td>
<td>mixed</td>
<td>English</td>
</tr>
<tr>
<td>Ellis et al. (2006)</td>
<td>mixed</td>
<td>English</td>
</tr>
<tr>
<td>Fiori-Agoren (2004)</td>
<td>missing</td>
<td>Spanish</td>
</tr>
<tr>
<td>Flynn (1986)</td>
<td>Spanish</td>
<td>English</td>
</tr>
<tr>
<td>Hameyer (1980)</td>
<td>mixed</td>
<td>German</td>
</tr>
<tr>
<td>Henning (1983)</td>
<td>Arabic</td>
<td>English</td>
</tr>
<tr>
<td>Li (2010)</td>
<td>English &amp; Korean</td>
<td>Chinese</td>
</tr>
<tr>
<td>Munnich et al. (1994)</td>
<td>Japanese</td>
<td>English</td>
</tr>
<tr>
<td>Myles (1996)</td>
<td>English</td>
<td>French</td>
</tr>
<tr>
<td>Ortega (2000)</td>
<td>English</td>
<td>Spanish</td>
</tr>
<tr>
<td>Perkins et al. (1986)</td>
<td>mixed</td>
<td>English</td>
</tr>
<tr>
<td>Roebuck et al. (1999)</td>
<td>Spanish &amp; Chinese</td>
<td>English</td>
</tr>
<tr>
<td>Thomas (2010)</td>
<td>Swedish</td>
<td>French</td>
</tr>
<tr>
<td>Van Boxtel et al. (2005)</td>
<td>German &amp; French &amp; Turkish</td>
<td>Dutch</td>
</tr>
</tbody>
</table>

4.4.2 Research Design

Table 4.5 demonstrates the design features of EI tasks. The studies designed EI tasks containing various numbers of sentences. The maximum number of sentences included in an EI task is 129 and the minimum is three with the mean of 31.2 and the standard deviation of 29.88. One study (Li, 2010) administered three EI tests, pre-, post-, and delayed EI test. The item numbers of the three tests vary from 23 in the first two EI and 40 in the delayed EI test.

The administration procedures of the primary studies are also different. Fourteen out of 20 (70%) unique samples tape recorded the stimuli sentences to control extraneous...
variables in the test administration. Three studies (15%) orally presented the stimuli sentences. One study (Munnich et al., 1994) compared the oral and taped administration and used both in the study. Two studies (10%) did not provide information on how they presented the stimuli sentences. Most studies (n = 15, 75%) presented each sentence once and five studies (25%) presented twice before the participants were required to repeat the sentence.

Fourteen studies (70%) included only grammatical sentences in the EI tests. Six out of 20 studies (30%) used ungrammatical sentences to investigate whether the participants had enough grammatical competence to be able to correct the ungrammaticalities in repeating these sentences in the EI tasks. Among the six studies, three instructed the participants to “repeat in correct English.” Two studies did not provide explicit instruction for the participants to correct the sentences, so it was left to the participants to notice the ungrammaticality in some items and to repeat them as heard or in corrected form. And Roebuck et al. (1999) instructed the participants to repeat “what you have heard as precisely and as clearly as possible” (p. 282) to tap the unconcious knowledge of the grammar and test the interlanguage of speakers of different native languages. Unfortunately, the last study of the six did not provide information on how the participants were instructed to repeat the sentences, even though ungrammatical sentences were included and studied.
Table 4.5
Design features of EI tasks I

<table>
<thead>
<tr>
<th>Studies</th>
<th># of items</th>
<th>sentence length</th>
<th>taped/oral presentation</th>
<th>times of presentation</th>
<th>ungrammatical sentences</th>
</tr>
</thead>
<tbody>
<tr>
<td>Flynn (1986)</td>
<td>3</td>
<td>10 words</td>
<td>oral</td>
<td>once</td>
<td>no</td>
</tr>
<tr>
<td>Thomas (2010)</td>
<td>27</td>
<td>10-11 syllables</td>
<td>taped</td>
<td>once</td>
<td>no</td>
</tr>
<tr>
<td>Burger &amp; Chretien (2001)</td>
<td>14</td>
<td>12-15 syllables</td>
<td>taped</td>
<td>once</td>
<td>no</td>
</tr>
<tr>
<td>Eisenstein, et al. (1982)</td>
<td>16</td>
<td>15 syllables</td>
<td>missing</td>
<td>once</td>
<td>no</td>
</tr>
<tr>
<td>Munnich et al. (1994)</td>
<td>24</td>
<td>15 syllables</td>
<td>oral &amp; taped</td>
<td>once</td>
<td>yes</td>
</tr>
<tr>
<td>Myles (1996)</td>
<td>20</td>
<td>15-20 syllables</td>
<td>oral</td>
<td>once</td>
<td>no</td>
</tr>
<tr>
<td>Wang (1998)</td>
<td>10</td>
<td>16-30 syllables</td>
<td>oral</td>
<td>twice</td>
<td>no</td>
</tr>
<tr>
<td>Henning (1983)</td>
<td>90</td>
<td>2-10 words</td>
<td>taped</td>
<td>once</td>
<td>no</td>
</tr>
<tr>
<td>Perkins et al. (1986)</td>
<td>18</td>
<td>3-18 syllables</td>
<td>taped</td>
<td>twice</td>
<td>no</td>
</tr>
<tr>
<td>Daugherity (2008)</td>
<td>15</td>
<td>4-17 words</td>
<td>taped</td>
<td>twice</td>
<td>no</td>
</tr>
<tr>
<td>Fiori-Agoren (2004)</td>
<td>30</td>
<td>5-11 words</td>
<td>taped</td>
<td>once</td>
<td>no</td>
</tr>
<tr>
<td>Batmanian et al. (2008)</td>
<td>56</td>
<td>6-9 syllables</td>
<td>taped</td>
<td>once</td>
<td>no</td>
</tr>
<tr>
<td>Ortega (2000)</td>
<td>30</td>
<td>7-19 syllables</td>
<td>taped</td>
<td>once</td>
<td>no</td>
</tr>
<tr>
<td>Thomas (1991)</td>
<td>6</td>
<td>missing</td>
<td>missing</td>
<td>once</td>
<td>no</td>
</tr>
<tr>
<td>Van Boxtel et al. (2003, 2005)</td>
<td>20</td>
<td>missing</td>
<td>taped</td>
<td>once</td>
<td>yes</td>
</tr>
<tr>
<td>Li (2010)</td>
<td>23</td>
<td>missing</td>
<td>taped</td>
<td>once</td>
<td>yes</td>
</tr>
<tr>
<td>Roebuck et al. (1999)</td>
<td>23</td>
<td>missing</td>
<td>taped</td>
<td>twice</td>
<td>yes</td>
</tr>
<tr>
<td>Ellis et al. (2006)</td>
<td>36</td>
<td>missing</td>
<td>taped</td>
<td>once</td>
<td>yes</td>
</tr>
<tr>
<td>Hameyer 1980</td>
<td>129</td>
<td>varying</td>
<td>taped</td>
<td>twice</td>
<td>no</td>
</tr>
</tbody>
</table>
Length of the cue sentences is also a focus of interest when designing EI tasks. Most studies were aware of the importance of this design factor and took control of the length of the cue sentences. The majority of the studies ($n = 14, 70\%$) reported the length of the sentences either in the part of description of instrument or by providing an appendix with the EI sentences. Five studies (25\%) failed to report the information. The researcher in one study (5\%) apparently was aware of the factor, but only reported the sentence lengths as varying. Of the 14 studies reporting the length, 11 contained sentences with varying length and 3 used sentences with fixed length. The unit of sentence length the studies chose to report was also different. Ten studies choose to report syllables as sentence length, and four report the number of words in a sentence. For the studies using syllables to report the length, the maximum syllables in a sentence is 30 and the minimum syllables is 3. For the word length, the maximum is 17 words in a sentence and the minimum is 2 words in a sentence.

There are two types of EI tasks, immediate and delayed. Eight studies (40\%) employed immediate EI tasks. The participants in the studies were required to repeat the sentence immediately after they heard the stimulus. Six (30\%) employed certain kinds of intervention between the stimulus sentence and the repetition to avoid any effects of rote memory. Still another 6 out of 20 studies (30\%) did not report this feature and leave readers in the dark about the intervention. Among the six studies which used delayed EI tasks, three required the participants to indicate whether the statement they just heard is true or false or whether they agree or disagree with the statement. The purpose of this requirement is to ensure that the participants process the sentences for comprehension first so as to avoid mechanic rote echoic repetition. Two out of the six studies used
articulatory suppression task to minimize the effect of rote memory. Specifically, in the study by Thomas (2010), after the stimulus sentence the numbers one to ten appeared on the screen for four seconds and the learner had to count in the target language. Burger and Chretien (2001) provided a pause between the stimulus and the repetition and, during the pause, a voice counted backwards to prevent students from rehearsing the sentence.

Few studies ($n = 3, 15\%$) controlled the time allowed for the participants to repeat a sentence. Three seconds were allowed for the participants to repeat sentences with 5 to 11 words; ten seconds were provided for the participants to repeat sentences with 15 syllables; and 20 seconds for participants to repeat sentences with 3 to 18 syllables. One study ($5\%$) (Roebuck et al., 1999) indicated that a short pause was provided after each sentence, but did not provide specific information on the short pause. Four studies ($20\%$) allowed the participants to take as long as they need, including one study indicate that the time given was more than enough. Most studies ($n = 12, 60\%$) did not see controlling the response time as an important factor in EI design, and therefore they did not report whether they made effort to control the time for the repetition (see Table 4.6).
<table>
<thead>
<tr>
<th>Studies</th>
<th>Pause before repetition</th>
<th>Intervention methods</th>
<th>Target grammar</th>
<th>Other measures</th>
<th>Contextual support</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ortega (2000)</td>
<td>yes</td>
<td>pause</td>
<td>no</td>
<td>grammar judgment production</td>
<td>none</td>
</tr>
<tr>
<td>Daugherity (2008)</td>
<td>missing</td>
<td>no</td>
<td>no</td>
<td>IELTS</td>
<td>story pictures</td>
</tr>
<tr>
<td>Perkins et al. (1986)</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>none</td>
<td>no</td>
</tr>
<tr>
<td>Henning (1983)</td>
<td>missing</td>
<td>no</td>
<td>no</td>
<td>production task</td>
<td>partial sentence repetition</td>
</tr>
<tr>
<td>Flynn (1986)</td>
<td>missing</td>
<td>yes</td>
<td>na</td>
<td>act out of sentences</td>
<td>none</td>
</tr>
<tr>
<td>Munnich et al. (1994)</td>
<td>no</td>
<td>yes</td>
<td>na</td>
<td>grammar judgment</td>
<td>none</td>
</tr>
<tr>
<td>Van Boxtel et al. (2005)</td>
<td>no</td>
<td>yes</td>
<td>na</td>
<td>grammar judgment</td>
<td>none</td>
</tr>
<tr>
<td>Li (2010)</td>
<td>yes</td>
<td>belief activity</td>
<td>yes</td>
<td>grammar judgment</td>
<td>vocab list</td>
</tr>
<tr>
<td>Ellis et al. (2006)</td>
<td>yes</td>
<td>belief activity</td>
<td>yes</td>
<td>grammar judgment</td>
<td>none</td>
</tr>
<tr>
<td>Fiori-Agoren (2004)</td>
<td>no</td>
<td>yes</td>
<td>na</td>
<td>grammar judgment</td>
<td>none</td>
</tr>
<tr>
<td>Thomas (1991)</td>
<td>missing</td>
<td>yes</td>
<td>na</td>
<td>metalinguistic knowledge test</td>
<td>none</td>
</tr>
<tr>
<td>Myles (1996)</td>
<td>no</td>
<td>yes</td>
<td>na</td>
<td>metalinguistic knowledge test</td>
<td>none</td>
</tr>
<tr>
<td>Batmanian et al. (2008)</td>
<td>no</td>
<td>yes</td>
<td>na</td>
<td>none</td>
<td>none</td>
</tr>
</tbody>
</table>
Sixteen out of 20 unique sample studies (80%) chose to test certain grammar structures in the EI tests. Four (20%) did not target specific grammar structures and relied on the global competence to finish the EI tasks. As part of the procedure of taking the EI tasks, three studies (15%) provided contextual support to the participants. In the only study to target L2 Chinese, Li (2010) provided annotation for some key words in each sentence. The annotation included the Chinese character, pinyin, and an English explanation. In the study by Thomas (2010), when learners repeated a sentence, a cue picture would appear on the screen as a help to recall the content. Daugherity (2008) used multiple methods to support the repetition. The sentences in the task formed a complete story and the story is illustrated by pictures. In addition, Daugherity (2008) also provided partial scripts for longer sentences and the participants only needed to fill in certain words that had been blanked out.

The majority of studies (n = 15, 75%) included a measure or measures other than EI to test the participants’ competence. Among the additional measures used in these studies, grammaticality judgments and oral production (n = 7, 35%) are the two mainly used techniques. Some studies employed one of them, and two employed both measures.

<table>
<thead>
<tr>
<th>Studies</th>
<th>Pause before repetition</th>
<th>Intervention methods</th>
<th>Target grammar</th>
<th>Other measures</th>
<th>Contextual support</th>
</tr>
</thead>
<tbody>
<tr>
<td>Roebuck et al. (1999)</td>
<td>no</td>
<td>yes</td>
<td>none</td>
<td>none</td>
<td>none</td>
</tr>
<tr>
<td>Hameyer (1980)</td>
<td>no</td>
<td>yes</td>
<td>none</td>
<td>none</td>
<td>none</td>
</tr>
<tr>
<td>Wang (1998)</td>
<td>missing</td>
<td>yes</td>
<td>production task</td>
<td>none</td>
<td>picture</td>
</tr>
<tr>
<td>Thomas (2010)</td>
<td>yes</td>
<td>articulatory suppression</td>
<td>yes</td>
<td>production task</td>
<td>picture</td>
</tr>
</tbody>
</table>
Metalinguistic knowledge tests were also used in three studies (15%) along with EI in order to test learners’ competence. One study (5%) (Daugherity, 2008) employed a standardized test, IELTS, to compare the two measures. One study (5%) (Flynn, 1986) chose to use an act-out task to measure comprehension competence, in addition to the production competence measured by EI task. Still there are five studies (25%) which did not use any additional measures other than the EI task.

The two qualitative studies also included different design features in EI tasks. Martohardjono and Flynn (1995) failed to report how many items were used in the EI task, but provided the information that all the sentences in the EI task were 15 syllables long. Kaplan (1993) used 16 items in the EI task and each statement varied from one to nine words. Both studies presented each stimulus sentence once, but the former was administrated through taped recording and the latter was orally presented in real time. Kaplan (1993) included ungrammatical sentences and did not instruct the participants to correct the sentences. Both studies failed to provide information regarding whether repetition was immediate or delayed and neither reported the time allowed for repetition. Both studies included specific grammar features in the design. Kaplan (1993) used a grammaticality judgment task in addition to the EI task. Finally, Martohardjono and Flynn (1995) gave a vocabulary list to assist the participants to repeat the sentences.

### 4.4.3 Scoring Procedures

The primary studies employed different scoring procedures. Six (30%) unique samples used a binary scoring procedure to grade the repetition samples collected from the participants. They used either correct or incorrect, or zero to one point, to indicate the
results of the repetition. Among the six studies, five (25%) only graded the target grammatical forms and ignored other lexical items or phonetic features. One (5%) (Perkins, Brutten, & Angelis, 1986) decided that one point is given only when each word is correctly repeated in a sentence. Six other studies (30%) employed a polytomous scoring system. The biggest range is from zero to 38 and the smallest range is from zero to three for every repetition. Among these studies, two (10%) scored the repetition samples based on whether the participant repeat sentences correctly and maintain whole or partial meaning of the sentence. One study (5%) only graded based on whether the sentence was correctly repeated lexically. One study (5%) graded based on the correctness of the target structure. Two studies (10%) looked at multiple aspects of language performance. Burger and Chretien (2001) gave points for an exact repetition and semantic equivalent, and additionally graded certain pronunciation, syntax, and discourse features for a maximum of 38 points for each sentence. Henning (1983) graded on fluency, pronunciation, and grammar on a scale of 0-9 for each repetition. Seven studies (35%) counted the frequency of either errors of the target form, correct target forms, or subjects with each group who consistently use the target form (Thomas, 1991). One study (5%) failed to report the scoring procedure in the paper.

4.4.4 EI Reliability

Even though 20 unique samples using EI as an instrument to investigate the research questions in their studies, very few \( n = 5, 25\% \) actually calculated and reported the reliability of the EI test. Two studies employed Kuder-Richardson 20 to report the reliability of the EI tasks. The remaining three studies used Cronbach Alpha to calculate.
the EI reliability. The reported reliability coefficients ranged from .778 to .96 (See Table 4.7).

<table>
<thead>
<tr>
<th>Studies</th>
<th>EI reliability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ellis (2005, 2006) &amp; Erlam 2006</td>
<td>Cronbach Alpha .88</td>
</tr>
<tr>
<td>Ellis et al. (2006)</td>
<td>Cronbach Alpha .779</td>
</tr>
<tr>
<td>Henning (1983)</td>
<td>K-R 20 .96</td>
</tr>
<tr>
<td>Ortega (2000)</td>
<td>Cronbach Alpha .95 (pre EI)</td>
</tr>
<tr>
<td></td>
<td>Cronbach Alpha .96 (post EI)</td>
</tr>
<tr>
<td>Perkins et al. (1986)</td>
<td>K-R 20 .778</td>
</tr>
</tbody>
</table>

### 4.4.5 Correlation of EI with Other Measures

As mentioned above, there are 15 studies which used a measure or measures in addition to EI. Among the 15 studies, six unique samples reported the correlation between EI and one other measures (Table 4.8). Table 4.8 shows that majority study samples achieved a correlation higher than $r = .5$ with the highest correlation of $r = .87$ between EI and Simulated Oral Proficiency Interviews (SOPI; see Stansfield & Kenyon, 1992), and the lowest correlation of $r = .17$ between EI and an oral production task on the target structure of “is”. The independent sample by Ellis (2006) and Erlam (2006) indicated a low correlation ($r = .28$) between EI and Metalinguistic knowledge test. The low correlation actually supported their research hypothesis that the correlations between EI and untimed Grammaticality Judgment test or with metalinguistic knowledge would be weak because the former is supposed to tap implicit but the latter two tap explicit knowledge. And the correlations between EI and timed Grammaticality Judgment test would be strong because they both were posited to tap implicit knowledge. Eisenstein et al. (1982) reported relatively low correlation between EI and the oral production test. They further examined the participants’ performance on the oral production test and
found that the results of the oral production “were quite erratic at early stages of learning” (p. 388) and learners tended to avoid use of certain structures.

Table 4.8  
Correlation between EI and pairing measure(s)

<table>
<thead>
<tr>
<th>Studies</th>
<th>Pairing measure</th>
<th>N</th>
<th>r</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SOPI</td>
<td></td>
<td>.87</td>
</tr>
<tr>
<td>Daugherity (2008)</td>
<td>IELTS listening</td>
<td>25</td>
<td>.52</td>
</tr>
<tr>
<td></td>
<td>Timed GJT</td>
<td></td>
<td>.58</td>
</tr>
<tr>
<td></td>
<td>Untimed GJT</td>
<td></td>
<td>.59</td>
</tr>
<tr>
<td></td>
<td>Metalinguistic knowledge</td>
<td></td>
<td>.28</td>
</tr>
<tr>
<td>Burger &amp; Chretien (2001)</td>
<td>Pre-oral discussion task</td>
<td>31</td>
<td>.5</td>
</tr>
<tr>
<td></td>
<td>Post-oral discussion task</td>
<td></td>
<td>.85</td>
</tr>
<tr>
<td>Eisenstein, et al. (1982)</td>
<td>oral production task on &quot;does&quot;</td>
<td>45</td>
<td>.28</td>
</tr>
<tr>
<td></td>
<td>oral production task on &quot;do&quot;</td>
<td></td>
<td>.28</td>
</tr>
<tr>
<td></td>
<td>oral production task on &quot;is&quot;</td>
<td></td>
<td>.17</td>
</tr>
<tr>
<td></td>
<td>oral production task on &quot;are&quot;</td>
<td></td>
<td>.23</td>
</tr>
<tr>
<td>Wang (1998)</td>
<td>guided oral interview</td>
<td>30</td>
<td>.45</td>
</tr>
</tbody>
</table>

4.5 Discussion

4.5.1 Participants’ Features

The primary studies are different in many aspects. In terms of participants’ features, the synthetic analysis shows that the n sizes of the majority studies are not big. Most studies reported the proficiency levels of the participants, which is important for readers to grasp the whole picture of the study and be able to replicate it in the future. The institutional status is the most commonly used method to assess participants’ proficiency levels because it is direct and convenient. However, different institutions design different curriculum for language courses, which can lead to unclear definition of the levels and leave readers in the dark as to how to really replicate the study accurately. Thomas (2006) suggested multiple assessments to interpret participants’ levels reliably.
Few studies have followed the suggestion and used multiple assessments. Future studies need to employ more than one assessment to clearly define the proficiency levels.

Almost all the studies in the present synthesis were conducted in colleges. This is because all the studies with children as participants were excluded from the synthesis study. English remains the most popularly studied language in the field of second language studies, followed by French and Spanish. Only one study targeted at Chinese language (Li, 2010). As Chinese is a growing language program in the United States, it is necessary and important to invite more studies targeting at Chinese language and provide more pedagogical implication and suggestions to the language teaching.

There are also different features in EI task design. As shown above, even though most studies tape-recorded the stimuli sentences and administered the test via playing the recording, there are still few studies chose to orally present the sentences. Munnich et al. (1994) compared the recorded and oral stimuli EI tasks and concluded there were no significant differences between these two types of EI administration. However, one study is not enough to draw a conclusion on this issue. When an experimenter orally presents a sentence many times to different participants, it is hard to maintain the same delivery speed, intonation, and clarity of articulation. As a result, all these factors may influence the performance of participants and may reduce the reliability of the task, and this may be particularly true for studies with large sample sizes. Some studies presented a cue sentence twice before requesting participants to repeat it. This allows extra time for learners to access their explicit knowledge and therefore such an EI task measures may include more prescriptive knowledge.
4.5.2 Most Important Areas for the Optimal Design of EI Tasks

Jessop, Suzuki, and Tomita (2007) proposed that EI is constructive and consists of three processes: "(a) processing a stimulus sentence, (b) reconstructing it with their own grammar, and (c) reproducing it" (p. 215). In order to make sure that language learners are reconstructing the heard sentence with their own grammar, rather than repeating it mechanically from memory, some studies employed ungrammatical sentences to monitor the process and investigate the reconstructive nature of EI tasks directly. Munnich et al. (1994) found that the participants consistently corrected the ungrammatical sentences of a certain structure of interest in the study (p. 235). They concluded that the results indicated the process of reconstruction and shows EI is not merely rote memory. They further pointed out that EI can be reliably used in both grammatical and ungrammatical sentence structures (p. 237). Erlam (2006) compared the performances of native speakers and second language learners and found that native speakers corrected 91% of ungrammatical sentences, which proves the reconstruction nature of the EI tasks. Further, she also reported that L2 learners did correct 35% of ungrammatical sentences, even though they were not explicitly told there were ungrammatical sentences in the task. Bley-Vroman and Chaudron (1994) pointed out that when “other things being equal, it appears easier to imitate grammatical sentences than ungrammatical strings of words” (p. 247). They further said that “it appears that grammatical processing somehow aids repetition” (p. 247). Their observation exactly supports the idea that repetition task is a reconstructive measure, but it also underscores the importance of considering ways of avoiding rote memory effects when designing good EI tasks. Major areas for concern when designing good EI tasks therefore include (a) rote memory effects, (b) serial position effects, and (c)
ceiling and floor effects (Bley-Vroman & Chaudron, 1994). The remainder of this section will discuss the caveats raised by Bley-Vroman and Chaudron and will apply them to the primary studies analyzed.

