2011-12-08

Gains in Fluency Measures during Study Abroad in China

Jeongwoon Kim
Brigham Young University - Provo

Gains in Fluency Measures during Study Abroad in China

Jeongwoon Kim

A thesis submitted to the faculty of
Brigham Young University
in partial fulfillment of the requirements for the degree of

Master of Arts

Dan P. Dewey, Chair
Dana S. Bourgerie
Jennifer Bown

Center for Language Studies

Brigham Young University

December 2011

ABSTRACT

Gains in Fluency Measures during Study Abroad in China

Jeongwoon Kim
Center for Language Studies, BYU
Master of Arts

This thesis study investigates gains in the speaking of China study abroad (SA) students from Brigham Young University. Pre-and post-program Simulated Oral Proficiency Interview (SOPI) tasks were used to generate multiple fluency measures, such as native judges' subjective fluency ratings, word count, number of unique words, number of filler words, mean pause length, tonal accuracy, etc. The study results display significant differences between pre- and post-tests for all fluency measures. In other words, China SA students were perceived to be more fluent in their speech by native judges after SA; their speech samples show more word (token and type) production, and shorter pauses in post-SOPI tasks than in pre-SOPI tasks. Participants used more filler words and had more unfilled pauses in post-measures than in pre-measures and they enhanced their tonal accuracy during the SA. Native judges' perception of task completion was also measured and the students were more capable of completing speech tasks in the post-program measures than in the pre-program measures. The OPI ratings indicate that some students made as much as two sub levels' improvement on the American Council on the Teaching of Foreign Languages (ACTFL) scale.

Keywords: fluency, Chinese, language gain, speaking, tones, tonal accuracy

ACKNOWLEDGEMENTS

I thank God our Heavenly Father for His guiding Hand throughout this thesis work. I am thankful for my family in Korea and Yoonho Kal for their unconditional love and support. I would like to express my gratitude for the following individuals: my Thesis Committee Chair, Dr. Dan P. Dewey, for his exemplary life as a disciple-scholar; Spencer Ring and Andrew Westover for their help with researching; Kendon Kurzer for his excellent help in editing; Dr. Dennis Eggett for his help in the statistical analysis; Dr. Michael D. Bush for his help in formatting.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

Recently, going on a study abroad (SA) has become a very common tool to acquire foreign language skills. According to the National Center for Education Statistics, the 262,400 U.S. students that studied abroad during the 2007-2008 academic year, quadrupled from the 62,300 study abroad students in 1987-1988 (National Center for Education Statistics, 2010).

As more foreign language learners choose to go on SA programs, an increasing number of researchers in the field of second language acquisition have been attracted to the topic of SA. Among the questions these researchers have been trying to answer are: Are study abroad programs beneficial to language learning? If so, in what ways do study abroad programs help foreign language learners increase their language skills?

Studies that examined SA with standardized proficiency measures generally accredit SA with increasing language proficiency (Carroll, 1967, O'Connor, 1988, Magnan, 1986, Liskin-Gasparro, 1984, Milleret, 1991). However, researchers in the field have consistently agreed that different SA participants reap varied results. Therefore, efforts to identify some of the factors that engender different levels of language gain followed these initial studies. So far, some of these factors include motivation (Isabelli-García, 2006; Douglass, 2006), attitude toward host culture (Isabelli-García, 2006; Wilkinson, 2002), language aptitude (Davidson & Ginsburg, 1995; Davidson, 2010), and social networks (Fraser, 2002; Whitworth, 2006).

As SLA researchers strove to find factors that contribute to successful language gain, fluency has received continued attention as part of language development during SA. In other words, fluency is often thought to be one of the most prominent benefits of SA (Freed, 1995b). Freed also argues that the term "fluency" is often used interchangeably with global language ability and oral ability (Freed 1995b, p. 124-5). Freed's (1995b) study contributed to the

identification of specific features of fluent speech and thus largely provides the operationalization of the construct of fluency for this study.

In the area of studying Chinese language acquisition and fluency, there has been relatively little research. However, becoming fluent in Chinese is the ambition of many non-native Chinese learners with the booming economy of China. According to Central Intelligence Agency (CIA) World Fact Book, China has become the world's largest exporter and had a 10% GDP growth in 2010. ACTFL's data, that compared 2004-2005 with 2007-2008 language course enrollment, shows that the Mandarin language has had the largest increase in enrollment, which was around 195%. The enrollment in Japanese language courses increased 18% from 2004-2005 to 2007-2008, while German, Russian, and Spanish increased 7%, 3%, and 2% respectively.

Brigham Young University-Provo has been sending its students to Nanjing, China for a study abroad program each year since 1985. While the fluency gains of these students and the impact of the China SA on their acquisition of the Chinese language are hoped for and expected, there has not yet been a study that measured China SA students' speaking gains, particularly in fluency. This thesis study, therefore, explores this new sphere of research.

The research questions that this thesis strives to answer are as follows: Do China SA participants' language skills change in terms of fluency during SA? If so, what measures of fluency do change over time?

Chapter 2 discusses some of the foundational studies for this thesis. Chapter 3 introduces how this study is designed to answer the research questions, along with the description of subjects and measurements. Chapter 4 shows the results of data analysis and the implications thereof. Chapter 5 summarizes the former chapters and provides conclusions of the study, implications to the field, and suggestions for further research.

Chapter 2: Review of the Literature

In order to comprehend the purpose of the current study, it is pertinent to understand the context that this study stems from. This chapter first introduces the history of study abroad (SA) research, from 1967 until now. It then investigates various factors that SA researchers have found to affect language gain. After that, it describes essential studies that examined fluency and SA in depth. Lastly, this chapter reveals a gap in SA research and concludes by establishing the need for the current study.

History of SA Research

John Carroll's 1967 study first showed attention to the role of SA in language acquisition. Carroll studied 2,782 graduating college seniors who were learning French, German, Italian, and Russian and measured their proficiency in four skills—listening, speaking, reading, and writing—through the MLA Foreign Language Proficiency Test. As a result, he proposed that time spent abroad is a strong predictor of language proficiency. He wrote:

> Time spent abroad is clearly one of the most potent variables we have found, and this is not surprising, for reasons that need not be belabored. Certainly our results provide a strong justification for a 'year abroad' as one of the experiences to be recommended for the language majors. Even a tour abroad, or a summer school course abroad, is useful, apparently, in improving the student's skill (Carroll, 1967, p. 137).

Carroll's large scale study, however, was not without limitations. His study did not compare language proficiency directly before and after the SA, but assessed college seniors' language proficiency at the time of graduation. Therefore, even with his large pool of subjects and strong correlation between SA and high achievement scores in standardized tests, it is unclear exactly how SA aided in increasing the language learners' language proficiency. On the

other hand, Carroll's study laid a foundation for further research on SA by describing what contributes to proficiency.

Carroll's study was then followed by research performed by some British researchers between the late 1960s and early 1970s. Willis, et al. (1977) also used standardized test scores to measure gains in speaking, listening, and writing skills of British students who spent over a year working or studying in Germany or France. Even though the authors acknowledged a lack of structure in their study, the results do support the role of a SA in initiating growth in the above-mentioned language skills.

While the work of Carroll and a number of British scholars on the benefits of SA was more objective and product-focused, Schumann and Schumann's study in 1977 marked the beginning of more subjective and process-based research. The Schumanns recounted their own language learning experience abroad, learning Arabic in Tunisia and Farsi in Iran. Their narrative account of language learning introduced individual factors, such as anxiety, motivation, cultural adaptation, personality, and learning strategies to SA research. Their studies were followed by other diary-based research by Ochsner and Baily (Ochsner, 1979; Bailey and Ochsner, 1983).

Other than a few case studies, such as one by Moehle and Raupach (1983), research on SA during the 1980s was predominantly affected by the "proficiency movement" (Kinginger, 2008, p. 41). With the foundation of the American Council on the Teaching of Foreign Languages (ACTFL) and Oral Proficiency Interview (OPI) guidelines, numerous studies on SA during this time used OPI ratings as a proficiency measurement. Studies cited by Freed (1995), by O'Connor (1988) and Magnan (1986) in French, Liskin-Gasparro (1984), Foltz (1991), and Veguez (1984) in Spanish, and Milleret (1991) in Portuguese discovered that students who went

abroad to study these languages (French, Spanish, and Portuguese) typically received OPI ratings of at least one sub-level higher than before. In addition, some of these studies support that SA students' OPI ratings were higher than those who received domestic, at home, instruction (AH).

Even though all of the above-mentioned studies that use pre- and post-program OPIs support the use of OPI ratings as an assessment tool of language proficiency, some researchers have acknowledged the limitations of OPI ratings. For example, progress made in language learning during a short period of time is not likely to be captured by OPI ratings. Both Freed (1990), who studied French learners in France for 6 weeks, and Milleret (1991), who studied Portuguese learners in Brazil for one summer, argue that OPI scores did not adequately reflect the progress of their participants. For example, Milleret saw more progress in advanced learners' language skills in informal observations than what was shown in the OPI. On the same note, Freed argued, "The OPI which utilizes one global holistic score for various aspects of language use is not sufficiently refined to capture growth in oral skills, particularly in a six-week period" (Freed, 1990, p. 475).

Some researchers argued that another limitation of the OPI ratings is the non-linear correlation between time spent abroad and OPI results. As Magnan (1986) studied 40 students learning French at the end of their first, second, third, and fourth years, their OPI proficiency ratings did not correspond exactly with their level of study, but rather overlapped. Thus, Magnan argues that OPI proficiency ratings do not show a credible relationship to level of study and are unpredictable. Freed (1995a) also argued that "because of non-linear construction, the OPI is often unable to discriminate progress made by students at the upper levels of the proficiency scale" (Freed, 1995a,16). However, Brecht et el.'s 1995 study indicates that time is a significant predictor of language gain, as measured by OPI or Interagency Language Roundtable (ILR)

scores. Thus, this argument is debatable. It is true, however, that the OPI cannot account for all language gains.

Starting in the late 1990s, some second language acquisition (SLA) researchers began to use language tests other than OPIs, such as the C-Test. A C-Test is a written test that measures one's overall language ability by providing test takers incomplete texts and having them reconstitute meaning. Coleman's 1996 study is a noteworthy study that made use of C-Tests. Coleman administered C-Tests to more than 25,000 students of French, Spanish, German, English, and Russian in the United Kingdom and other European countries in a large-scale project called the European Language Proficiency Survey. Coleman found that students who went on a SA scored significantly higher on the C-Test than those who did not. He also stated that the longer the participants were abroad, the higher the C-Test scores were.