4.5.2.1 Rote memory effects. How to use length of cue sentences to minimize any effects of rote memory is the focus of some research. The concern is that language learners may repeat a sentence from their rote memory without processing its meaning, unless the sentence is of some minimal length. In order to be able to repeat from rote memory alone, the language learner needs to store the information in short-term memory without retrieving data from his or her long-term memory. A learner may be able to do this within the limits of his or her working memory capacity. Naiman (1974) argues that when the cue sentence is with sufficient length, correct imitation cannot be achieved merely by rote memorization, but will require the involvement of comprehension and encoding of the structure (that is, repetition will be reconstructive). Miller (1956) pointed out that a person’s immediate memory span is “usually somewhat in the neighborhood of seven” units of information (p. 90). Perkins et al. (1986) also confirmed that seven to eight syllables was the threshold for their ESL participants. Kormos and Safar (2008) found that the participants were able to repeat 7.18 syllables. Hameyer (1980), in order to estimate the effect of rote memory, recruited three students who had zero knowledge of German and administered the EI test to them. The result shows that their best effort did not exceed 4.56 words per correctly imitated sentence. Therefore, a sentence consisting of around seven syllables may reach the limit of a person's capacity for processing L2 information in short term memory. As a result, researchers can control the number of
sylables in sentences used in an elicited imitation task if they want to avoid the possibility of rote memory.

Bley-Vroman and Chaudron (1994) argued that seven is the number of chunks and language learners can process the seven chunks within their rote memory, which may contain far more than seven syllables. In fact, Miller (1956) did point out that seven is the number of chunks that a person can process within the immediate memory capacity. However the number of information contained in each chunk may vary (p. 93). For example, “bed” can be considered a word and a chunk, but “bed time” may also well be one chunk, even though it made up of two words, for many fluent speakers of English. Therefore, a proficient language learner may be able to organize bits of information into chunks or varying sizes. The higher the proficient level is, the more information may be organized and packed into each chunk. As a result, the number of information contained in a chunk is correlated with proficiency levels, which is also supported by the argument of Munnich and his colleagues (Munnich et al., 1994). That is, the length of a sentence that can be repeated by sheer rote memory may increase as proficiency increases, and longer sentence stimuli will be needed to probe the upper limits of proficiency among high-proficiency and advanced learners. It is perhaps for this reason that many of the EI studies adopt stimulus sentences of varying length (11 of 14 studies reporting this information, as noted earlier).

A related issue is that among the primary studies, some reported the sentence length in the unit of words, and some reported as syllables. Based on Miller’s (1956) explanation on the chunk, syllable may be a better choice because it is a smaller unit than words and may organize together as a chunk. Bley-Vroman and Chaudron (1994)
presented their observation that the longer a sentence is, the less likely it is to be accurately repeated. This can well be explained by Miller’s theory (1956). The primary studies mostly contains more or less 10 syllables, which may well above the immediate memory span of seven and be able to distinguish certain proficiency levels. Bley-Vroman and Chaudron (1994) explained the process of EI tasks as when the learner gets the message from the stimulus and he or she will “try to run the message through the primitive grammar, and the result will be inaccurate” (p. 251). They also pointed out that “a slightly more advanced learner may be able to reduce the stimulus to a manageable number of chunks to fit into memory” (p. 251). On the other hand, two or three syllables in a sentence may be too short for a language learner that they can repeat by using rote memory. Therefore, in an EI task, seven may be a good number to start as minimum syllables in a sentence. In addition, in order to distinguish language learners at different proficiency levels, an EI task need to contain sentences with different number of syllables.

It is also in order to minimize the effect of rote memory that an intervention between the stimulus and the repetition was employed in some studies. McDade, Simpson, and Lamb (1982) found that language learners could not repeat a sentence they did not understand if there was a three-second interval between the stimulus and the imitation. Erlam (2006) supported McDade et al. by saying that if there is a time interval between the presentation of the stimulus and the elicited response, it would appear that memory is reconstructive instead of relying simply on rote repetition (p. 469). Li (2010), Ellis (2005, 2006, Ellis et al., 2006), and Erlam (2006) requested the language learners to indicate whether or not they agree with the statement to make sure they process the comprehension first and erase the possibility of rote memory. Burger and Chretien (2001)
and Thomas (2010) employed articulatory suppression to prevent participants from rehearsal phonetically in their mind. Ortega (2000) just provided a short pause to delay the repetition.

All these are possible methods to delay the imitation and minimize the effect of rote memory. However, each method has its advantages and disadvantages. A meaning-oriented distraction task between stimulus and response may increase the memory burden by adding more information and taking away attention, and reduce the accuracy of the measure of EI. A pause may generate the possibility that the learners are mentally rehearsing the sentences. Further studies are to be called on to investigate the topic.

As reported above, only three studies controlled time allowed for participants to repeat a sentence. Limited time helps to ensure that participants use online processing and do not have time to access their metalinguistic knowledge. However, how much time is appropriate for participants to repeat a sentence is still a question to be answered. Fiori-Agoren (2004) left three seconds for the participants to repeat sentences ranging from 5 to 11 words, which easily raised a question from readers whether three seconds is even enough for a native speaker to say a sentence of 11 words. Perkins et al. (1986) allowed 20 seconds for learners to repeat sentences from 3 up to 18 syllables. Twenty seconds may be more than enough for 3 syllables, but pressing for 18 syllables. Therefore, in order to better monitor the online process of repetition, it is a good idea to vary the limited time based on the number of syllables that each sentence contains.

4.5.2.2 Serial position effects. Bley-Vroman and Chaudron (1994) pointed out that the last word in a sentence is usually best repeated, followed by the first word in a sentence. The serial position effect may affect validity of an elicited imitation which is
designed to measure language learners' competence of a certain grammar feature, if the feature often appears at the end or beginning of the sentence, because it may be easier to repeat it from memory than if it appears in sentence-internal position. The present synthesis shows that the majority of the primary studies to date have targeted certain grammar structures. Twelve studies graded the repetition purely on the target grammar structures. In these cases, the serial position of the target structure in the sentence stimuli may have affected the validity of the EI tasks in unknown ways. Future researchers interested in using EI tasks for the purpose of measuring knowledge of certain grammatical subsystems need to take the positions of target structure into consideration when designing the tasks.

On the other hand, elicited imitation measuring global competence of language learning is not affected by the serial position. Bley-Vroman and Chaudron (1994) admitted that “a broad sampling of stimuli of various lengths and complexities should provide a reasonably good assessment of global proficiency” (p. 252). They also pointed out that “The more you know of a foreign language, the better you can imitate the sentences of the language. Thus, EI is a reasonable measure of global proficiency” (p. 247). Bley-Vroman and Chaudron (1994) summarized that EI involves (a) the speech comprehension system, (b) representation, (c) memory, (d) the speech production system. This, from another perspective, proves that EI is a good indicator for global proficiency since it requires the ability to comprehend and produce the L2 in real-time terms and holistically.

**4.5.2.3 Ceiling and floor effects.** Another concern for optimal EI design is that of ceiling effects that may occur when the sentence is so easy that almost all the language
learners can repeat it correctly, and conversely floor effects that may occur when the sentence is too hard to be successfully repeated by almost all the learners. Both ceiling and floor effects put a big question mark on the validity of elicited imitation task design that does not account for these issues, since its validity would be lost if the task failed to discriminate proficiency levels of different language learners. A carefully controlled design and a pilot study can be effective in addressing and avoiding ceiling and floor effects. In addition, an EI task containing sentences with varying difficulties may reduce the danger of ceiling and floor effects as well.

Another problem that Bley-Vroman and Chaudron (1994) pointed out is that sometimes, language learners do not accurately repeat a sentence; instead they may substitute some words to express the same meaning (p. 247). This fact needs to be taken consideration into the scoring system. It is important to establish a scoring rubric to distinguish learners who substitute some words but remain faithful to the intact meaning of the stimulus, from those who can perfectly repeat the sentence, and those who cannot repeat in a complete meaning. These differential performance abilities may reflect different degrees of global competence in the L2.

4.5.3 Reliability and Validity of EI Tasks

Based on the limited studies which reported the reliability of the EI tasks used, the three reliabilities of scores by Cronbach Alpha and two by Kuder-Richardson 20 derived from EI tasks seem to be relatively high. This indicates that EI is overall a reliable instrument. Munnich et al. (1994) did not report the reliability, but stated that EI can be reliably used in both grammatical and ungrammatical sentences. It is also recommended
that future studies report the reliability of an instrument to provide more information to readers.

Some studies calculated the correlation between EI and other measures used in the same study. EI correlated with grammaticality judgment tasks very well, but poorly correlated with a metalinguistic knowledge test. This is in fact a good sign suggesting construct validity, given the fact that metalinguistic knowledge tests aim at tapping what learners know about the language instead of tapping how well learners can actually use the language. EI tasks also correlated well with standardized tests, SOPI, and IELTS listening, as well as most oral production tasks. Daugherity (2008) concluded that the using a repetition test to measure basic language ability has been verified as valid. Although Munnich et al. (1994) did not provide the reliability of EI and grammaticality judgment task, they nevertheless speculated that with respect to grammatical sentences, the two tasks provide strikingly convergent results (p. 237). However, the study by Eisenstein et al. (1982) demonstrated a relatively low correlation between EI and oral production tasks on the four grammar features examined. As mentioned earlier, they closely examined the two tests and explained that it might be caused by problems in the production task. They pointed out that “the results of cued production in our study were quite erratic at early stages of learning. Amount of use was severely limited, and this limited use was further restricted by heavy reliance on prefabricated routines or a few high frequency verbs” (p. 388).

Daugherity (2008) also examined the correlation between the length of time in the United States and the scores on the repetition test. She found a fairly high correlation \((r = .67)\) between them and concluded that:
It is notable that people who have been in the U. S. a long time and who are known to have a near-native conversational ability score very highly. No really surprising scores, based on length of exposure to English, are seen in these results. This fact points toward the validity of the Repetition Test as a predictor of general English ability. (p. 116)

Henning (1983) compared three methods in oral proficiency test: interview, imitation, and completion. He reported highest overall validity for the imitation method in general. Gallimore and Tharp (1981), a study not included from the present synthesis due to its non-adult participants, administered EI tests of Standard English and Hawaiian Creole English twice a year to elementary students during five years. The results also indicated that EI tests yields “highly stable, test-retest correlation over a period of years (p. 378).

4.6 Conclusion

Based on the above discussion, it is clear that not only the studies with EI tasks, but also the studies in the field of second language acquisition need to shift more attention to less commonly-taught languages. The studies also need to provide more information on features of participants, especially their proficiency levels and methods to assess their levels. As for the studies with EI tasks, more detailed description regarding the administration and design of the tasks are needed for readers to get the whole picture and compare among studies.

Bley-Vroman and Chaudron (1994) raised three major areas in need of attention when designing good EI tasks, particularly for the purpose of measuring global oral
proficiency. Control the length of cues sentences is a good strategy to minimize the danger that a repetition task may become just a rote memory repetition task. A sentence exceeding seven syllables is likely exceeding the capacity of a person’s immediate memory span if he or she lacks the competence to combine syllables into chunks. In addition, the sentences need to be varying in length to provide a better measure for participants at various levels of proficiency. A delayed EI task can also reduce the effects of rote memory. However, what kind of intervention between stimulus and repetition best serves the purpose and will not increase the memory burden of the participants needs to be explored. The time allowed for participants to repeat a sentence should vary based on the length of the sentence stimulus, to correlate with learners’ proficiency levels. If EI is designed to measure a certain grammar structure and scoring is solely based on whether or not the participant can repeat the structure correctly, caution should be taken to avoid serial position effects. Sentences with various difficulties can also minimize ceiling and floor effects. The synthesis also suggests an optimal practice for EI administration is for researchers to tape recorded the stimuli sentences in order to provide a consistent input for the participants and to avoid the influences of possible extraneous variables caused by change in delivery speed or intonation.

In summary, EI overall is a reliable and valid instrument, especially in measuring global competence of a language. Jessop, Suzuki, and Tomita, (2007, p. 216) summarized that in an EI task, a wide range of L2 structures can be elicited relatively easily, including syntax, pronunciation, and discourse makers. Researchers have a degree of control over administration and analysis with the flexible and reliable processes of EI. EI can also be used with different age groups, different languages, and different populations. In the
present dissertation study, therefore, a carefully designed and administered EI task was
developed and pilot tested and then employed in the main study in order to measure
global oral proficiency in Mandarin Chinese, so as to then investigate the relationship
between WTC to L2 proficiency as a potential outcome mediated by the outcome of L2
frequency of use.
CHAPTER 5
RESEARCH DESIGN AND METHOD

The present research investigates the relation between L2 WTC and its affective factors, including communication anxiety and perceived competence. It also examines the causal relation between L2 WTC and the learners’ actual proficiency levels and frequency of L2 use. In addition, the differences of L2 WTC and its related affective factors between heritage and non-heritage language learners of Chinese are foci of the present study as well. The participants were recruited from five universities in the United States. The data for all five variables involved were collected from two instruments: a Mandarin EI task and a survey. The survey contains three major parts: background information, the affective factors in Mandarin learning, and frequency of L2 use. SPSS and SEM, Analysis of Moment Structures (AMOS, Arbuckle, 2011), are used to conduct analysis on the data.

5.1 Research Questions

The present study intends to seek answers for the questions related to the L2 WTC construct and differences arising between heritage and non-heritage language learners in terms of affective factors that may affect WTC differentially for the two populations of learners. It aims to find answers for the following research questions and provide pedagogical suggestions on increasing WTC in non-native language learning:

1. L 2 WTC Construct
   a. What is the relation of WTC with perceived competence and communication anxiety in L2?
b. What is the relation of L2 WTC with participants’ self-report frequency of L2 use?

c. What is the relation of WTC with participants’ proficiency levels in L2?

d. What are the differences in relation of L2 WTC with perceived competence and communication anxiety in varied individual L2 use contexts at home, at school outside the Mandarin classrooms and in the community?

2. Heritage and non-heritage language learners of Chinese

a. What is the difference of L2 WTC between heritage and non-heritage language learners of Chinese?

b. What are the differences of L2 perceived competence and L2 communication anxiety between heritage and non-heritage language learners of Chinese?

c. What are the differences of L2 frequency of use and actual proficiency levels between heritage and non-heritage language learners of Chinese?

5.2 Participants

The present research recruited 208 participants from five universities in the United States. The five universities are located at diverse parts of the United States: one each in Hawai‘i, Pennsylvania, and Wisconsin, and another two in California. All the participants are adult language learners of Mandarin at various proficiency levels, which were decided by their institutional levels and also measured by the elicited imitation task of Mandarin. As the data collection involved five universities and a large number of
participants were recruited, a few samples of non-English native speakers who were learning Mandarin were accidently included.

The initial sample of 208 participants was reduced to the final sample in the study of 179 participants who provided complete data sets. Table 5.1 summarizes the main reasons causing the loss of these 29 volunteers, which are also explained in detail here. First, in order to minimize the effects brought by linguistic and sociolinguistic influences from their distinct native languages, the research eliminated the participants who reported their strongest languages after the age of 5 being a language other than English. As a result, seven non-dominant English speakers, including three Vietnamese speakers, two Korean speakers, one Bosnian speaker, and one Italian speaker, were excluded from further analysis of the data. Second, the data collection procedure consisted of two parts and took approximately a total of 30 to 45 minutes, and some participants could not attend the full session of the data collection. Therefore, a number of students were allowed to either take the paper-based survey home and turn it in later or complete the survey online through Survey monkey. Unfortunately, this caused some uncollectable data due to no responses to any reminding emails sent by the researcher to complete the survey. Consequently, four participants were excluded because of incomplete surveys. Third, the researcher and the administrators collected data in person via recording the responses of participants in the EI task of Mandarin. One recording was cut off in the middle because of technical problems. Seven volunteers completed the online survey, but did not show up for the recording session. Ten participants from one of the universities completed all the procedures, but the data of the EI task were accidentally destroyed in the administrator’s computer. Therefore, unfortunately 18 participants were further
excluded from the sample pool because of incomplete or lost data of the EI task. In the end, altogether 179 participants were included in the data analyses for the dissertation (Table 5.1).

<table>
<thead>
<tr>
<th>Excluded reasons</th>
<th>n</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-English native speakers</td>
<td>7</td>
</tr>
<tr>
<td>Incomplete survey</td>
<td>4</td>
</tr>
<tr>
<td>Incomplete EI tasks</td>
<td>18</td>
</tr>
<tr>
<td>Total</td>
<td>29</td>
</tr>
</tbody>
</table>

5.3 Instruments

5.3.1 Design of Elicited Imitation Task

The present study uses EI to measure the global proficiency levels of English native speakers learning Mandarin as a foreign or heritage language. As a norm-referenced test, the Mandarin EI task aims to spread students out along a continuum of scores. The Mandarin EI task was developed as a parallel task closely following the stimuli created in EI tasks for English, German, Japanese, and Spanish developed by Ortega, Iwashita, Norris, and Rabie (2002). The internal consistency of the EI test scores in the original crosslinguistic study were .93 in English, .96 in German, .95 in Japanese, and .97 in Spanish. Perkins, Brutten, and Angelis (1986) reminded researchers that when creating parallel stimuli across languages, familiarity with the vocabulary, semantic plausibility, and schemata/expectations need to be taken into account. Bearing these principles in mind, my colleague Shu-Ling Wu and I each developed a parallel set of 30 EI sentence stimuli of Mandarin based on the 30 items of the original English EI task. We
then met and discussed the vocabulary, semantic plausibility, and appropriate structures in the task. We decided the final version of the 30 stimuli in Mandarin for the EI task upon discussion and mutual agreement. All the items were taped recorded with females’ voices for both English instruction and Mandarin stimuli (see the complete Mandarin EI task, including instructions and stimuli, in Appendix C; for the EI task scoring rubric, see Appendix D).

The present Mandarin EI task intends to measure the global competence of language learners of Mandarin. Therefore, any concerns regarding serial position is eliminated, since the learners’ performances are not judged based on the accuracy of certain grammar structures. On the other hand, in order to control the effects of rote memory and parallel the tasks in Ortega et al. (2002), the present EI task contains sentences varying from 7 to 19 syllables maximum. Based on previous research, seven is about the number of a person’s short term memory capacity (e.g., Miller, 1956). In addition, the task includes a wide range of sentence length to increase the ability of the EI to distinguish participants at different levels. Ortega (1997) drew a conclusion based on the data of 112 subjects in the studies by Scott (1989, 1994) that “(a) eight-syllable items discriminate well between L2 grammatical processing abilities and pure repetition strategies, (b) 10-syllable items are too taxing to repeat without accessing grammatical processing, (c) eight to 14-syllable items can be expected to be repeated with a minimum of 50% accuracy by mixed proficiency and mixed short-term memory capacity learners, and (d) most L2 learners reach their ceiling with 18-syllable items or longer” (p. 1).

A pause was inserted between the stimulus and the elicited repetition. A ring tone was placed at the end of the pause to signal the participants to start the repetition. The
pause between the end of each sentence and the onset of the tone lasts two seconds and the ring tone following it lasts 0.5 second. Therefore, the total delayed time before the sentence repetition begins is 2.5 seconds. Ortega, Iwashita, Rabie, and Norris (in preparation) explained that the decision of a 2.5-second interval was made based on the study by Chaudron and Russell (1990) that a 5-second delay may allow subjects to have chances to access their controlled processing strategies, such as rehearse the sentence in mind, and a three-second delay before repetition may be ideal in order to avoid mechanical repetition from short-term memory. The response time for participants is also carefully calculated based on the length of a sentence. It follows the natural pace of a native speaker repeating a sentence plus an extra delay to allow for the possibility that non-native speakers may hesitate, self-correct, or likely speak at a lower speech rate than native speakers. The length of the extra delay varies according to the number of the syllables that a sentence contains. Ortega et al. (in preparation) developed a method to calculate the response time for the participants. They first calculated the time taking a native speaker to speak the sentence, which is called \( t \). The shortest sentence in the EI task consists of seven syllables. Therefore, the allotted time for repetition of the sentence is \( t + 2 \) seconds. For sentences with each additional syllable, an additional 0.5 second was added to the \( t + 2 \) seconds time frame for the repetition. For example, for a sentence with eight syllables, the participants have \( t + 2.5 \) seconds to repeat the sentence, and for a sentence with nine syllables, \( t + 3 \) seconds were allocated for the repetition. The entire EI task lasts approximately 10 minutes.

The instructions require the participants to “try to repeat exactly what you hear” and “repeat as much as you can”. The instructions also remind the participants to start
repeating after the beep. The EI task starts with six L1 English sentences for participants to practice and make sure they do not start repetition before hearing the beep. When the participants finish the practice items, the recording informs the participants of the end of English sentences. The same instruction is given regarding Mandarin sentences. The Mandarin part includes 30 sentences from the minimum of 7 syllables to the maximum of 19 syllables. When all the sentences are finished, the participants will hear the instruction “This is the end of the repetition task. Thank you”. The entire EI was recorded on female voices on both English and Mandarin parts.

Ortega et al. (in preparation) developed a scoring rubric from zero to four points. The major distinction between each benchmark is as follows: (a) when there is a silence, unintelligible garbles, or only one word is repeated, zero points are awarded; (b) when half or barely half of lexical words and meaning are retained in the repetition, one point is assigned; (c) two points are assigned when at least more than half of the idea units are maintained; (d) when original and complete meaning is preserved, but some synonymous substitutions occurs without changing the meaning, three points are assigned; (e) four points are only given to perfect repetition, where both form and meaning are faithfully preserved and repeated. The complete scoring rubric with sample sentences is given in the Appendix D.