Another change in the 1990s in SA research is the increase in the use of subjective ratings to measure language gain during SA. Meara (1995), for example, asked 586 SA participants to rate their improvement of the four language skills—speaking, listening, reading, and writing—during SA. In response, most students reported increased confidence in interactive skills, namely speaking and listening. Yager's 1998 study used native Spanish judges to rate speech samples of students who stayed in Mexico for a summer program. The native judges gave holistic ratings as to nativeness, grammar, and the pronunciation of the participants. Yager reports that the majority of the students improved in at least one area. Kinginger (2008) asserts that Yager's study can be strengthened by defining "nativeness" and supporting native judgment of nativeness with specific examples of language use.

While research on SA prior to the 1980s focused on answering more simple questions such as whether SA benefits language learners, the research on this topic from the 1990s on

strived to identify specific factors that affect language gain. The following section of this chapter introduces some of those factors and critical research related to them. Fluency deserves to be mentioned first as it is the main topic of this study.

Fluency and Study Abroad

The section above consisted of a brief history of SA research. In this history, proficiency gain during SA has been the focus. This section emphasizes fluency during SA, which is a critical concept of this study.

The broad definition of fluency as smooth, flowing, clear speech performance was reevaluated and specified by Freed (1995b). By identifying seven factors of fluency, which are mostly temporal features of speech, Freed strives to focus the layman definition of fluency. These seven factors of fluency are as follows (Freed, 1995b, p. 130-131):

1. Amount of speech: raw frequencies of non-repeated words or semantic units
2. Rate of speech: number of non-repeated words or semantic units per minute
3. Unfilled pauses: number and length of silences that sounded dysfluent and are longer than .4 seconds
4. Frequency of filled pauses: number of lexical and non-lexical fillers, drawls or sound stretches
5. Length of fluent speech runs: length of continuous speech not interrupted by dysfluent pauses or hesitations

6. Repairs: number of repetitions, false starts, grammatical repair, etc.
7. Clusters of dysfluencies: two or more interruptions to the flow of speech

In Freed's study, six native judges listened to 30 students' OPI segments. Their ratings indicated that SA students exhibit a faster rate of speech than AH students. SA students' speech in general was characterized by longer speech runs, fewer unfilled pauses, and the desire to express more complex ideas.

Another study by Freed, et al., in 2004, defined fluency as a broader construct. In addition to temporal measures of fluency, the researchers included general features of speech performance in fluency measures. These general features include total word count, duration of speech, and length of longest run. Their 2004 study was unique as they not only compared SA and AH students, but also included those who participated in an immersive summer program (IM). The speech of 28 French students provided evidence that the AH group made no significant progress in fluency. The SA group made more statistically significant gains in speech fluidity than the AH group, but the IM group made the greatest gain in speech fluidity. In addition to speech fluidity, the IM group's speech presented the most increased number of unique words, highest gains in rate of speech, and longer mean length of fluent runs of speech. Accordingly, the authors concluded that SA is not the ultimate solution in order to improve fluency in language, but language learners can gain fluency by being immersed in a language learning environment that enables quality interactions and by being committed to use the second language (L2). They assert, it is "the nature of the interactions, the quality of the experiences, and the efforts made to use the L2 that render one context superior to another with respect to language gain" (Freed et al., 2004, p. 298).

Number of words and unique words produced have been considered measures of fluency by Freed (1995b) and Freed et al. (2004). These measures are sometimes expressed as tokens and types. Addressing vocabulary measurement, Read (2000) explains that the term token denotes the "total number of word forms, which means... individual words occurring more than once in the text" (Read, 2000, p. 18). He further explains that "the number of types is the total number of *different* word forms, so that a word which is repeated many times is counted only once" (Read, 2000, p. 18). Read notes that type-token ratio is used often as a measure of language development.

Dubiner, Freed, and Segalowitz (2007) studied how native judges perceive fluency. They asked nine native Spanish judges to write their definitions of fluency. The native judges also listened to the AH and SA participants' speech samples, and gave fluency ratings on a scale of 1 (lowest level) to 7 (highest level). The researchers found a discrepancy between native judges' written definition of fluency and their temporal fluency ratings. Most NS judges defined fluency more generally in writing, using terms such as clarity, comprehensibility, and grammatical accuracy. However, their actual ratings of speech samples show that they were more influenced by narrow aspects of fluency, such as filled and unfilled pauses and rate of speech. The pre-and post-test mean scores of SA and AH students show that the NS judges perceived the SA group as being significantly more fluent than the AH group.

Continued interest in fluency gain during SA has led some researchers to question whether gains in fluency correspond with gains in accuracy. Freed, So, and Lazar (2003) used SA and AH French learners' written essays to measure written fluency and grammatical accuracy. While native judges' subjective ratings identified the SA group's written fluency to be above that of the AH group, pre-post analysis of grammatical correctness showed no

improvement for either group. Freed, So, and Lazar admit that the improved written fluency might have resulted from the SA students' ambition to communicate their interesting stories during SA. Allen and Herron (2003) likewise found similar results for spoken language. They used picture description and OPI role-play to assess French learners' speech before and after their 6-week study in France. They reported that the participants' speech displayed significant advancement in fluency, but only minor growth in grammatical correctness.

Other researchers have deviated from the sole use of standardized proficiency tests or the OPI and have attempted other assessment tools. For example, Segalowitz and Freed (2004) attempted to compare 25 SA students in Spain and 18 AH students by administering lexical access (word recognition) tests. The participants were given computerized lexical access tests with English and Spanish words. The test takers looked at prompt words on a computer screen and had to choose whether a word was animate or non-animate. The researchers recorded the participants' speed of lexical access (reaction time) and efficiency of lexical access (accuracy of choice). Their results displayed a significant correlation between lexical processing and oral fluency measured by OPI scores.

Juan-Garau and Pérez-Vidal (2007) used role-play and narrative to measure the fluency of English learners in a longitudinal study. Their research data suggest that twelve Spanish-Catalan bilingual students who studied for three months in an Anglophone environment gained in fluency, made fewer grammatical and lexical errors, and produced more complex sentences. In addition, those students retained their fluency and accuracy gains even 15 months after the SA.

Factors that affect language gain during SA

Second language acquisition researchers have been striving to find out what factors account for different results of language learning during SA. This section introduces and

describes compelling studies on some of these factors: gender, aptitude, personality, individual goals, motivation, social interaction and language use.

Gender. Francine Schumman (1980) cited gender as an influencing factor in her own study of Farsi, saying that learning Farsi in Iran was a "far greater endeavor for a woman than a man" (p. 55). Whether this claim was true or not was investigated by other researchers, but the results have been controversial. Brecht, Davidson & Ginsburg (1995) studied 658 participants in a four-month study conducted by the American Council on the Teaching of Russian (ACTR) from 1984 to 1990. They administered the OPI, a listening and reading test (by Educational Testing Service), a grammar and reading test (by ACTR), and a short Modern Language Aptitude Test (MLAT). They discovered that previous experience of learning other languages, grammar and reading skills, and gender were the strongest predictors of language gain. Their study results revealed that men were more likely to succeed on the ETS Listening Test as well as to achieve an advanced level of speaking ability. The authors admitted a possible gender bias in the testing instrument and a skew in sample, but they posited that the gender difference likely arose from culturally different interactions between American and Russian men and women. Polanyi (1995) explains that this may in part be due to the negative reaction of female participants to aggressive Russian males, which limited opportunities for interaction and decreased female learners' language gain. Davidson has argued, however, that the collapse of the Soviet Union has changed the social norms in Russia, including the norms that dictate proper interaction between males and females (Davidson, 2010). In conclusion, the claim that gender is a predicting factor of language gain needs further investigation.

Aptitude. Scholars have investigated whether a learner's aptitude is related to successful language learning during SA. Brecht and Davidson (1991) report that language aptitude, as

measured by the Modern Language Aptitude Test (MLAT), is strongly related to learners' reading gains. Brecht, Davidson & Ginsburg's (1995) study further investigated the relationship between aptitude and language gain during SA. They reported that high language aptitude was a good predictor of gains in listening and reading, but was not a significant predictor of speaking gains.

Apart from language learning abilities, some researchers have further investigated the link between cognitive skills and fluency development. For example, Segalowitz and Freed's (2004) study presented data on the link between lexical processing and fluency gain. According to them, accurate and speedy lexical processing is significantly related to fluency gain. O'Brien, Segalowitz, Freed, and Collentine's (2007) study indicates a link between phonological memory and fluency gain. Davidson (2010) noted that a strong correlation between pre-listening proficiency and post-speaking proficiency exists for advanced level students.

Personality, individual goals, and motivation. There are other factors that some researchers profess to affect SA language gain, but these factors remain debatable. These include personality (DeKeyser, 1997), individual goals (Ginsburg & Miller, 2000), and motivation (Douglass, 2006). One of the challenges of verifying the importance of these factors is that they are highly individual and studies are not easy to replicate. In fact, most researchers who focus on individual goals or motivation perform qualitative case studies that are often difficult to generalize. In addition, factors such as motivation are often apt to change depending on learning environments and individual circumstances, as Peirce (1995) acknowledged. Wilkinson (2002) noted that motivation can increase or decrease by treatment they receive from host culture and the warmth of social interaction. Isabelli-García (2006) says that motivation is swayed by level of success in language learning.

Social interaction. Social interaction (i.e. social networking) has been attracting SA researchers' attention as many researchers closely follow SA participants through case studies. Some researchers note that the successful formation of social networks and accompanying meaningful interaction with native target language speakers may contribute to the development of language skills. Both Fraser (2002) and Whitworth (2006) note that students who take advantage of a variety of opportunities for social interaction, such as participating in community events, sports activities, and taking jobs, had more language gains than those who do not. Similarly, Isabelli-García (2006) also found that the SA participants' social networks predict language gains. In her study of Spanish SA language learners forming social networks with native Argentines, Isabelli-Garcia noted that those who formed first order zone social networks gained less from SA than those who formed second order zone social networks. First order zone social networks are formed with those who are *directly* linked to the participants while second order social networks are with those who are *distantly* connected to the participants (Milroy, 1987, p. 46). In other words, participants who successfully formed more extensive social networks gained more.

Language use. Freed, Segalowitz, and Dewey's (2004) study with the Language Contact Profile suggests that language use is related to fluency development. In their study, SA, IM, and AH students reported the amount of time they spent doing different activities in English and French. According to their study, IM students, who had significantly more contact hours in the classroom and doing extracurricular activities using French, made the greatest fluency gain.