The Mandarin EI task was uploaded and converted into a test that can be taken and recorded on an online server by Richard Medina, an IT Specialist at the Language Lab of the college of Language, Linguistics, and Literature (LLL) at the University of Hawai‘i. An online tool, Web Audio Utility (WAU), was used. WAU is used as a common tool in language classes at this university. It has the features that other academic
tools have, such as blackboard and Laulima. Teachers and students can upload documents, post comments, and submit homework there. However, WAU also allows students to submit sound files as homework and teachers can provide feedback by recording another sound file. This instrument is ideal for language classes, especially for speaking classes, by providing additional chances for students and teachers to interact with each other outside classrooms. Therefore, WAU was chosen as the platform for the online EI task. Richard Medina uploaded the EI recording files and did some technical changes to make it more appropriate for a testing task. The students needed to log in by their assigned participant numbers and once they clicked the play button, the recording started to play and at the same time, it started to record students’ responses. The pause, forward, and backward functions were all disabled in order to keep the timing of the test standardized across individuals. Therefore, once the EI task started, it continued to play without any interruption till the end. The participants needed to follow the pace preset in the recording. After the recording, the files were saved in the online database. They could also be downloaded in an mp3 format.

5.3.2 Piloting the EI Task

Before using the EI task in the main dissertation study, a pilot study was conducted by Zhou and Wu (2009) in order to examine that the task was appropriate for the target population and that there were no floor or ceiling effects (Bley-Vroman & Chaudron, 1994), as well as to provide preliminary evidence about the reliability and validity of the Mandarin EI task. For the pilot, twenty-three learners of Mandarin were recruited from first and fourth year of Mandarin classes at the university at Hawai‘i. They
finished both the background information survey and the EI task. Their EI recordings were scored based on the 0-4 points’ rubric. Two raters underwent training before starting to rate the EI data. They first studied the scoring rubric together and discussed the benchmark for each point. Then they rated a few sample items separately and compared the ratings. Differences in the ratings were discussed until agreements were reached and benchmarks were clarified. After the training session, the two raters independently rated the remaining Mandarin EI data for all the participants. The inter-rater reliability of the final coding was high ($r = .985$) and significant at 99% level ($p < .01$). The internal reliability of the EI task scores, which was estimated with Cronbach’s alpha was also satisfactory: $\alpha = .968$.

As mentioned, the sentences in the task varied in length, vocabulary, and grammar structures in order to minimize any floor or ceiling effects at the highest and lowest points of the learner proficiency continuum (since the items needed to challenge the very wide range of proficiencies targeted during sampling) and to make sure the task elicited reconstructive repetition rather than mechanic rote memory parroting. Item Discrimination (ID) analysis was conducted to examine the validity of the Mandarin EI task (See Table 5.2).
Table 5.2
*Item discrimination analysis of Mandarin EI task*

<table>
<thead>
<tr>
<th>Item #</th>
<th>IF upper</th>
<th>IF low</th>
<th>ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.81</td>
<td>0.66</td>
<td>0.15</td>
</tr>
<tr>
<td>2</td>
<td>1.00</td>
<td>0.52</td>
<td>0.48</td>
</tr>
<tr>
<td>3</td>
<td>0.92</td>
<td>0.55</td>
<td>0.37</td>
</tr>
<tr>
<td>4</td>
<td>0.92</td>
<td>0.57</td>
<td>0.35</td>
</tr>
<tr>
<td>5</td>
<td>0.88</td>
<td>0.30</td>
<td>0.58</td>
</tr>
<tr>
<td>6</td>
<td>0.73</td>
<td>0.41</td>
<td>0.32</td>
</tr>
<tr>
<td>7</td>
<td>0.88</td>
<td>0.45</td>
<td>0.42</td>
</tr>
<tr>
<td>8</td>
<td>0.75</td>
<td>0.27</td>
<td>0.48</td>
</tr>
<tr>
<td>9</td>
<td>0.67</td>
<td>0.30</td>
<td>0.37</td>
</tr>
<tr>
<td>10</td>
<td>0.75</td>
<td>0.57</td>
<td>0.18</td>
</tr>
<tr>
<td>11</td>
<td>0.50</td>
<td>0.09</td>
<td>0.41</td>
</tr>
<tr>
<td>12</td>
<td>0.60</td>
<td>0.36</td>
<td>0.24</td>
</tr>
<tr>
<td>13</td>
<td>0.63</td>
<td>0.30</td>
<td>0.33</td>
</tr>
<tr>
<td>14</td>
<td>0.33</td>
<td>0.14</td>
<td>0.20</td>
</tr>
<tr>
<td>15</td>
<td>0.50</td>
<td>0.23</td>
<td>0.27</td>
</tr>
<tr>
<td>16</td>
<td>0.42</td>
<td>0.14</td>
<td>0.28</td>
</tr>
<tr>
<td>17</td>
<td>0.52</td>
<td>0.16</td>
<td>0.36</td>
</tr>
<tr>
<td>18</td>
<td>0.33</td>
<td>0.23</td>
<td>0.11</td>
</tr>
<tr>
<td>19</td>
<td>0.60</td>
<td>0.25</td>
<td>0.35</td>
</tr>
<tr>
<td>20</td>
<td>0.42</td>
<td>0.25</td>
<td>0.17</td>
</tr>
<tr>
<td>21</td>
<td>0.48</td>
<td>0.09</td>
<td>0.39</td>
</tr>
<tr>
<td>22</td>
<td>0.44</td>
<td>0.11</td>
<td>0.32</td>
</tr>
<tr>
<td>23</td>
<td>0.35</td>
<td>0.18</td>
<td>0.17</td>
</tr>
<tr>
<td>24</td>
<td>0.50</td>
<td>0.11</td>
<td>0.39</td>
</tr>
<tr>
<td>25</td>
<td>0.50</td>
<td>0.25</td>
<td>0.25</td>
</tr>
<tr>
<td>26</td>
<td>0.40</td>
<td>0.27</td>
<td>0.12</td>
</tr>
<tr>
<td>27</td>
<td>0.31</td>
<td>0.11</td>
<td>0.20</td>
</tr>
<tr>
<td>28</td>
<td>0.54</td>
<td>0.16</td>
<td>0.38</td>
</tr>
<tr>
<td>29</td>
<td>0.58</td>
<td>0.23</td>
<td>0.36</td>
</tr>
<tr>
<td>30</td>
<td>0.40</td>
<td>0.14</td>
<td>0.26</td>
</tr>
</tbody>
</table>
According to Brown (2005), items with ID of .30 and above are usually considered acceptable items and those at .09 or below are poor items and need to be improved by revision (p. 75). Table 5.2 shows six items (Item 1, 10, 18, 20, 23, and 26) may need to be improved because they near the .09 cut-off point. Seventeen items show an acceptable or better ID index, and the remaining 7 are close to the .30 cut-off point for acceptability. Overall, the ID analysis shows that the items in the Mandarin EI task are reasonably able to distinguish high and low level of Mandarin learners, and there are no items which are too difficult or too easy to generate the floor and ceiling effects.

The participants were divided into two groups of low and high proficiency based on their institutional status: enrolled in first and fourth year courses, respectively. Table 5.3 shows the descriptive data of EI scores of the two groups.

<table>
<thead>
<tr>
<th>Group</th>
<th>n</th>
<th>M</th>
<th>SD</th>
<th>Min.</th>
<th>Max.</th>
</tr>
</thead>
<tbody>
<tr>
<td>low</td>
<td>11</td>
<td>33.55</td>
<td>12.895</td>
<td>16</td>
<td>56</td>
</tr>
<tr>
<td>high</td>
<td>12</td>
<td>70.58</td>
<td>13.097</td>
<td>42</td>
<td>90</td>
</tr>
<tr>
<td>Total</td>
<td>23</td>
<td>52.87</td>
<td>22.786</td>
<td>16</td>
<td>90</td>
</tr>
</tbody>
</table>

*Note.* Total maximum possible score was 120.

The pilot participants’ background information was quantified through a system of points that were added together in order to measure the amount of second language contact that a participant have experienced. The present point-system was admittedly developed out of researcher intuition. However, there is some precedent for L2 researchers needing to obtain a single score for amount of language contact through questionnaire answers to devise a point-system (for a recent example, see van der Hoeven & de Bot, 2012). This approach, albeit pre-theoretical, seems defensible given that, as van der Hoeven and de Bot suggest, no consensus has been reached in the field of second
language acquisition about how best to operationalize and measure amount of L2 contact.

The survey scoring system for amount of language contact was based on the following criteria:

1. Participants’ strongest language before or after age five:
   - 2 points: Mandarin or a Chinese dialect
   - 0 point: other language

2. Family members speak Chinese at home:
   - 4 points: both parents and grandparents
   - 3.5 points: both parents and one grandparent
   - 3 points: both parents
   - 2.5 points: one parent and at least one grandparent
   - 2 points: one parent
   - 1 point: both grandparents
   - 0.5 point: one grandparent

3. Age to start to hear or use Mandarin Chinese:
   - 2 points: before five
   - 0 points: after five

4. Length of the Mandarin classes that a participant attended:
   - Mandarin classes with at least eight-hour class sessions every week
     - 2 points: per year
   - Mandarin classes with three to seven-hour class sessions every week
     - 1 point: per year
   - Mandarin classes with less than three-hour session every week
0.5 point: per year

5. Length of visiting or living in a Chinese-speaking country:
   1 point: per year

   0.5 point: per six months

   0 point: less than one month

6. Individual language contexts (How much a participant hear or use Mandarin outside classroom):
   2.5 points: almost always

   2 points: frequently

   1.5 point: sometimes

   1 point: occasionally

   0 point: never

Table 5.4 shows the descriptive statistics for the participants’ L2 contact. A t-test was conducted to compare whether there is a significant difference between the participants of high and low proficiency groups on their performances of EI and their language backgrounds in terms of amount of language contact. The results show that the two groups have significantly different L2 contact ($p < .05$) from each other and also perform significantly differently on EI task ($p < .05$) from each other (Table 5.5). A correlation was conducted between the participants’ language contact and EI performance. The results show the two variables are moderately correlated at $r = .566$ ($p < .05$).

Table 5.4

<table>
<thead>
<tr>
<th>Group</th>
<th>$n$</th>
<th>$M$</th>
<th>$SD$</th>
<th>Min.</th>
<th>Max.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low</td>
<td>11</td>
<td>7.85</td>
<td>5.190</td>
<td>3</td>
<td>19</td>
</tr>
<tr>
<td>High</td>
<td>12</td>
<td>17.02</td>
<td>8.936</td>
<td>9</td>
<td>39</td>
</tr>
<tr>
<td>Total</td>
<td>23</td>
<td>12.63</td>
<td>8.609</td>
<td>3</td>
<td>39</td>
</tr>
</tbody>
</table>
Table 5.5
High-low proficiency group comparison of EI and L2 contact

<table>
<thead>
<tr>
<th>Variables</th>
<th>$t$</th>
<th>$p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Elicited imitation</td>
<td>-6.825</td>
<td>.000</td>
</tr>
<tr>
<td>L2 contact</td>
<td>-2.972</td>
<td>.007</td>
</tr>
</tbody>
</table>

*Note. $p < .05.$

In summary, the pilot study shows the Mandarin EI is a reasonably reliable and valid instrument to measure participants’ global proficiency levels in learning Mandarin as a heritage/foreign language.

5.3.3 Survey

An online survey was built by using Survey Monkey. The survey comprised the following parts five parts: (1) background information, (2) L2 WTC, (3) L2 perceived competence, (4) L2 communication anxiety, and (5) frequency of L2 use in Mandarin learning. They are explained in detail here (and shown in full in Appendix E).

1) Part 1 elicited *Background Information* and is based on the background survey developed by Chu (2008) and Kondo-Brown (2008) (it is shown in full in Appendix E). This part starts from general biographic data, including gender and age. It continues to ask about the participants’ strongest language before and after the age of five, any family members speaking Mandarin or one of the Chinese dialects, and the age when they started to hear or use Mandarin Chinese. This information was used to classify participants into heritage and non-heritage learner groups. Participants are also asked to provide information regarding their previous experiences in Mandarin learning in school, including school location (to decide whether they were in a second or foreign language learning context.
in the past), length of study, and hours of Mandarin classes per week (to better illustrate the different curriculum designs in different schools). The survey also asks for their past traveling experience in regions where Chinese is spoken. Responses from all these items related to background information were quantified into the variable of amount of second language contact through the point-system just presented section 5.3.2. Finally, the last item in part 1 of the survey is a five-point Likert item that asks how much Chinese the participants currently use at home, at school outside the Mandarin classroom, in the community, and at other occasions. The responses in this last item of the first part of the survey regarding participants’ individual language use contexts are used as a variable in later data analysis to investigate whether use context makes any difference in the relationship of L2 perceived competence and L2 communication anxiety with L2 WTC.

2) Part 2 was the WTC in L2 Mandarin. MacIntyre’s (MacIntyre et al., 2001) WTC scale was adopted to measure WTC in L2. A five-point scale is used for the participants to indicate from 1, almost never willing, to 5, almost always willing. The MacIntyre’s scale consists of two parts: inside and outside the classroom. Each part contains 27 items referring to WTC in four skill areas: speaking, comprehension, reading, and writing. Each skill area contains five to eight items providing some scenarios which require the use of Mandarin.
3) Part 3 was *Perceived Competence in L2 Mandarin*, and it was adopted from the perceived competence scale by MacIntyre et al. (2002). The 0-100 scale is used to let the participants indicate the range from completely incompetent and completely competent. Twelve items are included in the measure and ask the participants to indicate how competently they believe they would be able to communicate in each of the situation, including with friends, with acquaintance, with strangers, and at different occasions of in public, in group, and with an individual.

4) Part 4 measured *Communication Anxiety in L2 Mandarin*. The items come from MacIntyre’s CA scale (MacIntyre & Charos, 1996) and are used to measure communication anxiety in three occasions and three types of people. The 12 items in this part are identical to those in the perceived competence measure. However, the instructions ask the participants to indicate on a 0-100 scale the degree of anxiety they might feel in each situation.

5) Part 5 of the survey targeted *Frequency of Language Use in L2 Mandarin*, for which the items in MacIntyre’s study (MacIntyre et al., 2002) were adopted. The 12 items are the same as the ones in perceived competence and communication anxiety. The participants are asked to indicate at a seven-point scale of how often they have engaged in each of the situations from 1, never, to 7, many, many times.

It is important to clarify how three of the above mentioned variables differed from each other and constituted distinct constructs in the design of the study: L2 contact,
individual language use context, and frequency of Mandarin use. The L2 contact responses (i.e., the background information about past Mandarin exposure, instruction, and so on, collected through several items in Part 1 of the survey) was used to examine the amount of language contact that a participant has experienced in the past, and it was employed in the pilot and main study EI study to examine the validity of the Mandarin EI task (the assumption being that the more L2 contact learners have experienced in their history of learning, the more proficient they should be). Furthermore, L2 use contact investigated both active or passive contact that a participant has experienced in the past, for example, a participant may have been exposed to passive contact with Chinese language and culture when he or she was traveling in a Chinese-speaking region, but the participant might not necessarily actually have used Chinese as often. On the other hand, individual language use contexts (elicited in the last question of Part 1 of the survey) focused on current language use at the time of the data collection, independently from the contact that a participant experienced before, such as previous formal instruction or traveling experiences. Additionally, it targeted how often learners’ use the language productively in their present life in four general use contexts: at school outside the Mandarin classrooms, in the community, at home, and others. The answers were used to measure whether each language learner was in a high or low L2 use context at the time of the study. Finally, the frequency of language use in L2 Mandarin (Part 5 of the survey) investigated self-reported information about the current actual productive use of Mandarin of a participant with three specific types of interlocutors (friends, acquaintance, and strangers) and in three different types of occasion (in public, in group, and with an individual), all of which relate to the theoretical predictions of WTC. The purpose this
variable was therefore to examine how frequently L2 learners use the language with different interlocutors and at various occasions which are predicted by WTC theory to be related to how willing they are to engage in such use under that range of conditions.

5.4 Data Collection Procedure

The data collection procedure differs in different universities due to language lab availability, the students’, the teachers’, and the researcher’s schedules. At university A in Hawai‘i, the researcher went to the Mandarin classrooms from level 101 to 301, including an intensive Business Chinese class. The researcher briefly explained the purpose of the research and informed them the steps that the research included and the approximate time the entire data collection procedure needed. They were informed that they would receive a compensation of five dollars for their time. Then a sign-up sheet was distributed to each class. The students who would like to participate wrote down their names, email addresses, and optional phone numbers. After the researcher collected all the sign-up sheets, individual emails were sent to each participant to schedule a time for data collection. The whole data collection procedure took place at the language lab which is conveniently located at the first floor of the building which holds all Mandarin classes on campus. When the participants arrived, they first read and signed the consent form. They were then asked to fill out a hardcopy of Mandarin learning survey, which includes background information, the affective factors, and frequency of L2 use in Mandarin learning. Upon finishing it, they were required to put on a headset and start the online EI task via WAU. The WAU was already logged on using unique participant numbers by the researcher. The participants just needed to put on a headset and clicked
the start button whenever they were ready. They followed the instructions recorded in the EI task, which walked them through the whole EI task. The EI task took approximately ten minutes. When they finished the entire collection procedure, the researcher sincerely expressed the appreciation and gave them a five-dollar gift card of Starbucks to thank for their time. The entire procedure took approximately from 30 to 45 minutes.

Some data samples were collected from university A in Hawai‘i, but through different procedures. The researcher was not able to present at the location, so two instructors of the Mandarin courses at the university helped with the data collection. As the online survey was available through internet, it saves time for the instructors if the participants finish and submit the survey at their own time. However, unfortunately, the online version of the EI task was not successfully loaded on some home computers due to software requirements. The researcher decided to collect the EI task data through the help of the instructors and leave the online survey to the participants’ own time. Therefore, the instructors first elicited possible participants from their classrooms. Those who were willing to participate made an appointment with their instructor. At the same time, they received an email containing the link to the online survey hosted by surveymonkey.com. They were instructed to finish the online survey at their earliest convenient time. Some of them completed the online survey before conducting the EI task, and some did it after the EI task. During the appointment, they signed the consent form and were asked to do the EI task. For the convenience of the instructors, the EI task was not conducted at the language lab. Instead, it was conducted in a quiet room, usually in the instructors’ office or in a vacant classroom. A desktop or laptop was used to play the task and at the same time, a digital recorder was used to record the participants’ responses together with the
stimuli sentences. The participants were compensated for five-dollar cash at the end of the appointment. The recording data were sent to the researcher via emails. At university B in Pennsylvania and university C in Wisconsin, the researcher was not able to present at the location either and the other two instructors helped with the data collection procedures. Due to the technical problem of WAU, the EI task was conducted through a laptop and a digital recorder. The whole process was the same as the process described above for university A.

At university D in California, the researcher went to a Mandarin classroom consisting of 18 students and explained the purpose and procedures of the research. They were told they were welcomed to participate in the research and the whole process was completely voluntary. A sign-up sheet had been prepared which included time slot based on the students’ and the researcher’s schedules. The students who would like to participate in the research chose a time slot which worked best for them and signed their names. At the scheduled time, the participant and the research would find a vacant classroom and the participant was given a consent form to read and sign. The computer in the classroom was used to play the EI task and a digital recorder was used to record the responses. When they finished the task, they were compensated with five-dollar cash and also provided with a hardcopy of survey. They were told to complete the survey and turn it in to the researcher as early as possible. Those who did not submit the survey after one week was reminded and in the end only one student chose not to submit the survey.

University E in California is a few hundred miles away from where the researcher is. Therefore, during the week before the data collection, a sign-up sheet was emailed to the instructors of the seven Mandarin classes at the university. The sign-up sheet only
restricted the date and the participants could sign up at any time during these days. They also needed to provide their names, email addresses, and optional phone numbers. The collected sign-up sheets were scanned and emailed to the researcher. The researcher sent out emails to individual participant to confirm their appointment time or make slight changes to their appointment time if there was a conflict among them. Along with the email, the link to the online survey was also sent and they were required to complete the online survey before the appointment if they could. An office at the university was used as the venue for data collection. Over two-thirds of the students who signed up for the study actually appeared at the appointment. Most of them had completed the online survey before they went there. Upon their arrival, they were given a consent form to sign and then started the EI task. A laptop and a digital recorder were used to record the 10-minute task. After the task, five-dollar cash was provided for compensation. Those who had not completed the online survey were reminded to do so after the recording sessions.

The procedures of data collection vary slightly at different venues. However, it may balance the potential effects caused by certain sequence of data collection procedures. For example, if the EI task was administered before the survey of perceived competence and communication anxiety, some participants might rate lower at perceived competence and higher at communication anxiety due to difficulty and frustration encountered at the EI task. Contrastively, some participants might not be able to perform to their best in the EI task because of fatigue after spending approximately 30 minutes to complete the survey.
5.5 Summary

The present study includes 179 participants from five universities at the United States. Their proficiency levels are different from beginning to advance based on their institutional status and also measured by the Mandarin EI task. A pilot study confirmed the reliability and validity of the Mandarin EI task. The survey included a background information questionnaire, L2 WTC, L2 Perceived competence, L2 Communication anxiety, and Frequency of L2 use. The data collection procedures differed at the five different universities. The differences may have helped to reduce the influence that may be caused by certain sequence of collection procedure.
CHAPTER 6
ANALYSIS AND DISCUSSION

The present study included 179 participants. Among them, 101 are non-heritage language learners and 78 are heritage learners of Mandarin. The heritage learner group was defined in the present study as learners who have at least one parent (and in 6 cases in the sample, a grandparent only) speaking Mandarin or a Chinese dialect at home and receive linguistic and/or cultural influence of Chinese at home regardless whether or not they have certain linguistic proficiency levels before formal instruction of Mandarin at school (for discussion of how these definitions may impact observed language achievement and likely self-perceptions, please see Kondo-Brown, 2005; see also section 6.1.1 later in this chapter for more discussion). AMOS (Arbuckle, 2011) was employed to examine the causal relationship among the factors in the L2 WTC construct. Pearson product-moment correlation coefficient was conducted to examine the correlation between L2 PC, L2 CA and L2 WTC in different individual L2 use contexts as measured by the last question in background survey about how much they use Mandarin in four occasions (See Appendix E). One-way ANOVA was conducted to investigate the differences between heritage and non-heritage language learners of Chinese on their L2 WTC, L2 PC, L2 CA, frequency of L2 use, and L2 actual proficiency levels. The results are analyzed and discussed.
6.1 Analysis

6.1.1 Descriptive Findings of Participants

The 179 participants in the study include 77 (43%) females and 102 (57%) males. The minimum age of the participants is 18 years old and the maximum age is 63 years old with the mean of 23.04 and the standard deviation of 7.01. Some of them started to hear or use Mandarin or Chinese dialects since they were born because their family members speak the language at home. The latest age of starting to hear or use Chinese is 38 years old. Excluding one participant who failed to provide the information, the average age of the remaining 178 participants to start the contact with Chinese language is at 15.96 years old and the standard deviation is 7.61. The participants also differed in years of formal instruction of Mandarin in school. Some just started taking Mandarin courses for about one month at the time of the study and some have been studying Mandarin for over ten years. The average years of taking Mandarin courses is about two years and the standard deviation is 2.14 (see Table 6.1).

Table 6.1
Biographic data of the participants

<table>
<thead>
<tr>
<th>Items</th>
<th>N</th>
<th>Min</th>
<th>Max</th>
<th>M</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>179</td>
<td>18</td>
<td>63</td>
<td>23.04</td>
<td>7.01</td>
</tr>
<tr>
<td>Age to start contact w/Chinese</td>
<td>178</td>
<td>0</td>
<td>38</td>
<td>15.96</td>
<td>7.61</td>
</tr>
<tr>
<td>Mandarin at school (year)</td>
<td>179</td>
<td>0.1</td>
<td>10.3</td>
<td>1.951</td>
<td>2.14</td>
</tr>
</tbody>
</table>

Among the 179 participants, seven (3.9%) reported their strongest language before age of five is Mandarin. Forty-two (23.5%) indicated that their strongest language before age five is one of the Chinese dialects. The dialects they reported include (a) Cantonese, a dialect mainly spoken in Guangdong province in Mainland China and Hong Kong; (b) Hakka and Taishan dialects, which are mainly spoken in some regions in
Guangdong Province in China; (c) Minnan, a dialect spoken in southern Fujian Province and Taiwan; and (d) Hmong, a dialect of the Hmong minority nationality in China. In addition, eight (4.5%) participants reported they spoke another language better than English before age of five. Those languages include Spanish, Vietnamese, Japanese, and Arabic. However, almost all (n = 177, 98.9%) reported that their strongest language after the age of five is English. Only two (1.1%) indicated that their strongest language after five remains the Chinese dialects, one is Cantonese and the other is Hmong (see Table 6.2).

Table 6.2

<table>
<thead>
<tr>
<th>Items</th>
<th>Frequency</th>
<th>Percent</th>
<th>Cumulative Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Before age five</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>English</td>
<td>122</td>
<td>68.2</td>
<td>68.2</td>
</tr>
<tr>
<td>Mandarin</td>
<td>7</td>
<td>3.9</td>
<td>72.1</td>
</tr>
<tr>
<td>Chinese dialects</td>
<td>42</td>
<td>23.5</td>
<td>95.5</td>
</tr>
<tr>
<td>Other</td>
<td>8</td>
<td>4.5</td>
<td>100.0</td>
</tr>
<tr>
<td>Total</td>
<td>179</td>
<td>100.0</td>
<td></td>
</tr>
<tr>
<td>After age five</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>English</td>
<td>177</td>
<td>98.9</td>
<td>98.9</td>
</tr>
<tr>
<td>Chinese dialects</td>
<td>2</td>
<td>1.1</td>
<td>100.0</td>
</tr>
<tr>
<td>Total</td>
<td>179</td>
<td>100.0</td>
<td></td>
</tr>
</tbody>
</table>

In order to better understand the language background of the participants, the survey collected information on whether they have any family members speaking Mandarin or a Chinese dialect. The result shows that 78 (43.6%) participants have family members speaking Mandarin or one of the Chinese dialects, and 101 (56.4%) do not have. The family members in the present study only include parents or grandparents and exclude spouse, girlfriend, or boyfriend. As mentioned in Chapter 3, Kondo-Brown (2005) has pointed out that students with different heritage background differ in performances of...
the language. Specifically, her findings suggested that heritage language learners who just have one grandparent speaking the language but no parent (a) may have lower confidence and weaker linguistic skills than those who have at least one heritage-language-speaking parent, and (b) may even be more similar in their proficiency to foreign language learners than to heritage language learners with at least parent who speaks the language. Therefore, the present research first attempted to categorize the participants with heritage background into subcategories according to family language background. However, it turned out that only 6 heritage language learners in the sample had grandparents speaking the heritage language and the remaining 72 heritage language learners all had at least one parent speaking the language at home. Given the importance of the affective and linguistic differences uncovered by Kondo-Brown (2005) for these two profiles of heritage learners in her study, an exploratory ANOVAs were conducted in the present study comparing the two subgroups of heritage language learners (N= 72 vs. N=6) to assess any differences in their proficiency levels, frequency of use, WTC inside and outside classrooms, perceived competence, and communication anxiety in Mandarin. No significant differences between the two subgroups on any these variables were found. These results do not rule out the possibility that important differences of the kind identified by Kondo-Brown in her study in fact existed in the present sample, as the lack of significance in the ANOVAs may well be due to the small sample size of the grandparent-only group as well as to the very unequal size of the two groups compared. Nevertheless, it would have been impossible to distinguish between the two types of heritage language learners in the present study, when one profile would have been comprised of only 6 participants, and discarding these 6 heritage participants’ data...
seemed wasteful given the need to ensure a very large sample of participants in order to perform the needed powerful statistical analyses that would address the research questions. In the end, therefore, it was decided to consider heritage language learners as a single category (which included the six participants with only a grandparent but not a parent who speaks the heritage language) in all the analyses reported in this dissertation, and to compare them to the single category of foreign language learners. This decision, although not ideal, is defensible also from a pedagogical perspective in that it reflects the realities of the definition of heritage language learners, which in many programs and occasions includes language learners of only grandparent speaking the target language as heritage language learners. Nevertheless, two cautions about the findings for heritage language learners that will be reported in the present study are in order. First, the precision of these results may be less than ideal because of the 6 heritage learners with only a grandparent who were included in the total sample of 78 heritage learners. Second, and alternatively, it is possible that any reported findings should be interpreted as more directly relevant and representative of heritage language learners that belong to the (a) profile in Kondo-Brown (2005), that is, learners with at least one parent who speaks the heritage language.

Ninety-six participants (53.6%) have visited a region or regions where the target language is spoken for the purpose of travelling, studying, or visiting relatives. The places they have visited include mainland China, Hong Kong, Taiwan, Singapore, and Malaysia (see Table 6.3). Among them, some made multiple visits to the target language regions, and some visited just once. When adding the time of visit or multiple visits of each participant together, it turns out the length of their total stay in the places vary in a
big range. The shortest stay was one week (0.25 month) and the longest stay lasted 11 years (132 months) with the mean of 16.06 months and standard deviation of 27.69 (see Table 6.4).

<table>
<thead>
<tr>
<th>Visit</th>
<th>Frequency</th>
<th>Percent</th>
<th>Cumulative Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>No</td>
<td>83</td>
<td>46.4</td>
<td>46.4</td>
</tr>
<tr>
<td>Yes</td>
<td>96</td>
<td>53.6</td>
<td>100.0</td>
</tr>
<tr>
<td>Total</td>
<td>179</td>
<td>100.0</td>
<td></td>
</tr>
</tbody>
</table>

Table 6.4
Length of visits to the regions of the target language (month)

<table>
<thead>
<tr>
<th></th>
<th>n</th>
<th>Min.</th>
<th>Max.</th>
<th>M</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Length</td>
<td>96</td>
<td>0.25</td>
<td>132.00</td>
<td>16.06</td>
<td>27.69</td>
</tr>
</tbody>
</table>

Of the participants having been in a target language region, 27 (15.1%) have studied there in a second language learning environment. The minimum length of study is five weeks (0.1 year) and the maximum is five years with the mean of 1.42 and the standard deviation of 1.49 (see Table 6.5).

<table>
<thead>
<tr>
<th></th>
<th>n</th>
<th>Min.</th>
<th>Max.</th>
<th>M</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Length</td>
<td>27</td>
<td>0.10</td>
<td>5.00</td>
<td>1.42</td>
<td>1.49</td>
</tr>
</tbody>
</table>

In order to have a further understanding of participants’ language background, the survey required the participants to provide information of their individual language use contexts by answering the question of how often they use Mandarin in four different occasions, at home, at school outside the Mandarin classroom, in the community, and in other occasions. A five-point Likert scale is used with 1 indicating never, 2 indicating occasionally, 3 indicating sometimes, 4 indicating frequently, and 5 indicating almost
always. The result shows that almost half \((n = 88, 49.2\%)\) students have never spoken Mandarin at home, about one-third \((n = 65, 36.3\%)\) have never spoken Mandarin at school outside the Mandarin classroom, and about two-thirds \((n = 113, 63.1\%)\) of participants have never used Mandarin in the community. Thirty-six \((20.1\%)\) reported they occasionally speak Mandarin at home, 18 \((10.1\%)\) reported sometimes, 15 \((8.4\%)\) reported frequently, and only 22 \((12.3\%)\) reported always speaking Mandarin at home.

As for speaking Mandarin at school outside the Mandarin classroom, 73 \((40.8\%)\) reported occasionally, 33 \((18.4\%)\) reported sometimes, and eight \((4.5\%)\) reported frequent use of Mandarin. In the community, 41 \((22.9\%)\) said they used Mandarin occasionally, and 20 \((11.2\%)\) used it sometimes. Only four \((2.2\%)\) and one \((0.6\%)\) reported frequent and almost always use in Mandarin respectively (see Table 6.6).

### Table 6.6

<table>
<thead>
<tr>
<th>Items</th>
<th>Frequency</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>home</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Never</td>
<td>88</td>
<td>49.2</td>
</tr>
<tr>
<td>Occasionally</td>
<td>36</td>
<td>20.1</td>
</tr>
<tr>
<td>Sometimes</td>
<td>18</td>
<td>10.1</td>
</tr>
<tr>
<td>Frequently</td>
<td>15</td>
<td>8.4</td>
</tr>
<tr>
<td>Almost always</td>
<td>22</td>
<td>12.3</td>
</tr>
<tr>
<td>Total</td>
<td>179</td>
<td>100.0</td>
</tr>
<tr>
<td>at school outside the Mandarin classroom</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Never</td>
<td>65</td>
<td>36.3</td>
</tr>
<tr>
<td>Occasionally</td>
<td>73</td>
<td>40.8</td>
</tr>
<tr>
<td>Sometimes</td>
<td>33</td>
<td>18.4</td>
</tr>
<tr>
<td>Frequently</td>
<td>8</td>
<td>4.5</td>
</tr>
<tr>
<td>Total</td>
<td>179</td>
<td>100.0</td>
</tr>
<tr>
<td>in the community</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Never</td>
<td>113</td>
<td>63.1</td>
</tr>
<tr>
<td>Occasionally</td>
<td>41</td>
<td>22.9</td>
</tr>
<tr>
<td>Sometimes</td>
<td>20</td>
<td>11.2</td>
</tr>
<tr>
<td>Frequently</td>
<td>4</td>
<td>2.2</td>
</tr>
<tr>
<td>Almost always</td>
<td>1</td>
<td>0.6</td>
</tr>
<tr>
<td>Total</td>
<td>179</td>
<td>100.0</td>
</tr>
</tbody>
</table>
Regarding their L2 use contexts in occasions other than the above three, forty-one participants indicated they used Mandarin they were learning in other occasions, which can be categorized into six occasions: (a) watch Chinese movies or listen to Chinese radios, (b) talk to relatives or friends in Mandarin, (c) hear others talking Mandarin in public, such as in Chinatown, (d) doing homework for Mandarin courses, and (e) read Chinese materials online. Except for the eight participants who did not specify the occasions, majority participants (n = 13, 31.7%) indicated they used Mandarin to talk to relatives or friends either face to face or via online chatting. Ten (24.4%) participants emphasized they like to watch Chinese movies. Five (12.2%) indicated they would hear others talking in Mandarin in Chinatown or at a relative’s house. Four (9.8%) reported they need to use Mandarin to do homework or study for class. Only one participant (2.4%) indicated that he likes to read Chinese materials online because of his major requirement. Table 6.7 shows the frequency of use in the six categories under the context of other occasions.

Table 6.7

<table>
<thead>
<tr>
<th>Occasions</th>
<th>Frequency</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>missing</td>
<td>8</td>
<td>19.5</td>
</tr>
<tr>
<td>Movies/TV/Radio</td>
<td>10</td>
<td>24.4</td>
</tr>
<tr>
<td>Talk to relatives or friends</td>
<td>13</td>
<td>31.7</td>
</tr>
<tr>
<td>hear others talking</td>
<td>5</td>
<td>12.2</td>
</tr>
<tr>
<td>Homework</td>
<td>4</td>
<td>9.8</td>
</tr>
<tr>
<td>Read</td>
<td>1</td>
<td>2.4</td>
</tr>
<tr>
<td>Total</td>
<td>41</td>
<td>100.0</td>
</tr>
</tbody>
</table>

The differences of L2 individual use contexts (derived from answers to the last item in Part 1 of the survey) among heritage and non-heritage learners were investigated. The results showed heritage language learners (M at home = 3.14, M in the community =
1.72) used significantly more Mandarin at home and in the community than non-heritage language learners ($M_{\text{home}} = 1.38$, $M_{\text{community}} = 1.41$). The effect size of the difference of individual L2 use at home is $\eta^2 = .38$, which means 38% of the variance can be accounted for by the variable. However, the effect size of the difference in individual L2 use in the community is only $\eta^2 = .04$, which means only 4% of the variance can be explained by the variable. At school outside the Mandarin classrooms, on the other hand, the results surprisingly showed that non-heritage language learners ($M = 2.04$) actually used significantly more Mandarin than heritage language learners ($M = 1.74$), although only 3% variance can be explained by this variable ($\eta^2 = .03$). Regarding all other occasions for individual L2 use contexts in, there was no significant result found among the two groups (See Table 6.8).

Table 6.8
*Differences of heritage and non-heritage learners of Mandarin in individual L2 use contexts*

<table>
<thead>
<tr>
<th>Context</th>
<th>N</th>
<th>M</th>
<th>SD</th>
<th>Min</th>
<th>Max</th>
<th>p</th>
<th>$\eta^2$ squared</th>
</tr>
</thead>
<tbody>
<tr>
<td>at home</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Non-heritage</td>
<td>101</td>
<td>1.38</td>
<td>0.705</td>
<td>1</td>
<td>5</td>
<td>0.00</td>
<td>0.38</td>
</tr>
<tr>
<td>Heritage</td>
<td>78</td>
<td>3.14</td>
<td>1.501</td>
<td>1</td>
<td>5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>179</td>
<td>2.15</td>
<td>1.423</td>
<td>1</td>
<td>5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>at school outside the Mandarin classroom</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Non-heritage</td>
<td>101</td>
<td>2.04</td>
<td>0.894</td>
<td>1</td>
<td>4</td>
<td>0.02</td>
<td>0.03</td>
</tr>
<tr>
<td>Heritage</td>
<td>78</td>
<td>1.74</td>
<td>0.763</td>
<td>1</td>
<td>3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>179</td>
<td>1.91</td>
<td>0.85</td>
<td>1</td>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>in the community</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Non-heritage</td>
<td>101</td>
<td>1.41</td>
<td>0.71</td>
<td>1</td>
<td>4</td>
<td>0.01</td>
<td>0.04</td>
</tr>
<tr>
<td>Heritage</td>
<td>78</td>
<td>1.72</td>
<td>0.924</td>
<td>1</td>
<td>5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>179</td>
<td>1.54</td>
<td>0.823</td>
<td>1</td>
<td>5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Other</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Non-heritage</td>
<td>101</td>
<td>1.38</td>
<td>0.823</td>
<td>1</td>
<td>4</td>
<td>0.31</td>
<td>0.01</td>
</tr>
<tr>
<td>Heritage</td>
<td>78</td>
<td>1.51</td>
<td>0.964</td>
<td>1</td>
<td>5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>179</td>
<td>1.44</td>
<td>0.887</td>
<td>1</td>
<td>5</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Note.* $p < .05$. 

134
The participants’ proficiency levels range from beginning to advance. The current study used multiple assessments to decide the participants’ proficiency levels. First, the participants were from different institutional status. Many of them were from the first semester of Mandarin learning at school, some from second and third year of Mandarin courses. However, as the curriculum designs in different institutes are different greatly, the institutional status alone cannot accurately represent the proficiency levels of participants. Therefore, the study also obtained the information regarding how many class sessions they have every week for their Mandarin courses. In addition, their contact with Chinese speaking people is also considered one factor that influences their proficiency levels. Therefore, their heritage background and individual Mandarin use contexts are also recorded as described above. Besides these biographic data of the participants, the study also used an in-house assessment, the Mandarin EI task, to measure the participants’ proficiency levels.

6.1.2 Analysis on Instruments

6.1.2.1 Elicited imitation task. In order to ensure good inter-coder reliability levels, two raters went through the same training mentioned in the EI pilot study and subsequently rated the Mandarin EI tasks independently. The inter rater reliability estimated with Pearson correlation is high ($r = .988$) and significant at 99% level ($p < .01$). The scatter plot also shows the two ratings are highly positively correlated.
A Cronbach’s alpha coefficient was computed to investigate the internal reliability of the Mandarin EI task scores. The result shows Mandarin EI is highly reliable \((r = .975)\). In order to examine the validity of the Mandarin EI task, a correlation was computed between the EI task and the participants’ background information. The amount of L2 contact was quantified based on the background information collected in Part 1 of the survey and following a point-system, as explained in Chapter 5. After all the items were quantified, the scores of each item were added to achieve a total score, which is the score of the participants’ L2 contact. Table 6.9 shows the descriptive statistics of participants’ EI and L2 contact scores. A correlation was computed between 179 participants’ L2 contact and the Mandarin EI task. The result shows the correlation is moderate and significant \((r = .594, p < .01)\).

Table 6.9

<table>
<thead>
<tr>
<th>Variables</th>
<th>(M)</th>
<th>(SD)</th>
<th>(N)</th>
</tr>
</thead>
<tbody>
<tr>
<td>EI</td>
<td>36.15</td>
<td>25.517</td>
<td>179</td>
</tr>
<tr>
<td>L2 contact</td>
<td>7.767</td>
<td>5.7346</td>
<td>179</td>
</tr>
</tbody>
</table>

*Note.* Maximum total possible score on the EI task was 120; there was no maximum total score on L2 contact because the length of Mandarin instruction and visit in target language region varied on individuals.
6.1.2.2 L2 WTC. The survey consists of L2 WTC, L2 perceived competence, L2 communication anxiety, and L2 frequency of use. The internal reliabilities of each part of the survey are all very high. Table 6.10 shows the reliability of each instrument in the survey.

Table 6.10
Reliability of the instruments

<table>
<thead>
<tr>
<th>Instruments</th>
<th>N</th>
<th>K</th>
<th>Cronbach's Alpha</th>
</tr>
</thead>
<tbody>
<tr>
<td>L2 Frequency of use</td>
<td>179</td>
<td>12</td>
<td>.923</td>
</tr>
<tr>
<td>L2 WTC inside classroom</td>
<td>179</td>
<td>27</td>
<td>.955</td>
</tr>
<tr>
<td>L2 WTC outside classroom</td>
<td>179</td>
<td>27</td>
<td>.970</td>
</tr>
<tr>
<td>L2 Perceived competence</td>
<td>179</td>
<td>12</td>
<td>.985</td>
</tr>
<tr>
<td>L2 Communication anxiety</td>
<td>179</td>
<td>12</td>
<td>.948</td>
</tr>
</tbody>
</table>

L2 Mandarin WTC includes two parts, inside and outside the Mandarin classrooms. For the L2 WTC inside classroom, participants shows highest WTC in speaking ($M = 25.32$, $n = 179$), followed by writing ($M = 24.73$), reading ($M = 20.59$), and comprehension ($M = 18.43$). The L2 WTC outside classroom shows the same pattern that the participants indicated highest WTC in speaking ($M = 21.62$) and lowest in comprehension ($M = 15.56$) (see Table 6.11). Pearson correlation was calculated to examine the correlation between the four skills in WTC inside Mandarin classrooms. The results turn out that speaking inside classroom is significantly correlated with reading ($r = .531$), writing ($r = .612$), and comprehension ($r = .647$) inside classroom ($p < .01$). Reading is significantly correlated with writing ($r = .725$) and comprehension ($r = .596$) inside classroom. There was also a significant and strong positive correlation between writing and comprehension inside classroom ($r = .599$). According to Cohen (1988), correlation coefficient is considered large when it exceeds 0.5. In L2 WTC outside classroom, too, strong positive correlations were also found between the four skills ($n =$
The correlations of speaking with reading, writing, and comprehension outside classroom are $r = .685$, $r = .754$, and $r = .772$ respectively ($p < .01$). Reading and writing were also positively correlated ($r = .770$, $p < .01$). The correlation between reading and comprehension is 0.75. A strong correlation between writing and comprehension was also detected ($r = .768$, $p < .01$) (see Table 6.12).

**Table 6.11**

*Descriptive data of L2 WTC*

<table>
<thead>
<tr>
<th>Variables</th>
<th>$M$</th>
<th>$SD$</th>
<th>$N$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Inside classroom</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Speaking</td>
<td>25.32</td>
<td>7.840</td>
<td>179</td>
</tr>
<tr>
<td>Reading</td>
<td>20.59</td>
<td>6.347</td>
<td>179</td>
</tr>
<tr>
<td>Writing</td>
<td>24.73</td>
<td>8.837</td>
<td>179</td>
</tr>
<tr>
<td>Comprehension</td>
<td>18.43</td>
<td>5.025</td>
<td>179</td>
</tr>
<tr>
<td><strong>Outside classroom</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Speaking</td>
<td>21.62</td>
<td>8.822</td>
<td>179</td>
</tr>
<tr>
<td>Reading</td>
<td>17.75</td>
<td>7.342</td>
<td>179</td>
</tr>
<tr>
<td>Writing</td>
<td>19.31</td>
<td>9.165</td>
<td>179</td>
</tr>
<tr>
<td>Comprehension</td>
<td>15.56</td>
<td>5.970</td>
<td>179</td>
</tr>
</tbody>
</table>

**Table 6.12**

*Pearson correlation between L2 WTC*

<table>
<thead>
<tr>
<th>Inside classroom</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1) Speaking-inside</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2) Reading-inside</td>
<td>.531(*)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3) Writing-inside</td>
<td>.612(*)</td>
<td>.725(*)</td>
<td></td>
</tr>
<tr>
<td>4) Comprehension-inside</td>
<td>.647(*)</td>
<td>.596(*)</td>
<td>.599(*)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Outside classroom</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1) Speaking-outside</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2) Reading-outside</td>
<td>.685(*)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3) Writing-outside</td>
<td>.754(*)</td>
<td>.770(*)</td>
<td></td>
</tr>
<tr>
<td>4) Comprehension-outside</td>
<td>.772(*)</td>
<td>.750(*)</td>
<td>.768(*)</td>
</tr>
</tbody>
</table>

*Note. N = 179, $p < .01$.*

The result of comparing the means of different skills in L2 WTC inside and outside classroom shows an overall trend of the participants more willing to use Mandarin inside than outside classroom in all four skills. WTC in speaking inside classroom ($M = 25.32$) is significantly higher than the WTC in speaking outside
classroom \((M = 21.62) (p < .05)\). The participants also show significantly higher WTC in reading \((M = 20.59)\), writing \((M = 24.73)\), and comprehension \((M = 18.43)\) inside classroom than those outside classroom \((M = 17.75, \text{19.31, 15.56, } p < .05)\). According to the guidelines of the eta-squared values from Cohen (1988), a value greater than .14 indicates a large effect. Therefore, the differences in each skill contain large effect sizes between inside and outside classroom (.27 in speaking, .25 in reading, .35 in writing, and .33 in comprehension, .41 in total).