Ginsburg and Miller (2002) found that measuring only time-on-task does not predict language gains (Ginsburg & Miller, 2002, p. 245- 56). They found that even some students who spent significant time speaking the target language outside of the classroom and who formed

extensive social networks with native speakers did not improve linguistically, such as the case of one SA student, Simon. On the other hand, Juanita, who did not spend as much time speaking in the language, but successfully formed quality relationships with a few friends, showed more significant linguistic gains. Ginsburg and Miller (2002) therefore argue that a quality relationship that "goes beyond scripted and predictable types of conversation" contributes to meaningful language learning (Ginsburg & Miller, 2002, p. 254).

Recent SA research points out that SA participants face great obstacles to using their target language during SA. One of these obstacles is the development of communication technology that keeps SA students' venue for communicating in their native languages intact, thus making a full linguistic immersion in a SA challenging. Kinginger (2008) asserts that, "linguistic immersion is increasingly a matter of choice and even of struggle," for students who go on SA programs nowadays in this globalized world (p.108). She further argues that this choice and struggle to be fully immersed linguistically during a SA "requires a more profound and durable commitment than has been needed in the past" (Kinginger, 2008, p. 105).

Social networks with SA co-participants are often considered to be the foremost challenge to extensive language use during SA (Kinginger, 2007; Wilkinson, 2005). The majority of SA programs often assign students to travel, live, or spend time in other ways with co-participants from their native country. Since SA participants tend to face the same challenges, such as culture shock and difficulty acquiring the same language, it is often easier to deal with these issues as a group. Thus, they form "compatriot island[s]" and stay close to each other as described by Wilkinson (2005). Kinginger (2007) also note that students often maintain perfectly intact social networks that are "impervious to influences from the foreign culture" (Kinginger, 2007). SA participants who spend a lot of time interacting with co-participants or communicating

with family and friends back home may find less time to form meaningful relationships with native target language speakers that allow them meaningful interaction in the L2.

Acquiring Chinese Tones

[1]There are four Chinese tones. Relative pitch of these tones is shown in the diagram below. The first tone is high and has a stable pitch, the second tone is a rising one, the third is a dipping tone, while the fourth is a falling tone. Mispronouncing Chinese words in wrong tones (even if the pronunciation of pinyin sound is correct) changes the meaning, thereby hindering communication with native Chinese speakers. For example, "ma" can mean mother (the first tone), to bother (the second tone), horse (the third tone), or to scold (the fourth tone), depending on the tone.

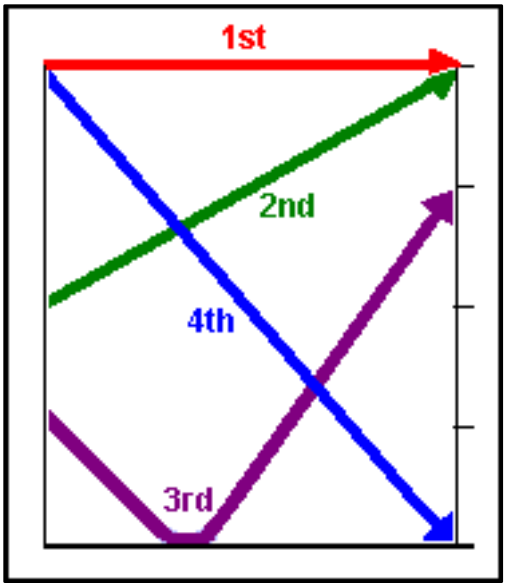

Figure 1. Four Chinese tones. The vertical line represents relative pitch of tones and the horizontal movement represents changes in the pitch.

[1] Retrieved from http://www.instantspeakchinese.com/pinyin/pinyinTones.cfm

The difficulty of acquiring Chinese tones has been discussed by many Chinese as a foreign language educators. Christensen and Warnick (2006) assert that "a system of four primary tones... is perhaps the most challenging part of Chinese phonology for the second language learner to master" (Christensen and Warnick, 2006, p. 86). Shen (1989) also reports that pitch and register errors are the most common errors that learners of Chinese as a foreign language make in their speech. Miracle (1989), who studied tonal errors in beginning Chinese learners' (who have been studying the language approximately for a year) speech, reports that the rate of tonal errors is 42.9%. Non-native Chinese learners seem to agree that tones are a challenging part of studying the Chinese language. Cheng's 2011 study also reports that 69% of 64 non-native Chinese learner subjects replied that tones are the hardest feature of pronunciation when asked to choose between vowels, medials/finals, or tones.

There have been debates on which of the four Chinese tones is the hardest to master. While Shen (1989) argues that the fourth tone is the hardest to speak, Miracle (1989) asserts that the second tone is the most challenging one. McGinnis (1997) and Christensen and Warnick (2006) agree that the second and the fourth tones are equally problematic tones for non-native learners of Chinese. White (1991) posits that acquisition of Chinese tones is especially difficult for native English speakers since native English speakers' pitch range is relatively narrower than that of Chinese speakers. Thus, it is easy for native English speakers to produce tones like intonation or word stress. Cheng's (2011) study of 64 non-native learners of Chinese with varied nationalities, studying in Hangzhou, China, shows slightly varied results. When asked, 49% of these students reported that the third tone is the hardest to learn and 41% of them replied the second tone is the hardest. Researchers used Praat software to analyze six randomly-picked

subjects' pronunciation of the second and the third tones. They found that the subjects often mixed up the second tone with the third tone and vice versa.