Table 6.13

<table>
<thead>
<tr>
<th>Pairs</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>M</td>
<td>SD</td>
<td>Eta squared</td>
<td>t</td>
<td>df</td>
<td>p</td>
</tr>
<tr>
<td>1 Speaking-inside</td>
<td>3.704</td>
<td>6.048</td>
<td>.27</td>
<td>8.193</td>
<td>178</td>
<td>.000</td>
</tr>
<tr>
<td>Speaking-outside</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2 Reading-inside</td>
<td>2.838</td>
<td>4.988</td>
<td>.25</td>
<td>7.613</td>
<td>178</td>
<td>.000</td>
</tr>
<tr>
<td>Reading-outside</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3 Writing-inside</td>
<td>5.425</td>
<td>7.425</td>
<td>.35</td>
<td>9.774</td>
<td>178</td>
<td>.000</td>
</tr>
<tr>
<td>Writing-outside</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4 Comprehension-inside</td>
<td>2.866</td>
<td>4.133</td>
<td>.33</td>
<td>9.277</td>
<td>178</td>
<td>.000</td>
</tr>
<tr>
<td>Comprehension-outside</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5 Total-inside</td>
<td>14.832</td>
<td>17.915</td>
<td>.41</td>
<td>11.077</td>
<td>178</td>
<td>.000</td>
</tr>
<tr>
<td>Total-outside</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

6.1.3 Findings of L2 WTC Construct

6.1.3.1 Model 1. The present study intends to seek answers of the relationship between the factors in the L2 WTC construct. In order to answer the research questions on the relationship between the affective factors in L2 Mandarin, AMOS (Arbuckle, 2011) was used to examine the structural equation modeling (SEM) of the L2 WTC. Based on previous literature, L2 perceived competence and L2 communication anxiety may be the two antecedents of L2 WTC. Communication anxiety contributes to perceived competence to some extent. Willingness to communicate is believed to lead to the actual
use of the language. Therefore, it is hypothesized that high WTC may cause the high frequency of L2 use. Practice makes perfect. As a result, frequency of L2 use may increase the proficiency level of the language. Therefore, the following model is established and tested by the variance-covariance matrix.

*Figure 6.2 Path Diagram of Model 1*

6.1.3.1 Model identification. Model identification is the key in AMOS to find out whether the SEM model has a solution and can be identified (Byrne, 2010). There are three levels of the model identification: under-identified, just-identified, and over-identified (Schumacker & Lomax, 2004, p. 64). When the model is just-identified or over-identified, there are one or more solutions to the model. On the other hand, an under-identified model suggests no solution and needs to be imposed additional constraints in the model. In order to determine the identification, the number of free parameters and the number of distinct values need to be calculated. When the former is equal or less than the later, the model is identified and has solutions. Through the following calculation, the number of parameters in the model is 15. The above model includes nine path coefficients, therefore, based on the formula \( p(p + 1)/2 = (9 \times 10)/2 \)
The distinct value is 45. The former is less than the latter. As a result, the current model is testable and has solutions.

Parameters:

1 exogenous variance
9 path coefficients
5 error variance for endogenous variables  =  15

15 < 45 → it is an over-specified model and has solutions

Before running the test of the hypothesized model, it is important to check the normality of the data. The Kurtosis value equal to or greater than seven is considered to be indicative of early departure from normality (Byrne, 2010, p. 103). Table 6.14 shows no variables have kurtosis greater than seven. Therefore, the data are multivariate normal. Further examination was conducted by checking the value of multivariate critical ration (c.r.). The result (c.r. = 4.013) again shows that the data is multivariate normal (c.r. < 5).

Table 6.14
Assessment of normality of Model 1

<table>
<thead>
<tr>
<th>Variables</th>
<th>min.</th>
<th>max.</th>
<th>skew</th>
<th>c.r.</th>
<th>kurtosis</th>
<th>c.r.</th>
</tr>
</thead>
<tbody>
<tr>
<td>L2 CA</td>
<td>.000</td>
<td>1200.000</td>
<td>-0.144</td>
<td>-0.788</td>
<td>-0.210</td>
<td>-0.574</td>
</tr>
<tr>
<td>L2 PC</td>
<td>.000</td>
<td>1140.000</td>
<td>0.281</td>
<td>1.537</td>
<td>-1.111</td>
<td>-3.034</td>
</tr>
<tr>
<td>L2 WTC inside</td>
<td>27.000</td>
<td>135.000</td>
<td>-0.334</td>
<td>-1.827</td>
<td>-0.344</td>
<td>-0.939</td>
</tr>
<tr>
<td>L2 WTC outside</td>
<td>27.000</td>
<td>135.000</td>
<td>0.189</td>
<td>1.035</td>
<td>-0.822</td>
<td>-2.245</td>
</tr>
<tr>
<td>L2 Freq. of use</td>
<td>12.000</td>
<td>70.000</td>
<td>1.845</td>
<td>10.080</td>
<td>4.556</td>
<td>12.444</td>
</tr>
<tr>
<td>L2 Proficiency level</td>
<td>2.000</td>
<td>120.000</td>
<td>0.925</td>
<td>5.052</td>
<td>0.246</td>
<td>0.671</td>
</tr>
<tr>
<td>Multivariate</td>
<td></td>
<td></td>
<td></td>
<td>5.878</td>
<td>4.013</td>
<td></td>
</tr>
</tbody>
</table>

To investigate whether there are cases whose scores are substantially different from all the others in a particular set of data, the squared Mahalanobis distance was computed. According to Bynes (2010), this set of statistics measures “the distance in standard deviation units between a set of scores for one case and the sample means for all
variables (centroids)” (p. 106). The calculation of Mahalanobis distance shows no scores are substantially different in a particular set of data.

Table 6.15 shows the estimate covariance of each pair of variables. When L2 CA goes up by one, L2 PC goes down by 0.315, with the standard error of 0.083. The largest covariance exists between L2 frequency of use and L2 proficiency level. When L2 frequency of use goes up by 1, L2 proficiency level goes up by 1.386. The parameters from L2 CA to L2 WTC inside classroom, from L2 CA to L2 WTC outside classroom, and from L2 WTC inside classroom to L2 frequency of use are not significant. Therefore, modification may be needed by reducing the three non-significant paths. Details will be discussed in the following part.

Table 6.15

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Estimate</th>
<th>S.E.</th>
<th>C.R.</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>L2 PC &lt;- L2 CA</td>
<td>-.315</td>
<td>.083</td>
<td>-3.810</td>
<td>***</td>
</tr>
<tr>
<td>L2 WTC inside &lt;--- L2 PC</td>
<td>.028</td>
<td>.005</td>
<td>5.528</td>
<td>***</td>
</tr>
<tr>
<td>L2 WTC inside &lt;--- L2 CA</td>
<td>-.009</td>
<td>.006</td>
<td>-1.598</td>
<td>.110</td>
</tr>
<tr>
<td>L2 WTC outside &lt;--- L2 WTC inside</td>
<td>.828</td>
<td>.060</td>
<td>13.777</td>
<td>***</td>
</tr>
<tr>
<td>L2 WTC outside &lt;--- L2 PC</td>
<td>.013</td>
<td>.004</td>
<td>2.833</td>
<td>.005</td>
</tr>
<tr>
<td>L2 WTC outside &lt;--- L2 CA</td>
<td>-.008</td>
<td>.005</td>
<td>-1.602</td>
<td>.109</td>
</tr>
<tr>
<td>L2 Freq. of use &lt;--- L2 WTC outside</td>
<td>.148</td>
<td>.036</td>
<td>4.071</td>
<td>***</td>
</tr>
<tr>
<td>L2 Freq. of use &lt;--- L2 WTC inside</td>
<td>.010</td>
<td>.043</td>
<td>.223</td>
<td>.823</td>
</tr>
<tr>
<td>L2 Proficiency level &lt;--- L2 Frequency of use</td>
<td>1.386</td>
<td>.167</td>
<td>8.285</td>
<td>***</td>
</tr>
</tbody>
</table>

Note. p < .01.

6.1.3.1.2 Goodness-of-fit indices of Model 1. The chi-square is equal to 88.278, with degrees of freedom of six, and a p value less than 0.01. The chi-square value is statistically significant (p < 0.01) and is not close in value to the degrees of freedom. This model fit index indicates that the initial model is not entirely adequate. However, the chi-square index is only a quick overview of model fit. As it is very sensitive to sample size, it tends to be statistically significant with a large sample size (Byrne, 2010). Both the
Goodness-of-Fit Index (GFI) and the Adjusted Goodness-of-Fit Index (AGFI) are measures of the relative amount of variance and covariance (Byrne, 2010, p. 77). The Goodness-of-Fit Index is .887, which is below the acceptable range of model fit (GFI > .95). The Adjusted Goodness-of-Fit Index (AGFI) is 0.605 for this model, not an acceptable level of fit (AGFI > .95). The comparative Fit Index (CFI) suggests whether the model fits the data well in the sense that the hypothesized model adequately described the sample data. Unfortunately, the CFI index (.794) again indicates an inadequate fit of the hypothesized model (CFI > .90). The root-mean-square error of approximation (RMSEA) answers the questions of “How well would the model, with unknown but optimally chosen parameter values, fit the population covariance matrix if it were available?” (Browne & Cudeck, 1993, pp. 137-138). RMSEA with the value less than .05 indicates a good fit, values ranging from .08 to 0.1 indicate a mediocre fit, and those greater than .10 indicate a poor fit (Byrne, 2010, p. 80). As a result, RMSEA of .278 shows unacceptable errors of approximation in the population, also below the typical acceptable level of model fit (RMSEA < .05). Therefore, the conclusion is that the hypothesized structural model is reasonable but some model modification might allow us to achieve a more acceptable model-to-data fit.

Table 6.16  
*Goodness-of-fit indices of Model 1*

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Chi-square</td>
<td>88.278</td>
</tr>
<tr>
<td>df</td>
<td>6</td>
</tr>
<tr>
<td>p value</td>
<td>.000</td>
</tr>
<tr>
<td>GFI</td>
<td>.887</td>
</tr>
<tr>
<td>AGFI</td>
<td>.605</td>
</tr>
<tr>
<td>CFI</td>
<td>.794</td>
</tr>
<tr>
<td>RMSEA</td>
<td>.278</td>
</tr>
</tbody>
</table>
6.1.3.1.3 Model misspecification. There are two types of information that are helpful in detecting model misspecification, the standardized residuals and the modification indices. The standard residuals with values greater than 2.58 are considered large enough to call attention for the significant discrepancy in the model. The table shows that two standard residuals between L2 PC and L2 frequency of use (4.797), and L2 PC and L2 proficiency level (6.648) are greater than the cut point 2.58. Therefore, these two pairs of variables may contain significant correlations between them.

<table>
<thead>
<tr>
<th>Variables</th>
<th>1)</th>
<th>2)</th>
<th>3)</th>
<th>4)</th>
<th>5)</th>
<th>6)</th>
</tr>
</thead>
<tbody>
<tr>
<td>L2 CA</td>
<td>0.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>L2 PC</td>
<td>0.000 0.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>L2 WTC inside</td>
<td>0.000 0.000 0.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>L2 WTC outside</td>
<td>0.000 0.000 0.000 0.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>L2 Frequency of use</td>
<td>-2.218 4.797 0.000 0.000 0.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>L2 Proficiency level</td>
<td>-1.147 6.648 1.643 1.946 0.000 0.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Modification Indices (MI) also reflect the possible modification of the hypothesized model. The MI represents the expected drop in the overall chi-square value if the parameter were to be freely estimated. The Par Change is the expected parameter change value. The results of modification indices suggest that a large Par change in regression weight exist between L2 WTC outside classroom and L2 proficiency level. There are also misspecification detected between L2 PC and L2 proficiency levels, L2 PC and L2 frequency of use. Based on the information in standardized residuals and modification indices, it is reasonable to proceed to modify the initial model.
6.1.3.2 Modified models. In order to modify the model, the paths suggested in the above MI need to be added one by one to test the model fit for every revision. The path between L2 WTC outside classroom and L2 proficiency level is first added to Model 2 because the Par change between them is the largest. The Goodness-of-Fit indices of GFI and CFI increase slightly, but still fail to reach the acceptable value, which indicates the model is still very weak in its fit.

The three paths suggested in the above MI of the initial model remain in the model 2 as shown in Table 6.20 The largest Par change exists in the path from L2 PC to L2 proficiency of use.

<table>
<thead>
<tr>
<th>Parameters</th>
<th>M.I.</th>
<th>Par Change</th>
</tr>
</thead>
<tbody>
<tr>
<td>L2 Freq. of use &lt;--- L2 CA</td>
<td>6.263</td>
<td>-.006</td>
</tr>
<tr>
<td>L2 Freq. of use &lt;--- L2 PC</td>
<td>30.113</td>
<td>.011</td>
</tr>
<tr>
<td>L2 Proficiency level &lt;--- L2 PC</td>
<td>23.324</td>
<td>.024</td>
</tr>
<tr>
<td>L2 Proficiency level &lt;--- L2 WTC outside</td>
<td>5.543</td>
<td>.136</td>
</tr>
</tbody>
</table>
Based on the result yielded in Model 2, the path from L2 PC to L2 proficiency level is added. The Goodness-of-fit indices are all improved slightly, but the RMSEA which carries most information is still below the acceptable level. The modification indices suggest two more paths and the larger one is from L2 PC to L2 frequency of use. After adding the path from L2 PC to L2 Frequency of use, the Goodness-of-fit indices have been improved significantly. The chi-square value is 4.493 with degree of freedom of three \((p > .05)\). A non-significant chi-square closer to the value of degree of freedom indicates a good fit of the model. The values of GFI, AGFI, and CFI are all above the cut point .95, indicating an excellent fit. In addition, RMSEA is around .05. It is concluded that the model fits the data well.

Table 6.21

<table>
<thead>
<tr>
<th>Goodness-of-fit indices of Model 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chi-square</td>
</tr>
<tr>
<td>(df)</td>
</tr>
<tr>
<td>(p) value</td>
</tr>
<tr>
<td>GFI</td>
</tr>
<tr>
<td>AGFI</td>
</tr>
<tr>
<td>CFI</td>
</tr>
<tr>
<td>RMSEA</td>
</tr>
</tbody>
</table>

The results obtained from Model 4 indicate four non-significant parameters in the model: the parameters from L2 CA to L2 WTC inside classroom, L2 CA to L2 WTC outside classroom, L2 WTC inside classroom to L2 frequency of use, and L2 WTC outside classroom to L2 proficiency level. The parameter from L2 WTC outside classroom to L2 proficiency level was added in the first step of model modification. However, after the overall modification, the parameter shows to be non-significant. As a result, the final step of the modification is to reduce the non-significant parameters from the model.
Table 6.2

AMOS output for parameter estimates of Model 4

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Estimate</th>
<th>S.E.</th>
<th>C.R.</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>L2 PC</td>
<td>&lt;---</td>
<td>L2 CA</td>
<td>-.315</td>
<td>.083</td>
</tr>
<tr>
<td>L2 WTC inside</td>
<td>&lt;---</td>
<td>L2 PC</td>
<td>.028</td>
<td>.005</td>
</tr>
<tr>
<td>L2 WTC inside</td>
<td>&lt;---</td>
<td>L2 CA</td>
<td>-.009</td>
<td>.006</td>
</tr>
<tr>
<td>L2 WTC outside</td>
<td>&lt;---</td>
<td>L2 WTC inside</td>
<td>.828</td>
<td>.060</td>
</tr>
<tr>
<td>L2 WTC outside</td>
<td>&lt;---</td>
<td>L2 CA</td>
<td>-.008</td>
<td>.005</td>
</tr>
<tr>
<td>L2 WTC outside</td>
<td>&lt;---</td>
<td>L2 PC</td>
<td>.013</td>
<td>.004</td>
</tr>
<tr>
<td>L2 Frequency of use</td>
<td>&lt;---</td>
<td>L2 WTC outside</td>
<td>.094</td>
<td>.033</td>
</tr>
<tr>
<td>L2 Frequency of use</td>
<td>&lt;---</td>
<td>L2 WTC inside</td>
<td>-.021</td>
<td>.039</td>
</tr>
<tr>
<td>L2 Frequency of use</td>
<td>&lt;---</td>
<td>L2 PC</td>
<td>.014</td>
<td>.002</td>
</tr>
<tr>
<td>L2 Proficiency level</td>
<td>&lt;---</td>
<td>L2 Frequency of use</td>
<td>.634</td>
<td>.189</td>
</tr>
<tr>
<td>L2 Proficiency level</td>
<td>&lt;---</td>
<td>L2 WTC outside</td>
<td>.071</td>
<td>.060</td>
</tr>
<tr>
<td>L2 Proficiency level</td>
<td>&lt;---</td>
<td>L2 PC</td>
<td>.034</td>
<td>.006</td>
</tr>
</tbody>
</table>

Note. p < .01.

After the deletion of the non-significant paths, the final model is presented below.

The number of parameters in the model is 14. The model includes eight path coefficients. Therefore, the distinct value is 36. The former is less than the latter. As a result, the current model is testable and has solutions.

Parameters: 1 exogenous variance +
8 path coefficients +
5 error variance for endogenous variables = 14

16 < 36 → it is an over specified model and has solutions.
Figure 6.3 Path diagram of Model 5 (Final Model)

All the parameters estimates in the table below are significant. There are no suggestions for modification indices. Therefore, all the paths in the final model are significant.

Table 6.23
AMOS output for parameter estimates of the final model

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Estimate</th>
<th>S.E.</th>
<th>C.R.</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>L2 PC &lt;--- L2 CA</td>
<td>-.315</td>
<td>.083</td>
<td>-3.810</td>
<td>***</td>
</tr>
<tr>
<td>L2 WTC inside &lt;--- L2 PC</td>
<td>.031</td>
<td>.005</td>
<td>6.162</td>
<td>***</td>
</tr>
<tr>
<td>L2 WTC outside &lt;--- L2 WTC inside</td>
<td>.839</td>
<td>.060</td>
<td>13.967</td>
<td>***</td>
</tr>
<tr>
<td>L2 WTC outside &lt;--- L2 PC</td>
<td>.014</td>
<td>.004</td>
<td>3.210</td>
<td>.001</td>
</tr>
<tr>
<td>L2 Frequency of use &lt;--- L2 WTC outside</td>
<td>.081</td>
<td>.023</td>
<td>3.530</td>
<td>***</td>
</tr>
<tr>
<td>L2 Frequency of use &lt;--- L2 PC</td>
<td>.014</td>
<td>.002</td>
<td>7.005</td>
<td>***</td>
</tr>
<tr>
<td>L2 Proficiency level &lt;--- L2 Frequency of use</td>
<td>.691</td>
<td>.183</td>
<td>3.766</td>
<td>***</td>
</tr>
<tr>
<td>L2 Proficiency level &lt;--- L2 PC</td>
<td>.036</td>
<td>.005</td>
<td>6.585</td>
<td>***</td>
</tr>
</tbody>
</table>

*Note.* $p < .01.$

The Table of Goodness-of-fit indices indicates a good fit of the final model with GFI of .98, AGFI of .941, and CFI of .989. The RMSEA is close to .05 (see Table 6.24).
Table 6.24
*Goodness-of-fit indices of the final model*

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Chi-square</td>
<td>11.265</td>
</tr>
<tr>
<td>df</td>
<td>7</td>
</tr>
<tr>
<td>p value</td>
<td>0.127</td>
</tr>
<tr>
<td>GFI</td>
<td>0.980</td>
</tr>
<tr>
<td>AGFI</td>
<td>0.941</td>
</tr>
<tr>
<td>CFI</td>
<td>0.989</td>
</tr>
<tr>
<td>RMSEA</td>
<td>0.059</td>
</tr>
</tbody>
</table>

Figure 6.4 L2 WTC construct in the Final model

Note. p < 0.01, N = 179.

The L2 WTC construct in the Final model appears to be slightly different from what is suggested in the literature. L2 CA predicts L2 PC directly (\( r = -.27 \)), and L2 PC is the only direct antecedent of L2 WTC. Both L2 PC and L2 WTC inside classroom are antecedents of L2 WTC outside classroom. However, L2 WTC inside classroom (\( r = .71 \)) is a much stronger predictor of L2 WTC outside classroom than L2 PC (\( r = .16 \)). Only L2 WTC outside, instead of inside classroom predicts learners’ L2 frequency of use (\( r = .24 \)). In turn, L2 proficiency level of learners is the outcome of L2 frequency of use (\( r = .26 \)).
In addition, learners’ L2 PC predicts learners L2 frequency of use \( (r = .47) \) and actual L2 proficiency levels \( (r = .46) \). The path coefficient from L2 WTC inside to outside classrooms is the highest in the model, followed by the path from L2 PC to L2 Frequency of use, L2 actual proficiency level, and L2 WTC inside classrooms.

### 6.1.4 Findings of the Affective Factors in Individual L2 Use Contexts

Individual differences of L2 use context was measured by the last question in the background information questionnaire. The participants were asked to rate from one to five on how often they use Mandarin outside classroom in four different situations. The total score varies from four to 20. The midpoint eight is used as the cutoff point for high and low L2 use context. Therefore, the participants who indicated a total score equal or higher than eight are assigned to the group of high L2 use context and the other participants are in the group of low L2 use context. It turns out that 120 participants are using Mandarin in an L2 low use context and 59 are in a high L2 use context.

The relationship between L2 PC, L2 CA and L2 WTC inside and outside classroom in individual low and high L2 use context was investigated using Pearson product-moment correlation coefficient. The results show that L2 PC is significantly correlated with L2 WTC in both low and high L2 use context, and L2 CA is significantly correlated with L2 WTC in the high use context only. In the both low and high L2 use context, L2 PC is more correlated with L2 WTC inside and outside classroom than L2 CA. However, the correlations between L2 CA and L2 WTC increased in the high L2 use context.
Table 6.25

Correlation between L2 PC, L2 CA and L2 WTC in individual L2 use context

<table>
<thead>
<tr>
<th>L2 use context</th>
<th>Low (n = 120)</th>
<th>High (n = 59)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>L2 WTC inside</td>
<td>L2 WTC outside</td>
</tr>
<tr>
<td>1) L2 PC</td>
<td>.307*</td>
<td>.310*</td>
</tr>
<tr>
<td>2) L2 CA</td>
<td>-.158</td>
<td>-.176</td>
</tr>
</tbody>
</table>

Note. *p < .05

6.1.5 Findings on Heritage Language Learners of Mandarin

Among the 179 participants, 101 do not have any heritage background on Chinese language and 78 have at least one parent (or in 6 cases, a grandparent only) who speaks Chinese at home. Both heritage and non-heritage learners show higher L2 WTC inside than outside classroom. ANOVA analyses were conducted to inspect any differences across variables between the two groups. Alpha level was set at a stringent $p < .008$, by dividing the overall conventional alpha level (.05) by the number of variables (6). As shown in Table 6.27, the differences between the heritage and non-heritage language learner groups of Mandarin are non-significant in all the six variables compared. However, at the descriptive statistics level (see Table 6.26), a tendency is seen for the heritage language learner group towards lower L2 WTC both inside ($M = 84.71$) and outside ($M = 69.27$) classroom, when compared to the non-heritage language learner group ($M$ inside = 92.46, $M$ outside = 78.09). Heritage language learners ($M = 714.9$) reported a trend for higher communication anxiety than non-heritage language learners ($M = 625.95$). No significant difference was found on L2 perceived competence between heritage and non-heritage Chinese learners, but heritage learners ($M = 482.77$) showed slightly higher PC than non-heritage learners ($M = 479.07$). Interestingly, non-heritage language learners ($M = 23.24$) reported a tendency for higher frequency of L2 use than
heritage language learners \( (M = 19.42) \). Based on the result of the Mandarin EI task, heritage language learners \( (M = 38.72) \) did perform slightly better than non-heritage language learners \( (M = 34.17) \) on Mandarin EI task, but the difference is small and not significant. In interpreting these findings, readers are reminded that the precision of these comparisons between heritage and non-heritage language learner groups in the present study may be less than ideal because of the 6 heritage learners with only a grandparent who were included in the total sample of 78 heritage learners. Future studies with large enough samples of maximally homogenous heritage language learning profiles may need to revisit the question of whether/which profiles of heritage language learners are different from foreign language learners of Mandarin.

Table 6.2

Descriptive data of heritage and non-heritage language learners

<table>
<thead>
<tr>
<th>Variables</th>
<th>Groups</th>
<th>n</th>
<th>M</th>
<th>SD</th>
<th>Min.</th>
<th>Max.</th>
</tr>
</thead>
<tbody>
<tr>
<td>L2 WTC inside</td>
<td>Non-heritage</td>
<td>101</td>
<td>92.46</td>
<td>24.343</td>
<td>37</td>
<td>135</td>
</tr>
<tr>
<td></td>
<td>Heritage</td>
<td>78</td>
<td>84.71</td>
<td>22.477</td>
<td>27</td>
<td>121</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>179</td>
<td>89.08</td>
<td>23.797</td>
<td>27</td>
<td>135</td>
</tr>
<tr>
<td>L2 WTC outside</td>
<td>Non-heritage</td>
<td>101</td>
<td>78.09</td>
<td>29.629</td>
<td>27</td>
<td>135</td>
</tr>
<tr>
<td></td>
<td>Heritage</td>
<td>78</td>
<td>69.27</td>
<td>25.641</td>
<td>27</td>
<td>126</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>179</td>
<td>74.25</td>
<td>28.228</td>
<td>27</td>
<td>135</td>
</tr>
<tr>
<td>L2 PC</td>
<td>Non-heritage</td>
<td>101</td>
<td>479.07</td>
<td>321.660</td>
<td>0</td>
<td>1100</td>
</tr>
<tr>
<td></td>
<td>Heritage</td>
<td>78</td>
<td>482.77</td>
<td>330.855</td>
<td>9</td>
<td>1140</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>179</td>
<td>480.68</td>
<td>324.781</td>
<td>0</td>
<td>1140</td>
</tr>
<tr>
<td>L2 CA</td>
<td>Non-heritage</td>
<td>101</td>
<td>625.95</td>
<td>286.558</td>
<td>0</td>
<td>1200</td>
</tr>
<tr>
<td></td>
<td>Heritage</td>
<td>78</td>
<td>714.90</td>
<td>271.335</td>
<td>0</td>
<td>1200</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>179</td>
<td>664.71</td>
<td>282.731</td>
<td>0</td>
<td>1200</td>
</tr>
<tr>
<td>L2 Freq. of use</td>
<td>Non-heritage</td>
<td>101</td>
<td>23.24</td>
<td>10.925</td>
<td>12</td>
<td>70</td>
</tr>
<tr>
<td></td>
<td>Heritage</td>
<td>78</td>
<td>19.42</td>
<td>7.388</td>
<td>12</td>
<td>48</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>179</td>
<td>21.58</td>
<td>9.709</td>
<td>12</td>
<td>70</td>
</tr>
<tr>
<td>L2 Proficiency level</td>
<td>Non-heritage</td>
<td>101</td>
<td>34.17</td>
<td>25.290</td>
<td>2</td>
<td>112</td>
</tr>
<tr>
<td></td>
<td>Heritage</td>
<td>78</td>
<td>38.72</td>
<td>25.743</td>
<td>5</td>
<td>120</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>179</td>
<td>36.15</td>
<td>25.517</td>
<td>2</td>
<td>120</td>
</tr>
</tbody>
</table>
Table 6.27
ANOVA results of heritage and non-heritage language learners

<table>
<thead>
<tr>
<th>Variables</th>
<th>Between Groups</th>
<th>df</th>
<th>F</th>
<th>p</th>
<th>Eta squared</th>
</tr>
</thead>
<tbody>
<tr>
<td>L2 WTC inside</td>
<td></td>
<td>1</td>
<td>4.767</td>
<td>.030</td>
<td>.026</td>
</tr>
<tr>
<td>Within Groups</td>
<td></td>
<td>177</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>178</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>L2 WTC outside</td>
<td></td>
<td>1</td>
<td>4.378</td>
<td>.038</td>
<td>.024</td>
</tr>
<tr>
<td>Within Groups</td>
<td></td>
<td>177</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>178</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>L2 PC</td>
<td></td>
<td>1</td>
<td>.006</td>
<td>.940</td>
<td>.000</td>
</tr>
<tr>
<td>Within Groups</td>
<td></td>
<td>177</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>178</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>L2 CA</td>
<td></td>
<td>1</td>
<td>4.440</td>
<td>.037</td>
<td>.024</td>
</tr>
<tr>
<td>Within Groups</td>
<td></td>
<td>177</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>178</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>L2 Freq. of use</td>
<td></td>
<td>1</td>
<td>7.023</td>
<td>.009</td>
<td>.038</td>
</tr>
<tr>
<td>Within Groups</td>
<td></td>
<td>177</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>178</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>L2 Proficiency level</td>
<td></td>
<td>1</td>
<td>1.402</td>
<td>.238</td>
<td>.008</td>
</tr>
<tr>
<td>Within Groups</td>
<td></td>
<td>177</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>178</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note. $p < .008$. The overall alpha level (.05) was divided by the number of variables (6), and alpha was set at $p < .008$ in order to maintain an experiment-wise alpha level of $p < .05$.

6.1.6 Summary of the Findings

The participants in the current study carried various features. They are adult language learners at different ages and their contact with Chinese differed in the aspects of family heritage background, experiences of visiting regions speaking Chinese, length of formal Mandarin instruction at school, and individual L2 contexts. They also varied in their proficiency levels of Mandarin. Some were at the beginning levels and some were at advanced levels measured by their institutional status and the Mandarin EI test. They were significantly more willing to communicate in Mandarin inside than outside classrooms. In the both situations of inside and outside a Mandarin classroom, these
language learners were more willing to communicate in speaking and least willing to communication in comprehension skill.

The structure equation model was calculated through AMOS (Arbuckle, 2011). Model modifications were made based on the interpretations of Goodness-of-fit and modification indices. The final model finds that L2 CA directly predicts L2 PC, that is, learners with higher L2 communication anxiety would perceive themselves with lower competence in L2. In turn L2 PC predicts L2 WTC inside more than outside classrooms. Learners who believe they have higher L2 competence would be more willing to communicate in L2 both inside and outside classrooms. However, L2 PC is a stronger predictor in L2 WTC inside than outside classrooms. L2 WTC inside classrooms leads to L2 WTC outside classrooms. L2 WTC outside classroom is the antecedent of the learners’ frequency of use of the L2 and the frequency of L2 use leads to the learners’ actual proficiency levels. The more willing a language learner is to communicate in L2 inside a language classroom, the more willing he or she would be to communicate in that language outside the classroom. If a language learner is willing to communicate in L2 outside classroom, he or she would actually use L2 frequently and as a result his or her proficiency levels would improve correspondingly. At the same time, L2 PC also directly predicts L2 frequency of use and L2 actual proficiency levels, that is, a learner with higher L2 perceived competence would also actually use L2 more frequently and his or her actual proficiency levels would usually be higher. Learners reported higher or lower individual L2 use in the contexts of at home, at school outside the Mandarin classroom, in the community, and others, but the L2 PC remains a more important antecedent in L2 WTC comparing to L2 CA in L2 WTC. However, for the learners who reported higher
L2 use contexts, the weight of their communication anxiety in L2 WTC becomes heavier. Heritage language learners showed a tendency for lower L2 WTC than non-heritage language learners both inside and outside the classroom. They also indicated slightly higher levels of L2 CA than non-heritage learners. Heritage learners perceived themselves with slightly higher competence and they did have higher actual proficiency levels, but neither of these two differences were significant. As the non-heritage language learners reported slightly high L2 WTC, they did tend to use the L2 more frequently. The results observed for the direct comparison of heritage and non-heritage language learners are to some extent limited by the fact that in the heritage learner group (N=78), 6 heritage learners had a grandparent but no parent who spoke the heritage language, and this difference in family background may have affected the outcomes of the overall comparisons.

6.2 Discussion

6.2.1 L2 WTC Construct

Research question 1a. What is the relation of WTC with perceived competence and communication anxiety in L2?

The model shows that perceived competence directly influence the WTC both inside and outside a second language classroom. However, perceived competence weighs more in L2 WTC inside classroom than outside classroom. On the other hand, communication anxiety does not directly influence WTC, and this is true both inside and outside classroom. Instead, it influences WTC indirectly through perceived competence. The result confirms the conclusion brought by the previous studies (e.g., Clément et al.,
which stated that perceived competence is a stronger indicator of WTC than anxiety. As shown in the synthesis study of WTC in Chapter 2, the studies (Baker & MacIntyre, 2000; MacIntyre et al., 2003; Yashima et al., 2004) in low L2 use contexts reported higher correlations in perceived competence with WTC than communication anxiety. In general, the participants in the current study were all recruited from the United States, where Mandarin is learned as a foreign language. Therefore, the results conform to the previous studies that perceived competence has a stronger correlation with WTC in prototypically low L2 use contexts, such as foreign language settings.

Few previous studies distinguished the correlations of L2 WTC with other factors inside and outside the classroom. The study by MacIntyre and Doucette (2010) affirms the result obtained in the current study that both L2 PC and L2 CA are more correlated with L2 WTC inside than outside classroom. It confirms the importance for a language teacher to facilitate language learners with their perceived competence and reduce their communication anxiety in a language classroom. Perceived competence plays a more important role in WTC inside than outside classroom and communication anxiety directly influences perceived competence. Therefore, the higher correlation between L2 PC and L2 WTC inside classroom can be accounted for by the influence of communication anxiety to perceived competence. Young (1991) proposed a list of potential sources of language anxiety (p. 427): (a) personal and interpersonal anxieties, (b) learner beliefs about language learning, (c) instructor beliefs about language teaching, (d) instructor-learner interactions, (e) classroom procedures, and (f) language testing. Four among the six resources are directly related to classroom learning and teaching. Language anxieties
stemming from the first three sources may seem not to be directly attached to classroom learning, but are also frequently observed in Mandarin classrooms. In addition, as a less-commonly-taught language, Mandarin may actually provoke higher language anxiety than commonly-taught languages do. Le (2004) found that Mandarin learners possessed the highest anxiety level compared to the learners learning other languages in the previous studies using the Foreign Language Classroom Anxiety Scale (FLCAS) (Horwitz et al., 1986). The high level of language anxiety may be also well accounted for by Mandarin learners and instructors’ beliefs, as well as the interaction between them. The gap between the Chinese culture and western cultures, and the characteristics of Chinese culture itself may also contribute to high language anxiety.

Shi’s study (2007) reported the strongest factor provoking the learners’ anxiety was the belief that Mandarin was a difficult language, especially vocabulary and grammar. Le (2004) also reported an overall belief among learners that Mandarin is a “very difficult” or “difficult” language. Mandarin is an isolating language which contains no inflections or case markers for learners to figure out sets of rules to follow (Norman, 1988). When learners find they need more time and effort to learn the grammar in Mandarin, not to mention to speak it fluently, they may become anxious. The four tones in Mandarin generate an almost “mission impossible” for accurate pronunciation, especially at the early stage of learning. According to Le (2004), actually a higher percentage of learners valued an excellent pronunciation in Mandarin than that of American students learning German, French and Spanish. As a result, the difficulty caused by the tone system in Mandarin increases anxiety of learners who believe in the accuracy pronunciation in speaking. Instructors’ beliefs in language teaching are also a
major source of Mandarin learners’ language anxiety. A Mandarin teaching classroom, especially one with an instructor of Chinese culture background, may possibly be influenced by the Chinese value system, where teachers play the role of the academic authority in the classroom (Lai, 1994). Therefore, instructors in Mandarin classrooms may tend to believe the necessity of intimidation in teaching and more strongly value learners’ performances on quizzes and tests. Instructors influenced by the teaching system in China may tend to act as a “drill sergeant” instead of a facilitator. The classroom is likely to be more teacher-centered, instead of learner-centered, and calling on students to answer a question is possibly a typical way for a teacher to ensure attention from the students. These potential factors may contribute to high language anxiety in a Mandarin learning classroom.

As a result, Mandarin classrooms may generate higher communication anxiety, which leads to lower perceived competence and in turn, perceived competence influence WTC inside classroom more than outside classroom.

Research question 1b. What is the relation of WTC with participants’ self-report frequency of use of L2?

It is interesting to find out that L2 WTC inside classroom highly predicts L2 WTC outside classroom. Only L2 WTC outside classroom leads to frequency of use of the language. The higher the WTC outside classroom is, the more frequently language learners use the language. The result confirms the claim by MacIntyre and Doucette (2010) that it is important to promote L2 WTC in the classroom, which prepares learners to get ready for using the language outside of the classroom. The finding also supports
the ultimate goal of language classroom teaching that language learners are expected to actually use the language to communicate with people in real life.

On the other hand, L2 perceived competence also highly directly predicts language learners’ frequency of L2 use. The participants who believe themselves with higher competence of the language tend to actually use the language more frequently. McCroskey and McCroskey (1988) stated that many decisions made by people in communication are actually decided by self-perceived competence, instead of actual one (p. 110). Therefore, it explains why people who believe they have the competence to perform the communication task would actually carry out the communication.

Research question 1c. What is the relation of WTC with participants’ proficiency levels in L2?

L2 WTC is not an immediate antecedent of learners’ proficiency levels or, conversely seen, higher L2 proficiency is not a direct outcome of higher WTC levels. Instead, WTC may have an impact on the proficiency levels indirectly through frequency of L2 use. As expected, high frequency of language use leads to better proficiency levels. The more a language learner report to use the language, the higher proficiency level the learner exhibits. Theoretically speaking, we may want to generalize and say that the more someone uses a target language, the higher proficiency someone will attain. In order to increase the proficiency level, therefore, language learners need to maximize their opportunities to use the language. And as discussed above, enhancing learners’ L2 WTC outside classroom would help them to actually take action to grasp every possible opportunity to use the language. At the same time, learners’ perceived proficiency levels
directly predict their actual proficiency levels. This may be also the result of increased frequency of use caused by high perceived competence.

Research question 1d. What are the differences on relation of L2 WTC with perceived competence and communication anxiety in varied individual L2 use contexts?

The perceived competence is significantly correlated to L2 WTC in both low and high individual L2 use contexts. L2 CA is significantly correlated to L2 WTC only in high use context. In both contexts, the perceived competence is the factor that weighs more than L2 CA in L2 WTC inside and outside classroom. However, even though the individuals’ L2 use contexts vary, all the participants were recruited from the context of learning Mandarin as a foreign language. Therefore, in a broad sense, the participants are all in a low L2 use context. The result actually follows the conclusion in the previous studies (e.g., Clément et al., 2003; MacIntyre et al., 2002; Yashima et al., 2004) that L2 PC is a stronger predictor of L2 WTC in low use context. On the other hand, the correlation between L2 CA and L2 WTC has shown an increasing trend from low to high individual L2 use contexts. When L2 use context for language learners continue to increase and language learners have opportunities of using the target language in a second language context, L2 CA may become a stronger predictor than L2 PC. The result yielded in the present study did not disapprove the previous finding that L2 CA is a stronger predictor in high L2 use context (Baker & MacIntyre, 2000; MacIntyre et al., 2003).
6.2.2 Heritage and Non-heritage Language Learners of Chinese

a. What is the difference of L2 WTC between heritage and non-heritage language learners of Chinese?

b. What are the differences of L2 perceived competence and L2 communication anxiety between heritage and non-heritage language learners of Chinese?

c. What are the differences of L2 frequency of use and actual proficiency levels between heritage and non-heritage language learners of Chinese?

The heritage language learners show a tendency for lower L2 WTC than non-heritage learners. It was hypothesized that heritage language learners would have higher L2 WTC than non-heritage language learners, because the former were presupposed to have higher L2 contact opportunities and higher proficiency levels. The results show that heritage language learners do have somewhat higher self-perceived competence and actual proficiency levels, but the differences between them are not significant. These non-significant differences between them may be explained by the definition of heritage language learners in this study. The present study defined all the language learners who have a family or family members speaking Chinese at home as heritage language learners. This definition does not restrict heritage language learners as those who have certain proficiency levels prior to the formal instruction of Mandarin at school. In fact, for both the heritage and the non-heritage group the sampling by expected proficiency was purposefully wide (from first-year courses to third-year courses) in order to ensure that there would be very large variability and full (rather than truncated) ranges of proficiency in the full sample as well as within each group of both non-heritage and heritage learners. As a result, the heritage language learners in this study include those who may start from
the very beginning at their Mandarin classes at school. In addition, Kondo-Brown (2005) showed that those who have only grandparents speaking the heritage language at home were more similar to the non-heritage language learners in terms of affective and linguistic measures than to heritage language learners who have at least one heritage-language-speaking parent. In the present study, only six heritage language learners fall into the subgroup of heritage background with grandparents and ANOVA analyses showed non-significant differences between them and the rest of the heritage language learner subsample, so the study did not distinguish them from other heritage language learners. Nevertheless, the inclusion of even this small portion of learners in the heritage group might have dragged down the average performance of the overall heritage language learner sample, making the difference statistically non-significant when it may have been important in observations from a more tightly homogeneous group of heritage learners. The other possible reason may be because the range of proficiencies within the heritage and within the non-heritage groups were purposefully wide, we cannot expect that just membership as heritage or non-heritage trumps differences in proficiency (e.g., enrollment in first year vs. third year) and instead, in this sample, proficiency trumped learning background (i.e., heritage or foreign learning profiles). The above factors explain why the heritage and non-heritage language learners perform with non-significant difference on the Mandarin EI task (see Table 6.27).

On top of the non-significant difference of actual proficiency levels between the two learner groups, the learners in the heritage group also didn’t have significantly higher perceived competence than the learners in the non-heritage group. The results on L2 WTC construct prove that language learners’ perceived competence is highly correlated
with their actual proficiency levels. Therefore, the non-significant difference of perceived competence between heritage and non-heritage language learners corresponds to the findings on their actual proficiency levels. On the other hand, the differences of the cultural values between the two regions also play a role in the result. Asian cultures value modesty and inwardness (Le, 2004). The heritage language learners inheriting at least part of Chinese culture from their family may tend to be modest and perceive themselves with less competence. Non-heritage language learners in this study come from American culture which values self-confidence. They may experience various difficulties and anxieties in learning Mandarin but still be able to perceive themselves with relatively similar competence with heritage language learners.

The tendency for a higher L2 CA of heritage language learners may contribute to the lower L2 WTC of heritage language learners. The high communication anxiety of heritage language learners is supported by the previous findings. It was widely discussed in the previous studies (e.g., Carreira & Kagan, 2011; Comanaru & Noels, 2009; Lacorte & Canabal, 2003; Le, 2004; Li, 2006) that heritage language learners might feel anxious because of high expectations from teachers, classmates, parents, and themselves. Language teachers and classmates usually believe that it is relatively easier and effortless for a heritage learner to learn the language because he or she has already started hearing or using the language at home. Therefore, they would expect that heritage learners spend less time and achieve higher proficiency levels than other students. The heritage language learners in the present study include those with heritage background of Chinese dialects. Hsiao (2010) pointed out that those dialect heritage learners also felt high expectations from teachers and classmates, but they experienced difficulties and interferences brought
by their dialects in learning Mandarin. The high expectations imposed on them interact with the interferences to generate high anxiety levels in Mandarin learning. Parents of heritage language learners receive at least partial influence from Chinese culture. An important component in Chinese culture is the belief and value invested in education (Li, 2006). Therefore, Chinese parents may set high expectation for their children in school education, including Mandarin learning. At the same time, heritage language learners also set a high goal for themselves. Carreira and Kagan’s survey (2011) found out that majority of heritage language learners wished to use Mandarin in their professional life in future. These high expectations proposed in the previous studies sufficiently support the finding of high anxiety levels in heritage language learners in the present study.

In Chapter 5, the comparison of individual L2 use contexts (that is, current levels of use of Mandarin reported in the last item of Part 1 of the survey) between heritage and non-heritage learners of Mandarin shows heritage learners have significantly higher L2 use contexts at home and in the community, but significantly lower L2 use contexts at school outside classrooms. It is also interesting to find out that even though a non-significant difference was found in their frequency of L2 use (levels of current use of Mandarin according to WTC-specific scenarios self-reported in Part 5 of the survey) in the three different modes of occasions and three types of interlocutors, heritage language learners actually use the language slightly less frequently. This finding confirms the result obtained in the L2 WTC construct that the frequency of actual use the language is determined by whether or not learners have the willingness to use the language. If a person has low L2 WTC, he or she will tend to hesitate or be reluctant to use the
language even when plenty of opportunities are provided. This, in turn, will negatively affect the development of proficiency.