Even though there are differing results about which Chinese tones are the most challenging ones to learn, suffice it to say that without proper pronunciation of tones, achieving fluency and proficiency in the Chinese language is almost impossible. Anyone who would like to become truly fluent and proficient in Mandarin Chinese should master the four tones, which requires a consistent and dedicated effort. Just like Cheng (2011) advises, Chinese teachers must understand that teaching tones is and should be a long-term endeavor and a part of the whole curriculum in Chinese language education.

Need for Current Study

Most of the research on SA has been with Indo-European languages with a few exceptions. In non-Indo-European language research, Huebner (1995) compared Japan-based students of Japanese language with U.S.-based Japanese learners in their listening and reading skills. He found that the Japan-based students outperformed the U.S.-based students significantly in reading comprehension, but for character (Kanji) recognition, no significant difference was detected. Hayden's (1998) study used the Computer Adaptive Test for Reading Chinese in assessing the reading skills of 21 English-speaking students who spent a semester in China. Most students gained one sub-level or more on the ACTFL scale.

A series of studies by Dewey on reading and vocabulary acquisition are currently the most advanced ones in the Asian languages and study abroad research. In his 2004 study, Dewey reports that the SA group felt significantly more confident about reading in Japanese than the domestic immersion (IM) group in post-test measures. The IM group demonstrated significantly less monitoring of reading comprehension and more affective reaction to texts than the SA

group. In his 2006 study, he further finds that the SA group had out gained the AH group, but the IM group demonstrated an increased depth of vocabulary knowledge when compared to the SA group. Summarizing these two studies, Dewey suggests that SA can "promote vocabulary and reading development not only in European languages but also in Asian languages" (Dewey, 2006, p. 77).

Bourgerie's 2010 presentation during the 8th International conference on Chinese Language Pedagogy reports the language gain of U.S. Chinese flagship students. His data shows that 44 % of these advanced learners of Chinese, after an overseas capstone experience (four months of direct enrollment in a Chinese university and three to five months of an internship in a Chinese company), exhibited a gain of one or two sub-levels on the ACTFL OPI. 50% of students, however, do not gain at all while 59% of these no-gainers approached Superior level at the pre-capstone point. Therefore, Bourgerie warns of the "ceiling effect" of the OPI. Interagency Language Roundtable (ILR) scores for these same students showed that among the students who completed an overseas capstone, 70% of them gained oned level, 10% of them gained two levels and 20% did not gain.

Other than the above-mentioned studies, there is a general gap in the field of SLA on the topic of Chinese acquisition and SA. Personal correspondence with Associated Colleges in China (ACC) revealed that ACC just began to collect pre and post proficiency data this year despite that it has been sending students to China for several decades. In addition, acquisition of the Chinese tones during SA has not been studied much, with the exception of a few studies by native Chinese scholars written in Chinese (Cheng, 2011; Wang, 2006). In order to fill this gap in the research, this study seeks to use quantitative measures to study Chinese SA participants' speaking gain. The following chapter, Chapter 3, lays out the design of the current study.

Chapter 3: Design of the Study

Chapter 3 describes how this study was designed to answer the following research questions: Do China study abroad (SA) participants' language skills change in terms of fluency during SA? If so, what measures of fluency change over time? This chapter first describes the subjects, and then the assessment tools for fluency measures are described. The last section of this chapter describes how data for this study were analyzed.

China Study Abroad—Dynamics

Participants. There were a total of twenty-four participants during the China Study Abroad Program. Among them, twenty students, fourteen males and six females, agreed to participate in this study voluntarily. Nineteen participants spoke English as their first language while one participant's first language was Spanish. The participants' language background varied. Five male participants spoke Mandarin Chinese extensively for two years (three in Taiwan, one in England, and one in Canada; those who lived in England and Canada mainly worked with Chinese immigrants). Four male participants spoke Cantonese extensively in Hong Kong for two years. Two male participants spoke Spanish in Spanish-speaking countries for two years. Four female participants studied Chinese since high school. For two male participants, this was their second China SA after three years.

The participants were required to finish Chinese 202 (intermediate Chinese) before participating in this program, but some students had advanced as far as the Chinese 301 course before the fall departure. Two students came from BYU-Idaho and one from BYU-Hawaii.

The students came from a variety of majors, as the program was open to anyone who completed intermediate Chinese courses. Even though students' grades were not a decisive factor, they were interviewed and chosen by the program director and therefore it is assumed that

their overall grades, motivation, and attitudes are higher than the applicants who were not accepted for the program. Once these students were chosen, they were enrolled in a weekly two-hour pre-SA class for two months where they prepared themselves to live and travel in China by learning and discussing differences in American and Chinese culture, Chinese history, and geography.

China SA classes. All students were assigned to one of three levels (A, B, or C class, A being the highest) after they took a placement test in the beginning of the semester. The placement test was an extensive reading and writing test. The test assessed both spoken (口语 kǒuyǔ) and written (书面语 shūmiànyǔ) languages of Chinese (differences in these two forms can be drastic in the Chinese language).

The China SA classes were composed of 20 hours of language instruction and 4 hours of Chinese culture instruction a week. The Chinese Culture class was conducted mostly in English. Students could take additional courses of their choice, in Chinese, for one to three hours a week.

The structure of each level was the same, composed of speaking and reading classes except for the A class, which had an additional writing class once a week. Speaking classes focused heavily but not exclusively on spoken Chinese while the reading classes focused heavily but not exclusively on written Chinese. The teachers were all native Chinese speakers with at least a few years of experience in teaching Chinese as a second language, except for one beginning teacher. The textbooks were chosen and sometimes written by Nanjing University.