6.3 Summary

The present study confirms the findings in previous research. L2 perceived competence is a stronger predictor in L2 WTC in low L2 use context. L2 communication anxiety contributes to learners’ perceived competence directly, and the correlation between communication anxiety and L2 WTC increases as the L2 use contexts change from low to high end. L2 WTC outside classroom is decided by L2 WTC inside classroom and leads to L2 frequency of use. Therefore, it is important for a language teacher to enhance language learners’ WTC inside a language classroom. The more frequently a learner uses the language, the higher proficiency levels one can achieve. In order to increase L2 WTC and frequency of L2 use, and finally improve learners’ proficiency levels, it is important to help learners to believe themselves with higher perceived competence. One way to achieve this purpose is to reduce learners’ anxiety levels of the language. The sources of language anxiety are various, including the six sources proposed by Horwitz (Horwitz et al., 1986). The unrealistic beliefs of learners and instructors, and the interaction between them are the factors contributing most to the high language anxiety in a Mandarin classroom. Chinese culture can be the additional source for high language anxiety in Mandarin learning.
CHAPTER 7
CONCLUSION

The present study supports the previous findings regarding language learners’ L2 WTC and its antecedents and outcomes. It also proves the importance of enhancing learners’ L2 WTC in a non-native language classroom. The study also provides significant findings regarding heritage and non-heritage language learners of Mandarin on their L2 WTC, its antecedents, and its outcomes. This chapter will review the significance of the present study, provide pedagogical implications to improve language learners’ willingness to communicate, point out limitations of the present study, and finally propose suggestions for future studies.

7.1 Significance of the Present Study

The findings in the present study support the results in the previous studies. It lends empirical support to the claim that a language learners’ perceived competence directly predicts their willingness to communicate and their communication anxiety contributes to the willingness to communicate indirectly through perceived competence. It also supports the claim that in low L2 use contexts, learners’ perceived competence is a stronger predictor in L2 WTC construct and the contribution from L2 communication anxiety to L2 WTC increases along with the increased individual L2 use context. The present study further proves the significance of L2 WTC inside a language classroom by showing that L2 WTC ultimately leads to higher proficiency level of a language learner. It shows that L2 WTC inside classroom would enhance the L2 WTC outside the
classroom. When a language learner feels willing to communicate outside the language classroom, he or she would actually communicate in the language and the proficiency level of the language learner would improve due to frequent practice in the target language. Learners’ self-perceived competence is found to be important in directly relating to their frequency use of the language and actual proficiency levels. The present study supports and extends the L2 WTC model in previous research. It not only confirms the importance of increasing language learners’ perceived competence and reducing communication anxiety in learning a foreign language, but also proves the importance for language teachers to enhance language learners’ willingness to communicate in the L2 classrooms as it will lead to their willingness to communicate in L2 outside the classrooms and eventually the improvement of language learners’ proficiency levels.

The study provides valuable insights regarding heritage and non-heritage learners of Mandarin, although tempered by the need to acknowledge that the precision of the results in this part of the study may be less than ideal (because of the 6 heritage learners with only a grandparent who were included in the total sample of 78 heritage learners with at least one parent). Conversely, it may need to be acknowledged that the findings could be more directly relevant and representative of heritage language learners with at least one parent who speaks the heritage language, that is, the profile of heritage learners who, according to Kondo-Brown (2005) is likely to exhibit higher confidence and stronger linguistic skills overall than other heritage profiles involving grandparent only connections with the ancestor language. Future studies with large enough samples of maximally homogenous heritage language learning profiles will need to revisit the
question of whether (and which) different profiles of heritage language learners are different from foreign language learners of Mandarin.

Based on the available evidence gleaned in the study, it is posited that heritage and non-heritage learners of Mandarin share non-significant differences in their perceived competence and actual proficiency levels. The definition of heritage language learners and the differences between two culture values were discussed to explain the results. The heritage learners reported a tendency of lower frequency of L2 use, L2 WTC, and higher communication anxiety in Mandarin learning than the non-heritage learners. It can be explained by the fact that heritage language learners of Mandarin mostly only have heritage background of Chinese dialects. The tendency for low frequency of L2 use and L2 WTC can be accounted for by the high communication anxiety among heritage learners of Mandarin, even though the difference is not significant. The findings may contradict to the assumption of many language teachers and learners. Therefore, it is crucial for language teachers to realize the importance of helping heritage language learners to reduce their anxiety levels in learning Mandarin.

Another contribution of the present study is developing the Mandarin EI task based on the EI tasks by Ortega et al. (2002). The design of the Mandarin EI task successfully minimized the three effects proposed in previous literature that may diminish the reliability of EI tasks. The Mandarin EI in the study proved to be reasonably reliable and valid in measuring a second language learner’s global competence of the language for the norm-referenced testing and research purpose of distributing widely ranging proficiency individuals along a continuum of global oral competencies. The EI
task provides a convenient and quick proficiency measure for future researchers in their studies on second language learners of Mandarin.

7.2 Pedagogical Implication

Enhancing language learners’ WTC in a second language classroom helps to increase their WTC outside the classroom. When language learners are more willing to communicate in the second language outside the classroom, they will actually use the language more frequently. Practice makes perfect. Therefore, when they grasp more opportunities to actually use the language, their proficiency levels increase. In order to increase language learners’ WTC inside a second language classroom, teachers need to help learners to increase their self-perceived competence. At the same time, when learners believe they have enough competence to use the language, they would actually use the language and their proficiency levels would increase. Lewin (1951) proposed that there are two types of forces: Driving forces and Restraining forces. Lewin noted that it is easier to modify a person’s action by reducing the restraining forces than increasing the driving forces. Therefore, in L2 WTC construct, self-perceived competence is the driving force and communication anxiety is the restraining forces. In order to increase learners’ self-perceived competence and enhance their L2 WTC, it may be easier for teachers to reduce language learners’ communication anxiety in a classroom.

7.2.1 Identifying Sources of Language Anxiety

Based on the understanding of language anxiety, identification of its sources in a foreign or second language classroom is essential in finding ways to cope with language
anxiety and enhance learners’ perceived competence and L2 WTC. Language anxiety generally relates to the difficulties and frustration learners experienced during the process of language learning (MacIntyre, 1999). At the early stage of language learning, learners may encounter various kinds of difficulties and problems in pronunciation, vocabulary, grammar, and other areas. Facing these difficulties and problems, some learners may develop anxiety, which leads to poor performance, and poor performance may in turn generate more anxiety. This forms a vicious circle.

Young (1991) proposed a list of potential sources of language anxiety (p. 427): (a) personal and interpersonal anxieties, (b) learner beliefs about language learning, (c) instructor beliefs about language teaching, (d) instructor-learner interactions, (e) classroom procedures, and (f) language testing. Personal and interpersonal anxieties relate to learners’ low self-estimate, low self-perceived ability levels, and competitiveness between the learner and others (Young, 1999). As indicated in the present and previous findings, self-perceived ability levels are closely related to language anxiety. Learners with self-perceived low ability levels are found to possess higher anxiety levels (e.g., Clément et al., 1994). Bailey (1983) emphasized the impact of competitiveness on language anxiety. She pointed out that over self-comparison of language learners with others or with self-expectation would develop anxiety.

The unrealistic beliefs of learners about language learning arouse language anxiety when the beliefs come into collision with reality. Learners in Price’s (1991) study believed that language learning needed a special aptitude. Some learners put great stress on speaking with an excellent accent, or are over confident that two year is enough to speak a language fluently (Young, 1991). Also some students in Chen and Chang’s (2004)
study believed the target language was too difficult and they could not handle it. These beliefs provoke anxiety in language learning. Instructors’ beliefs about language teaching may also provoke anxiety of language learners. Chen and Chang (2004) reported that instructors believing in the teaching styles of lecturing, talking fast, and refraining from using the blackboard would make students anxious. Instructors who believe that students need to be corrected constantly at any of their mistakes may also easily provoke language anxiety among students. Students may become anxious when their instructor takes the role more like a drill sergeant than a facilitator (Young, 1991). The interaction between instructors and learners contributes to language anxiety, too. The interaction which most easily provokes anxiety is the manner of error correction of instructors (Young, 1991). Correcting errors in front of the class or making learners sound “dumb” will easily make them anxious. Instructors need to find a less intimidating way for error corrections, especially with anxious students. A friendly and relaxing classroom atmosphere also helps to alleviate language anxiety and increase learners’ willingness to use the language and interact with the instructor.

Classroom procedures which may provoke anxiety includes having students speak in the target language in front of the class, giving frequent oral quizzes, and calling on students to answer questions unprepared (Young, 1999). Language testing is also a possible source of language anxiety. When students find the test content is not what they have been studying for, or test tasks are unfamiliar to them, they may become anxious (Young, 1999).

Competitiveness can be one of the major sources of language anxiety for learners without Chinese background in the same classroom as heritage learners. Heritage learners
usually speak some dialect of Chinese or have contact with Mandarin speakers in their families. Kondo-Brown (2005) found a substantial better performance of heritage learners with parents speaking the target language in grammatical knowledge and the skills of listening and reading compared to non-heritage students. This fact is very likely to provoke high language anxiety among non-heritage learners sitting in the same classroom. Nevertheless, on the other hand, as discussed previously, heritage language learners, especially those with dialect background or with zero proficiency levels prior to formal Mandarin classes, experience high anxiety levels due to high expectations from teachers, classmates, parents, and themselves and frustration caused by interferences from their dialects in learning the language.

Culture background contributes to the high language anxiety in Mandarin learning, especially in high contact of the target language and its culture (Le, 2004). Learners without an East Asian background may find Chinese culture is far different from their own cultures, which may arouse additional anxiety, especially when they find they have to deal with the language learning and cultural adjustment at the same time, but the anxiety may be reduced along with the increased contact with the target language and the familiarization with the culture. While for learners with an East Asian background, the background itself may arouse high anxiety (Le, 2004). East Asian cultures value the characteristics of shyness and inwardness, and attach great importance to protecting one’s face, that is, public image (Peng, 2007). The value system in East-Asian cultures may influence learners of such background to be easily anxious in language learning.
7.2.2 Alleviating Language Anxiety in Mandarin Classrooms

7.2.2.1 Alleviating language anxiety on proposed sources. In order to alleviate learners’ language anxiety stemming from personal and interpersonal factors, it is necessary to first make learners aware of the existence of language anxiety in the classroom and directly address this problem. Crookall and Oxford (1991) suggested addressing anxiety directly to demonstrate an attitude “which considers that both students and teacher can share the problem, and both can deal with it together” (p. 144). The simplest method to achieve this purpose is to adopt a questionnaire on anxiety, such as FLCAS items by Horwitz et al. (1986), and let the students do the questionnaire either on the first day of the class or sometime during the semester. From these items, learners can realize the existence of anxiety in a language classroom and feel easier to address and discuss this problem with the teacher or the peers. Following the questionnaire, the instructor can share the results with the class to talk about which items may be most likely to arouse language anxiety among them. This activity helps learners to understand that experiencing anxiety is common in a language classroom and their classmates have the same feeling as they do. Competitiveness is also one of interpersonal factors. It is helpful if instructors can help learners to establish practical and short-term goals. Comparing with the short-term goal may release language anxiety caused by competitiveness. Another effective way to help learners to reduce the sense of competition and build confidence in a Mandarin classroom is to separate heritage and non-heritage learners into two different classes.

Helping learners to develop realistic beliefs of language learning is also a way to cope with language anxiety. Campbell and Ortiz (1991) suggested an activity named
“The myths and the realities of foreign language learning” in their Foreign Language Anxiety Workshop. It is a questionnaire containing items examining a number of popular notions about the study of a foreign language. However, some items especially addressing Mandarin learning should be added, such as the beliefs about accuracy of tones and Chinese orthography. Students need to know that language learning is a process and it is inevitable to make mistakes on pronunciation, including tones, and Chinese characters. Instructors, also as language learners, can share their own language learning experiences with the students. I always share my experience of learning English with my students to encourage them to learn and speak without fearing of making mistakes.

Developing instructors’ beliefs also helps to alleviate learners’ language anxiety in Mandarin classrooms. The first thing an instructor can do is to understand the importance of a learner-centered classroom and adopt a role of a facilitator instead of a drill sergeant. Although the Chinese saying says “a good apprentice always has a strict master”, instructors in Mandarin classrooms need to understand that mistakes are part of language learning and allow students to make mistakes. Instructors need to emphasize more on conveying meaning than grammatical accuracy, especially in speaking and writing practices. Mandarin instructors also need to develop appropriate beliefs and expectations on heritage language learners. It is important to understand that heritage learners of Mandarin may only possess background of a Chinese dialect and such background may contribute less than expected into their Mandarin learning.

Regarding methods for giving feedback, instructors may need to give students more positive feedback and avoid explicitly correcting students in front of the class. Recasts can be a less anxiety-provoking way to give feedback in front of the class.
According to Nicholas, Lightbown, and Spada (2001), recasts are a type of implicit feedback which repeats a learner’s utterance by providing the correct forms of the target features and at the same time maintain the central meaning contributed by the learner. Therefore, recasts may provide necessary feedback, minimize the possibility to provoke language anxiety, and maintain the conversation flows at the same time. The activity of Mistake Panel from Crookall and Oxford (1991, p. 147) can be another good way to handle mistakes and help learners to adopt an amusing attitude at the errors. They suggested either the teacher or the students collect mistakes over a few class sessions or make up mistakes. Each mistake is written on one card. Then the learners are divided into groups and start rating each mistake based on the following criteria: amusement (how funny they are), creativity (how original they are), logic (how reasonable is the mistake in terms of the language), communicative intelligibility (how well can one still understand the message, despite the error), and learning contribution (how much can be learned from this error). The game helps learners to learn to be at ease with errors and at the same time also learn from these errors.

Classroom procedures can also be changed to minimize language anxiety. Pair or group works may help to reduce language anxiety. Long and Porter (1985) conducted a study on group work and confirmed that group work helped to promote a positive affective climate in a language classroom. Learners may feel less competitive when working in pairs or groups, and feel less anxious when reporting in class because they may feel they are representing the group. Quizzes and exams in language classrooms are usually achievement tests to examine whether the learners have acquired what they are supposed to learn. Therefore, it is important to always test what have been taught in class.
Donley (1999) suggested always informing students of the format of a test and percentage of points of each part. Learners may feel less anxious when they know what and how they are going to be tested.

Explaining Chinese culture to learners may shorten the time and reduce the difficulty to get accustomed to the target culture in the context of learning Chinese as a second language. It is also important for instructors to understand the potential high anxiety caused by heritage learners of East Asian background, and work more on alleviating language anxiety at various aspects as mentioned above and below.

**7.2.2.2 Alleviating language anxiety on grammar and vocabulary.** Le (2004) and Shi (2007) reported that grammar and vocabulary were the two areas which Mandarin learners attach most importance to. Tobia’s (1980) model of three stages showed that language anxiety started to take effect at the input stage and acted as a filter. Simple input may reduce language anxiety and allow input to enter into the processing stage as much as possible. Lee and Van Patten (2003) suggested the characteristics of simplified input for second language learners were slower rate, high-frequency vocabulary, and repetition of grammar structures. Therefore, when a new vocabulary or grammar pattern is introduced, follow-up activities need to be employed more than once. Visual aids also provide additional exposure of vocabulary and grammar. Instructors can choose to write on a blackboard, prepare handouts, or use computer aids. Other reasons for grammar to cause anxiety are too much material, frequent grammatical errors, and/or lack of grammar instruction (Van Pattern & Glass, 1999). Grammar material needs to be reduced if learners become anxious over the overwhelming material they have been given. Again, it is important for instructors to deal with grammatical errors wisely. Recasts and
the Mistake Panel mentioned above can be two good ways to handle the errors and minimize language anxiety. On the other hand, lacking grammar instruction can provoke anxiety, too. Most students believe that grammar is necessary and beneficial and may become anxious if instructor does not provide grammar instruction. Therefore, in a Mandarin classroom where communicative teaching method dominates, grammar input is also necessary and important.

Koch and Terrell (1991) asked the learners to rate classroom activities based on the feeling of comfort or nervous. Using pictures to present vocabulary and acquisition of vocabulary through association are the two techniques rated as comfortable. Pictures may provide additional help by building a connection between the concrete image and the abstract Chinese orthography, especially for visual learners. Take the Chinese word 豐盛 fengsheng ‘sumptuous’ as example, a picture of a feast with a full table of dishes may help learners to associate the orthography with the image and easily recall its meaning. Another way to learn vocabulary is to help learners to establish association among different characters. For example, for the word 脸 lian ‘face’, the radical 月 means flesh or body parts. Instructor can associate it with other previously learned words carrying the same radical and meaning body parts, such as 肚 du ‘stomach’, 腹 fu ‘belly’, 腿 tui ‘leg’, 脚 jiao ‘foot’. A simple explanation of etymology may help learners to understand the character and remember the orthography. The character 豐 feng ‘harvest’ is composed of 山 shan “mountain” and double 方 feng “grains” at the top, meaning mountains of grains, and 豆 dou ‘a vessel’. The whole character means mountains of grains in a vessel; therefore, it is a harvest (Wieger, 1965). The etymology may also help learners feel
easier and more interesting to learn characters, and may alleviate possible language anxiety in vocabulary learning.

7.2.2.3 Alleviating language anxiety on four skills. Listening is one of the areas that are most likely to provoke anxiety. Campbell (1999) reported that both males and females were more anxious about listening than other skills. A slower speaking rate is an effective way to reduce language anxiety in listening. Visual aids as mentioned above can also help to alleviate anxiety. Another effective way to reduce language anxiety in listening is to provide pre-listening activities to make input comprehensible (Vogely, 1999). The pre-listening activities can be a summary or outline of the text to be heard, a skim of the comprehension questions before the listening, or a discussion about what learners have known about the topic. For example, before listening to the passage about insurances in America (Liu & Yao, 2006, p. 259), an instructor can assign the two tasks as a pre-listening activity: (a) Please work in pairs and talk about what kinds of insurances you or your family have purchased; and (b) please discuss whether it is necessary to purchase these insurances, and Why? The pre-listening activity may recall learners’ memory of the related vocabulary and sentence patterns, which helps to ease the listening and reduce language anxiety. Vogely (1999, p. 115) also listed some sample listening comprehension strategies to facilitate listening and reduce anxiety: (a) guess the meaning of words through context, (b) use existing knowledge of the world, (c) make use of visual cues, (d) recognize cognates, (e) listen for specific information, (f) listen for the main idea, (g) recognize the structural format of the discourse, (h) recognize the function of the discourse, and (i) recognize transitional words and phrases.
Speaking easily provokes anxiety, but, on the other hand, it is an important aspect in language learning. Instructors sometimes feel at a dilemma of increasing speaking opportunities and reducing language anxiety. In order to make learners feel at ease when speaking, it is important to create a sense of a community in the class for learners (Price, 1991). Opportunities for learners to get to know each other may help them to establish cohesion and feel they are working together as one community. Group work can be another way to practice speaking and at the same time to provide a secure feeling for learners. Representing a group to speak in class reduce the feeling of being confronted and evaluated. The time allowed learners to prepare before speak up in class also helps to reduce language anxiety. In speaking, instructors need to focus more on what are being conveyed than the correctness of pronunciation and grammar. Overt error corrections may increase anxiety and interrupt the streams of communication. Instructors can employ recasts to give feedback and at the same time keep the flows of conversation. Overt corrections of tones will also increase anxiety. Sanders (2008) claimed that Chinese teachers should try to provide students with “the widest currency of acceptance throughout the Mandarin-speaking world” (p. 15). Therefore, again emphasis needs to be put more on meaning, instead of the forms of pronunciation.

To alleviate language anxiety in writing, instructors need to emphasize the process instead of the outcome of writing. It is a good way to teach students to start from brainstorm and outlines to help them to get prepared for writing tasks (Leki, 1999). Two or more drafts for one writing task may make learners feel less anxious about evaluation, as they can improve themselves through drafts. Allowing students to write pinyin occasionally for the words they don’t recall helps learners to focus on meaning
expression and less worry about the forms. Instructors can provide options of topics for learners to write, so that they can choose a familiar and interesting topic and less worry about not having enough to write about. Computer may also be a way to write with low language anxiety. Chatting rooms, MSN can allow learners to express their opinions in an informal way without feeling being confronted and evaluated based on the correctness of forms.

Reading may provoke anxiety, too. Students sometimes complain that they understand each vocabulary, but fail to understand the sentences or passages. This may be because language anxiety blocks their processing of the meaning of the sentences. Reading is not just answering comprehension questions (Lee, 1999). Reading skills are one way to enhance learners’ reading competence and reduce anxiety. Instructors can first let students to identify key words in each paragraph of the reading sample, and then ask them to use one sentence to summarize the paragraph. The learners can also discuss in groups about the key words and sentences they have written and learn from each other.

7.3 Limitations of the Study

The participants in the present study were recruited from the universities in the United States. Even though their individual L2 use contexts vary, the overall L2 use context of the participants is the same. They are learning the language as a foreign language and do not really have too many opportunities to use the language in daily life. Therefore, even though the study found that L2 CA contributed more in L2 WTC when the individual L2 use contexts increased, L2 PC remains the stronger predictor in L2 WTC in the present study. As a result, the study cannot provide direct support on the
previous findings that L2 CA is a stronger predictor in L2 WTC construct when participants are learning the language in high L2 use context.

The study did not recruit large $n$ size of participants who have only a grandparent or grandparents speaking the target language at home. Therefore, the differences between subgroups of heritage language learners proposed by Kondo-Brown (2005) were not detected. The present study adopted an all-inclusive definition of heritage language learners of Chinese. The background information survey did not distinguish those with dialect heritage background from those with Mandarin heritage background. The participants who have family members speaking a Chinese dialect were all defined as heritage language learners. Even though this is the commonly adopted method and makes participants recruiting procedure easier, it may lead to some discrepancies or lack of precision in the results.

There were two main instruments in the literature to measure L2 WTC. Even though most studies used McCroskey’s WTC scale, the present study used the scale by MacIntyre. In order to make the data collection procedure feasible for the participants, the present study did not use both scales and did not have the opportunity to compare the two scales. In addition, the frequency of L2 use is measure by a self-reported scale. Actual observed frequency of L2 use of language learners would be more accurate. However, because of the large $n$ size of the data samples, it is impossible for the researcher to observe their actual frequency of L2 use.
7.4 Suggestions for Future Studies

The results yielded in the present study illustrated the L2 WTC construct in low L2 use context. Future studies can recruit participants from Mandarin-speaking regions and compare whether the L2 WTC and its relation to the affective factors are the same in high L2 use context. Futures studies can distinguish the groups of Chinese dialects and Mandarin family background by adding a relevant question in the background survey, and compare differences between heritage language learners who have background of Mandarin-speaking families and those whose family members speaking a Chinese dialect to find out whether these two types of heritage language learners share the same pattern in the L2 WTC construct. Furthermore, future studies can examine the differences between two subgroups of heritage language learners, those with grandparents speaking the language and those with parents speaking the language. If feasible, future studies can employ two types of L2 WTC scales and investigate any differences caused by different scales. It will provide suggestions for future studies of how to choose the appropriate scale to measure L2 WTC. Future studies can also employed observed frequency of L2 use to test whether the L2 WTC construct would remain the same. The Mandarin EI task provides a quick and reliable measure for assessing global proficiency levels of Mandarin learners for research needs such as spreading individuals along a continuum of proficiency or testing proficiency differences between study groups. However, future studies can improve the six items with ID lower than .19 to further improve its reliability and validity. Furthermore, the present study used a point-system to qualitify the L2 contact that a participant has experienced. The system was developed out of researcher intuition. Therefore, future studies should make effort to establish a more reliable system...
to compare learners’ L2 contact. Last but not least, since the present study proves the importance of L2 perceived competence, future studies can explore more antecedents of L2 perceived competence and provide more suggestions for language teaching on enhancing L2 perceived competence, L2 WTC, and eventually improve language learners’ proficiency levels.
APPENDIX A

CODING VARIABLES FOR L2 WTC

I. Publication characteristics
   A Author(s):
   B Journal:
   C Year:
   D Type of publication:
      1. Journal article
      2. Dissertation
      3. Book chapter

II. Participants
   A Total $n$ size _________________
   B Proficiency levels of participants
      0. Zero knowledge
      1. Beginning
      2. Intermediate
      3. Advanced
   C Proficiency assessment types
      0. Impressionistic judgment
      1. Institutional status
      2. In-house assessment
      3. Standardized test
      4. Self-assessment
   D School status
      0. Primary school
      1. Secondary school
      2. College

III. Research Design
   A First language _________________
   B Target Language _________________
C  Experiment venue________________
D  L2 use context
   0. High L2 use
   1. Low L2 use

IV. Instrument
A  WTC Scale
   0. McCroskey’s scale
   1. MacIntyre’s scale
   2. Other _______________
B  Perceived competence scale
   0. McCroskey’s scale
   1. MacIntyre’s scale
   2. Other _______________
C  Communication anxiety scale
   0. McCroskey’s scale
   1. MacIntyre’s scale
   2. Horwitz’s scale
   3. Other _______________
D  Frequency of L2 communication
   0. self-report
   1. observation

V. Research purposes
A  To examine the relation between L2 WTC and its antecedents;
B  To examine the relation between self-reported WTC and self-reported or observed communication behavior in L2;
C  To examine the differences of WTC in L2 low and high use context;
D  To examine the effects of sex, age, and proficiency levels on L2 WTC;
E  To examine the differences of WTC in L1 and L2.
APPENDIX B

CODING VARIABLES FOR EI TASK

I. Publication characteristics
   A  Author(s):
   B  Journal:
   C  Year:
   D  Type of publication:
       1. Journal article
       2. Dissertation
       3. Book chapter

II. Participants
   A  Total $n$ size _________________
   B  Proficiency levels of participants
       0. Zero knowledge
       1. Beginning
       2. Intermediate
       3. Advanced
   C  Proficiency assessment types
       0. Impressionistic judgment
       1. Institutional status
       2. In-house assessment
       3. Standardized test
       4. Self-assessment
   D  School status
       0. Primary school
       1. Secondary school
       2. College
III. Research Design
   A. First language
   B. Target Language

IV. Research questions:

V. Elicited Imitation task design
   A. Total items
   B. Sentence Syllables
   C. Administration: (how the stimuli are presented)
      0. Oral
      1. Taped
      2. Oral & Taped
      3. Missing
   D. Use of ungrammatical sentences
      0. Yes - Instruction given to correct sentences
         a. Yes
         b. No
         c. missing
      1. No
   E. Pause between stimuli and repetition
      0. Yes
      1. No
      2. Missing
   F. Length given to repeat the sentences
   G. Targeted grammatical features
   H. Other special features about EI:
      I. Purpose
      J. Other measures:

VI. EI reliability

VII. EI scoring
   a. Score range:
   b. Scoring criteria:

VIII. Conclusion
APPENDIX C

MANDARIN REPETITION TASK¹

You are going to hear several sentences in English. After each sentence, there will be a short pause, followed by a tone sound {TONE}. Your task is to try to repeat exactly what you hear. You will be given sufficient time after the tone to repeat the sentence. Repeat as much as you can. Remember, DON'T START REPEATING THE SENTENCE UNTIL YOU HEAR THE TONE SOUND {TONE}. Now let's begin.

I like flowers.
I am writing a letter.
I don’t think I need a big car.
As it is raining, I don't go out.
The little girl hurt herself and started to cry.
As soon as I returned home, I watched TV with my sister.

That was the last English sentence.

Now, you are going to hear a number of sentences in Mandarin. Once again, after each sentence, there will be a short pause, followed by a tone sound {TONE}. Your task is to try to repeat exactly what you hear in Mandarin. You will be given sufficient time after the tone to repeat the sentence. Repeat as much as you can. Remember, DON'T START REPEATING THE SENTENCE UNTIL YOU HEAR THE TONE SOUND {TONE}. Now let's begin.

1 我得去剪头发了。
2 红色的书在桌子上。
3 这个城市街道很宽。
4 他每天早上都要洗澡。
5 我听说明天可能会下雨。
6 你刚才说你今天在做什么？
7 我不觉得他开车开得很好。
8 晚饭以后，我好好地睡了一觉。
9 我喜欢看有快乐结局的电影。
10 这些房子好是好，就是太贵了。
11 昨天死了小猫的小男孩很伤心。
12 那家饭馆的中国菜应该很不错。
13 你真的很喜欢听流行音乐，对不对？
14 她刚把公寓所有的房间都漆完了。
15 在红绿灯那过马路，然后一直往前走。

我希望别墅能便宜一些，我才买得起。
我现在交往的那个人非常有幽默感。
我想要一个我的宠物可以住的大房子。
我希望今年的天气会比去年暖和一点。
我的一个好朋友老帮他的邻居看孩子。
他得先把房间打扫干净才可以出去玩。
我最快乐的回忆就是那次跟你去看戏。
那个被警察抓到的小偷长得又高又瘦。
据统计，每年吸烟的人数越来越多了。
这次考试根本没有你跟我说的那么难。
她点菜的时候只点有肉的，从来不点青菜。
你昨天喂的那只黑猫就是被狗追的那只。
可以麻烦您把桌子上的那本书递给我吗?
我不知道十点半的火车是不是已经开走了。
为什么还是有很多人早上什么东西都不吃呢？

This is the end of the repetition task. Thank you.
## APPENDIX D

### EI TASK SCORING RUBRIC

#### SCORE 0

<table>
<thead>
<tr>
<th>Criteria</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Nothing (Silence)</td>
<td></td>
</tr>
<tr>
<td>• Garbled (unintelligible, usually transcribed as XXX)</td>
<td></td>
</tr>
<tr>
<td>• Minimal repetition, then item abandoned:</td>
<td></td>
</tr>
<tr>
<td>- Only 1 word repeated</td>
<td>- 我喜欢..... (12/#9)</td>
</tr>
<tr>
<td>- Only 1 content word plus function word(s)</td>
<td>- 你昨天挨摸…什么什么.. (01/#27)</td>
</tr>
<tr>
<td>- Only 1 content word plus function word(s) plus extraneous words that weren’t in the original stimulus</td>
<td>- 昨天..校劳 (36/#11)</td>
</tr>
<tr>
<td>- Only function word(s) repeated</td>
<td>- 我…. (113/#1)</td>
</tr>
</tbody>
</table>
| **NOTE: with only, just, yet (meaningful adverbs), score 1**             | - 她刚把…做完了…(116/#14)  
                           | (score 1)                                      |

#### SCORE 1

<table>
<thead>
<tr>
<th>Criteria</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>• When only about half of idea units are represented in the string but a lot of important information in the original stimulus is left out</td>
<td>- 住的大房子. (88/#18)</td>
</tr>
<tr>
<td>• When barely half of lexical words get repeated and meaningful content results that is unrelated (or opposed) to stimulus, frequently with hesitation markers</td>
<td>- 晚饭以后，我… (56/#8)</td>
</tr>
<tr>
<td>• Or when string doesn’t in itself constitute a self-standing sentence with some (targetlike or nontargetlike) meaning (This may happen more often with shorter</td>
<td>- 在红绿灯…(61/#15)</td>
</tr>
<tr>
<td></td>
<td>- 那个被警察…(61/#23)</td>
</tr>
<tr>
<td></td>
<td>- 她点菜的时候…(107/#26)</td>
</tr>
<tr>
<td></td>
<td>- 我…想快…电影 (71/#9)</td>
</tr>
<tr>
<td></td>
<td>- 我希望…便宜…(107/#16)</td>
</tr>
<tr>
<td></td>
<td>- 那家饭….还的不错(30/#12)</td>
</tr>
<tr>
<td></td>
<td>- 我听说的夏天下雨 (66/#5)</td>
</tr>
<tr>
<td></td>
<td>- 你真的…对不对？ (112/#13)</td>
</tr>
<tr>
<td></td>
<td>- 可是好是好，快了 (102/#10)</td>
</tr>
</tbody>
</table>

---

items, where if only 2 of 3 content words are repeated and no grammatical relation between them is attempted, then score 1)

- Also when half of a long stimulus is left out, and the sentence produced is incomplete

<table>
<thead>
<tr>
<th>SCORE 2</th>
<th>Examples</th>
</tr>
</thead>
</table>
| Criteria | -我想我的宠物住的一个大房子 (51/#18)  
-这个东市道很还宽 (97/#3)  
-我能希望...便宜... (55/#16)  
-考试真多的...那么难 (55/#25)  
-考试...那么难 (102/#25)  
-你昨天...那黑猫 (92/#27)  |
| -left out “要，可以”; changed the position of “一个”)  
-这房子的好，太贵了 (6/#10)  
-<left out the measure word “些” and cohesive “就是”; incorrect use of the pattern “好是好”)  
-我不觉得..开得很好 (97/#7)  
-<left out “他开车” and made the sentence ambiguous) |

<table>
<thead>
<tr>
<th>SCORE 3</th>
<th>Examples</th>
</tr>
</thead>
</table>
| Criteria | -我不觉得他开车得开很好(1/#7)  
-他每天都早上都要洗澡 (56/#4)  |
| -Original, complete meaning is preserved as in the stimulus. Strings which are quite ungrammatical can get a 3 score, as long as exact meaning is preserved. Some synonymous substitutions are acceptable.  
-Examples of acceptable substitutions (SCORE 3): 好好=好好地；桌子=桌  
-Anything with ‘很’ can be substituted with ‘好’  
-Examples of unacceptable substitutions or omissions (SCORE 2):  
-个 > 些  
-在 > 想  
-听说 > 听到  
-很 > 最  
-才 > 然后 |
| -晚饭以后，我<好>好睡了一觉 (11/#8) (Score 3)  
-红色的书在<上> (59/#2) (Score 3)  
-昨天死了小猫的小男孩<好>伤心 (57/#11) (Score 3)  
-这些考试根本没有你跟我说的那么难 (41/#25) (Score 2)  
-你刚才说你今天想想做什么 (11/#6) (Score 2)  
-我<听到>明天可能会下雨 (56/#5) (Score 2)  
-我不觉得他开车开得<好> (104/#7) (Score 2) |
- 递给 > 给

- Changes in grammar that affect meaning should be scored as 3. For example, a present progressive tense repeated as past or as future should be scored as meaning change (score 2).

- Similarly, singular/plural differences between stimulus and repeated string change the meaning, not only the grammar (score 2).

- Ambiguous changes in grammar that COULD be interpreted as meaning changes from a NS perspective should be scored as 2. That is, as a general principle in case of doubt about whether meaning has changed or not, score 2.

| SCORE 4 |
|---|---|
| Criteria | Examples |
| • Exact repetition: String matches stimulus exactly. Both form and meaning are correct without exception or doubt. | -他得先把房间打扫干净然后可以出去玩 (41/#21)(Score 2)  
-可以麻烦您把桌子上的那本书给我吗(67/#28)(Score 2)  
-你刚才说你今天要做什么.(41/#6)(Score 2)  
-晚饭以后，我好好地睡一觉. (41/#8)(Score 2)  
-这些考试根本没有你跟我说的那么难(41/#25)(Score 2)  
-这些房子好是好，但太贵了 (14/#10)(Score 2)  
(但 refers to a turn in meaning, but 就是 not only refers to a turn in meaning, but also points out the only flaw.) |
APPENDIX E

SURVEY³

Participant number __________

Gender: □ Male □ Female    Age: __________________________

Q1. What was your first or strongest language before your age of 5?
□ English □ Other (specify) ________

Q2. What is your strongest language now?
□ English □ Mandarin Chinese □ Other (specify) ________

Q3. Check if your parents, grandparents, or anyone else in your immediate/extended family is a native speaker of Mandarin Chinese or a Chinese dialect. □ Mother □ Father □ Maternal grandparent(s) □ Paternal grandparent(s) □ Other (specify) ________

Q4. At what age did you start to hear or use Mandarin Chinese? __________

Q5. Mandarin learning inside classroom
1. How long (in years) in total have you studied Mandarin at school? ______________

2. List the following information for any previous Mandarin studies (e.g., college, high school, intermediate/elementary school, Chinese language school, private language institute, private tutor, etc.).

Please also include the current study program.

School 1: __________________ (school name) in __________________ (country name)
Start year: _______ End year: _______ Hours of Mandarin class per week _______

School 2: __________________ (school name) in __________________ (country name)
Start year: _______ End year: _______ Hours of Mandarin class per week _______

School 3: __________________ (school name) in __________________ (country name)
Start year: _______ End year: _______ Hours of Mandarin class per week _______


L2 Perceived competence scale and Frequency of L2 use score were from MacIntyre, P. D., Baker, S., Clément, R., & Donovan, L. (2002). Sex and age effects on willingness to communicate, anxiety, perceived competence, and L2 motivation among junior high school French immersion students. Language Learning, 52, 537-564.

Q4. Have you visited/lived in a Chinese-speaking country?
- No
- Yes (if YES, see below)

(At what age: _________; For ___________ [length of the stay]; Location: _________________)
(At what age: _________; For ___________ [length of the stay]; Location: _________________)
(At what age: _________; For ___________ [length of the stay]; Location: _________________)

Q7. How much do you hear or use Mandarin outside classroom?
1: never 2: occasionally 3: sometimes 4: frequently 5: almost always

- speak Mandarin at home
- speak Mandarin at school outside the Mandarin classroom
- speak Mandarin in the community, e.g. at work, etc.
- other, please specify

<table>
<thead>
<tr>
<th>Activity</th>
<th>Never</th>
<th>Rarely</th>
<th>Sometimes</th>
<th>Often</th>
<th>Frequently</th>
<th>Many times</th>
<th>Almost always</th>
</tr>
</thead>
<tbody>
<tr>
<td>speak Mandarin at home</td>
<td>N/A</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>N/A</td>
</tr>
<tr>
<td>speak Mandarin at school</td>
<td>N/A</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>N/A</td>
</tr>
<tr>
<td>speak Mandarin in the community</td>
<td>N/A</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>N/A</td>
</tr>
<tr>
<td>other, please specify</td>
<td>N/A</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>N/A</td>
</tr>
</tbody>
</table>

8. Below are twelve situations in which you need to communicate. Please indicate how often you have engaged in each of the situations described below.

1=Never 2=Rarely 3=Sometimes 4=Often 5=Frequently 6=Many times
7=Many, many times

1. Have a small-group conversation in Mandarin with acquaintances.
2. Give a presentation in Mandarin to a group of strangers.
3. Give a presentation in Mandarin to a group of friends.
4. Talk in Mandarin in a large meeting among strangers.
5. Have a small-group conversation in Mandarin with strangers.
6. Talk in Mandarin in a large meeting among friends.
7. Talk in Mandarin to friends.
8. Talk in Mandarin in a large meeting with acquaintances.
9. Talk in Mandarin to acquaintances.
10. Give a presentation in Mandarin to a group of acquaintances.
11. Talk in Mandarin to a stranger.
12. Talk in Mandarin to a small group of friends.

9. This part is composed of statements concerning your feelings about communication with other people, in Mandarin. Please indicate in the space provided the frequency of time you would CHOOSE to speak in Mandarin in each classroom situation.

1=Almost never willing 2=Sometimes willing 3=Willing half of the time 4=Usually willing 5=Almost always willing

**Speaking in class, in Mandarin**

1. Speaking in a group about your summer vacation.
2. Speaking to your teacher about your homework assignment.
3. A stranger enters the room you are in, how willing would you be to have a conversation if he talked to you first?
4. You are confused about a task you must complete, how willing are you to ask for instructions/clarification?

5. Talking to a friend while waiting in line.

6. How willing would you be to be an actor in a play?

7. Describe the rules of your favorite game.

8. Play a game in Mandarin, for example Monopoly.

Reading in class (to yourself, not out loud)

1. Read a novel.

2. Read an article in a paper.

3. Read letters from a pen pal written in native Mandarin.

4. Read personal letters or notes written to you in which the writer has deliberately used simple words and constructions.

5. Read an advertisement in the paper to find a good bicycle you can buy.

6. Read reviews for popular movies.

Writing in class, in Mandarin

1. Write an advertisement to sell an old bike.

2. Write down the instructions for your favorite hobby.

3. Write a report on your favorite animal and its habits.

4. Write a story.

5. Write a letter to a friend.

6. Write a newspaper article.

7. Write the answers to a “fun” quiz from a magazine.

8. Write down a list of things you must do tomorrow.

1=Almost never willing  2=Sometimes willing  3=Willing half of the time  4=Usually willing  5=Almost always willing

Comprehension in class

1. Listen to instructions and complete a task.

2. Bake a cake if instructions were not in English.

3. Fill out an application form.

4. Take directions from a Mandarin speaker.

5. Understand a Mandarin movie.

10. Sometimes people differ a lot in their speaking, reading, and so forth in class and outside class. Now we would like you to consider your use of Mandarin outside the classroom. Again, please tell us the frequency that you choose to use Mandarin in the following situations. Remember, you are telling us about your experiences outside the classroom this time. There are no right or wrong answers.

1=Almost never willing  2=Sometimes willing  3=Willing half of the time  4=Usually willing  5=Almost always willing
**Speaking outside class, in Mandarin**

1. Speaking in a group about your summer vacation.
2. Speaking to your teacher about your homework assignment.
3. A stranger enters the room you are in, how willing would you be to have a conversation if he talked to you first?
4. You are confused about a task you must complete, how willing are you to ask for instructions/clarification?
5. Talking to a friend while waiting in line.
6. How willing would you be to be an actor in a play?
7. Describe the rules of your favorite game.
8. Play a game in Mandarin, for example Monopoly.

**Reading outside class, in Mandarin**

1. Read a novel.
2. Read an article in a paper.
3. Read letters from a pen pal written in native Mandarin.
4. Read personal letters or notes written to you in which the writer has deliberately used simple words and constructions.
5. Read an advertisement in the paper to find a good bicycle you can buy.
6. Read reviews for popular movies.

**Writing outside class, in Mandarin**

1. Write an advertisement to sell an old bike.
2. Write down the instructions for your favorite hobby.
3. Write a report on your favorite animal and its habits.
4. Write a story.
5. Write a letter to a friend.
6. Write a newspaper article.
7. Write the answers to a “fun” quiz from a magazine.
8. Write down a list of things you must do tomorrow.

**Comprehension outside class**

1. Listen to instructions and complete a task.
2. Bake a cake if instructions were not in English.
3. Fill out an application form.
4. Take directions from a Mandarin speaker.
5. Understand a Mandarin movie.

**11. Below are twelve situations in which you might need to communicate. People's abilities to communicate effectively vary a lot, and sometimes the same person is more competent to communicate in one situation than in another. Please indicate **HOW COMPETENT** you believe you are to communicate in each of the situations described below. Indicate in the space provided your estimate of your competence by putting a number between 0 and 100.**

196
Presume 0=completely incompetent and 100=competent.

________ 1. Have a small-group conversation in Mandarin with acquaintances.
________ 2. Give a presentation in Mandarin to a group of strangers.
________ 3. Give a presentation in Mandarin to a group of friends.
________ 4. Talk in Mandarin in a large meeting among strangers.
________ 5. Have a small-group conversation in Mandarin with strangers.
________ 6. Talk in Mandarin in a large meeting among friends.
________ 7. Talk in Mandarin to friends.
________ 8. Talk in Mandarin in a large meeting with acquaintances.
________ 9. Talk in Mandarin to acquaintances.
________ 10. Give a presentation in Mandarin to a group of acquaintances.
________ 11. Talk in Mandarin to a stranger.
________ 12. Talk in Mandarin to a small group of friends.

12. Below are the same twelve situations in which a person feels different degrees of anxiety. Please indicate in the space below what degrees of anxiety you might feel in the following situations.

0 (I don't feel anxious at all) ------------------------ 100 (I always feel anxious)

________ 1. Have a small-group conversation in Mandarin with acquaintances.
________ 2. Give a presentation in Mandarin to a group of strangers.
________ 3. Give a presentation in Mandarin to a group of friends.
________ 4. Talk in Mandarin in a large meeting among strangers.
________ 5. Have a small-group conversation in Mandarin with strangers.
________ 6. Talk in Mandarin in a large meeting among friends.
________ 7. Talk in Mandarin to friends.
________ 8. Talk in Mandarin in a large meeting with acquaintances.
________ 9. Talk in Mandarin to acquaintances.
________ 10. Give a presentation in Mandarin to a group of acquaintances.
________ 11. Talk in Mandarin to a stranger.
________ 12. Talk in Mandarin to a small group of friends.
REFERENCES


---

4 Note. References marked with * indicate quantitative studies included in the synthesis study of L2 WTC. References marked with ** indicate qualitative studies included in the synthesis study of L2 WTC. References marked with # indicate quantitative studies included in the synthesis study of L2 EI. References marked with ## indicate qualitative studies included in the synthesis study of L2 EI.


Byrne, B. M. (2010). *Structural equation modeling with AMOS* (2nd ed.). New York, NY: Taylor and Francis Group, LLC.


Retrieved from ProQuest Dissertations and Theses database. (AAT 3297511, Capella University)


Le, J. Y. (2004). *Affective characteristics of American students studying Chinese in China: A study of heritage and non-heritage learners’ beliefs and foreign*
language anxiety (Doctoral dissertation). Retrieved from ProQuest Dissertations and Theses database. (AAT 3150676, The University of Texas at Austin)


#Li, S. F. (2010). Corrective feedback in perspective: The interface between feedback type, proficiency, the choice of target structure, and learners' individual
differences in working memory and language analytic ability (Doctoral dissertation). Retrieved from ProQuest Dissertations and Theses database. (AAT 3432414, Michigan State University)