Out-of-classroom learning. All China SA participants lived off campus with other study abroad participants. Nanjing University aided students' out-of-the-classroom learning by assigning each participant to a study buddy in the beginning of the semester. There was no formal follow up with the study buddy program. Two tutors helped students with their

homework and class work twice a week, with each group tutoring session lasting for two to three hours at a time. Other than the study buddy program and the teaching assistants, students were responsible for seeking their own opportunities to interact with native Chinese people.

Students took trips to various parts of China such as Guilin, Huangshan, Xi'an, and Beijing, with each trip lasting anywhere from a few days to a week. The total travel time to other cities was about a month, or approximately a third of the total time of the China SA. The participants also took weekend trips to various historical sites in Nanjing.

Study Design

This study incorporates quantitative measures in order to discover SA students' language gains in fluency. Quantitative data for this study were composed of two pre- and post-program assessments: a short, modified version of the Simulated Oral Proficiency Interview (SOPI) tasks and complete ACTFL OPIs (American Council on the Teaching of Foreign Languages [ACTFL], 2011).

SOPI (Simulated Oral Proficiency Interview). Three items from the Simulated Oral Proficiency Interview (see Appendix) were given at the beginning and at the end of the participants' stay in China to measure for speaking gains. The questions involved describing and expressing opinions. A total of twenty-four students took the three SOPI items.

The three SOPI items were chosen to provide consistency in data in difficulty and tasks. Since this study data was collected as a part of larger study that compares those SOPI items across various languages, these identical questions were given to all subjects to allow one to better compare language across individuals and languages (compared to the OPI,where questions asked vary greatly among subjects). In addition, these SOPI items allowed students to

demonstrate some linguistic ability via a much shorter measure, compared to the full OPI. The three items varied in difficulty.

Oral Proficiency Interviews (OPIs). Another measure used to assess speaking gain was the OPI. The OPI, as outlined in the ACTFL website, is "a standardized procedure for the global assessment of functional speaking ability" (ACTFL, 2011). Thus this measure was chosen to assess the increase in the global speaking ability of China SA students. Students were encouraged to take the ACTFL Oral Proficiency Interview (OPI) before beginning the China SA program and right before it was completed. All the OPI tests were administered over the phone and were rated twice by certified raters. A total of six students responded and participated in the pre- and post-OPI tests. Such low level of participation is due to the fact that participation was voluntary. Few participants found adequate time to participate in both pre-and post-OPIs, as they were busy preparing for departure to China over the summer and for finals during last few weeks in China.

Data Analysis

The above section discussed what types of data are used for this study. This data analysis section describes how the SOPI tasks and the OPIs were used to measure fluency gains.

Analyzing quantitative data—SOPI tasks. SOPI items were analyzed extensively using a series of fluency measures. We first attempted to use all seven of Freed's (1995b) factors of fluency, but found very few instances of repairs and clusters of dysfluencies, so we eliminated them. This is because there were many students' speech recordings that simply did not display repairs or clusters of dysfluencies due to nature of the task. This resulted in many zeros in the raw data, which would have lead to a skewed analysis.

In addition to the five fluency measures from Freed's (1995b) study, three other fluency measures were added to this study—native speaker ratings of overall fluency, task completeness, and tonal accuracy. Overall fluency ratings were added to see if there was any difference in native Chinese speakers' holistic perception of China SA participants' speech in the pre-and post-measures. Holistic task completeness ratings were added to see if the students' ability to perform tasks in the target language changed along with their fluency. Tonal accuracy, measured as the number of tonal errors, was added as an additional measure, to see if the SA participants' speech accuracy was affected by a SA.

The results of each participant's fluency measures from the three SOPI tasks were averaged and then combined as a group total and for each fluency measure. Then the averaged pre- and post-measures for all three SOPI items as a group were compared. The three native judges' ratings for each participant were averaged for overall fluency and task completeness, as well as the number of tonal errors per minute.

A series of mixed linear model ANOVAs was conducted (blocking on individuals) to determine whether there were significant pre-to-post differences on each of the measures studied. Likewise, mixed linear model ANOVAs were used to investigate whether there were differences in all the measures across the three SOPI tasks. Since time and task can have simultaneous effects on the measures of this study, the interaction effect between task and time was evaluated through mixed linear model ANOVAs. In other words, the ANOVAs were used to analyze whether the three tasks showed different patterns of development over time.

Fluency measures. The following fluency measures were used to assess China SA participants' fluency gains. The data analysis follows procedures from Freed (1995), even though not all of Freed's measures were used. Most measures below are per-minute measures since

students recorded for different durations of time. Full-length recordings were used since selecting a portion of speech may interfere with task completeness and fluency subjective ratings.

Word count (Tokens). The ability to produce more words within a set amount of time is a measure of fluency. The researcher used an online Chinese word list application (zhtoolkit.com) to measure the total word count in each SOPI item. Since Chinese words are the basic unit of meaning, not characters per se, it was assumed that counting words is more useful than counting characters. After counting the total number of words, the researcher then calculated word count per minute by dividing the number of words and unique words by the total length of recording time.

Unique word count (Types). Freed's (1995b) study and Freed et al.'s (2004) study both measured raw frequencies of unique words and the number of unique words per minute as the amount of speech and the rate of speech. Thus, this measure was chosen as one way of gauging students' fluency in SOPI tasks. The researcher used zhtoolkit.com for a unique word count and then unique words per minute was calculated through the same process as described above for word count per minute.

Filler word count. Freed's (1995b) study employed the frequency of filled pauses as a measure of fluency. This denotes lexical and non-lexical fillers. Lexical fillers are words such as "那个" (nà ge: that; Chinese filler word), "就是" (jiùshì: that is; Chinese filler word). Non-lexical fillers are meaningless words such as "uh," "um," "hmm," etc., that are found in between meaningful phrases or sentences. The total number of these filler words was counted and then divided by the total length of recording time.

Unfilled pauses. Unfilled pauses are defined as "silence which sounded dysfluent" (Freed, 1995, p.130). This data was analyzed using acoustic analysis software, Praat. Since the methods of silent pause analysis of this study is modeled after the methodologies of De Jong and Wempe (2009) and Segalowitz (2010).This study adopts these two studies' definition of pause as silence longer than .2 seconds, rather than Freed's (1995b) .4 seconds. For this analysis, all speech samples that had too poor of sound quality (i.e. too much noise in the background, making it hard to understand speech) were excluded.

Mean pause length. The mean length of filled and unfilled pauses was calculated through Praat. Poor sound quality speech samples were excluded for this analysis as well.

Holistic fluency rating. Three native Chinese raters gave holistic ratings of fluency and task completeness for each speech sample on how fluent the participant sounds on a scale of 1 to 10 (1 being no fluency and 10 being perfect fluency). The raters were asked to listen to each speech sample and first rate on fluency. The raters were asked to listen to each speech sample and first rate on fluency. During the rater training meeting, the researcher used the term 流利 (liúlì: fluent) to denote fluency. Other than the simple scale, no other explanation on fluency was given.

Holistic task completeness rating. Three native Chinese raters gave holistic ratings on task completeness after they gave fluency ratings. They did so based on how well the participants accomplished the task, on a scale of 1 (no task completion) to 10 (perfect task completion). Task completeness is not a fluency measure that has been widely used by researchers before, but this researcher saw the possibility that task completeness will affect raters' perception of fluency (a student who answered a question more completely and thus more successfully accomplished a language task may sound more fluent to the native speaker judges than a student who only

partially answered a question and incompletely fulfilled a language task)..The raters were given two piles of SOPI transcriptions (pre- and post-program) at a time and were aware of which transcriptions they were analyzing.

Tonal accuracy. Freed's definition can include accuracy, tonal accuracy was choseAfter the native Chinese judges gave holistic ratings for fluency and task completeness, they were asked to measure tonal accuracy. Each rater was asked to circle each word that a participant pronounced in inaccurate tones on the transcription and put correct tonal marks on the top of each word. These raters and the researcher counted the total number of tonal errors, other mistakes such as word usage or grammatical errors in each item for each participant.

The raters were given SOPI transcriptions and were asked to compare the speech samples with the transcriptions. The SOPI transcriptions were prepared by a native Chinese speaker first and then an advanced non-native Chinese speaker so that all the attempts made by participants to speak could be preserved and recorded. Each sample was transcribed twice: a native Chinese speaker transcribed first and then an advanced non-native Chinese speaker (whose proficiency level is close to that of native Chinese person) transcribed secondly, transcribing all the attempts to speak that could not be captured by the native Chinese speaker. SOPI transcriptions depicted the participants' speech as close as possible. In other words, all the speech errors and usage of English words were recorded.

Analyzing quantitative data—the OPIs. The differences between pre- and post- OPI ratings were compared, employing the procedures of quantifying the OPI results used by Meredith (1990) and Rifkin (2005). First, a number from 1 to 10 was assigned to sub-levels of OPI ratings (1 for Novice-Low and 10 for Superior). Then the researcher subtracted the pre-OPI

rating number from that of post-OPI for each student. This was used to gauge how much students' language proficiency has increased.

Chapter 4: Study Results and Discussions

This section of the chapter shows the quantitative results of mixed linear model ANOVAs, least squares means, standard errors, and significance values for SOPI tasks. The pre- and post-program OPI ratings are also reported. The discussion of the study results— its importance and implication follows.

Results of Fluency Measures

The results of fluency measures—holistic fluency ratings, word count, unique word count, filler words count, unfilled pauses, mean pause length, task completion, and tonal accuracy—in SOPI tasks are reported below.

Word count (Tokens). Table 1.1 shows the least squares means and standard errors for word count on each of the questions for both pre and post-testing.

Table 1.1

Least Squares Means and Standard Errors (in Parentheses) for Word Count

	Pre-Test	Post-Test	Combined Pre and Post Average
Task 1	73.38 (4.884)	83.33 (4.884)	78.35 (4.653)
Task 2	68.83 (4.884)	82.78 (4.884)	75.81 (4.653)
Task 3	68.58 (4.939)	81.05 (4.939)	74.81 (4.688)
Average of 3 Tasks	70.27 (4.580)	82.38 (4.580)	

As Table 1.2 depicts, only time was significant for word count, but task and the interaction between time and interaction were not. This means that students produced more words in post-SOPI tasks than pre-SOPI tasks in general, but the word production across the tasks was not significantly different.

Table 1.2

Mixed Model ANOVA Results for Word Count

	Num DF	*Den DF*	*F*	*p*
Time	1	101	48.37	< .0001
Task	2	101	1.44	0.2408
Time * Task	2	101	0.46	0.6296

Unique word count (Types). The least squares means and standard errors for unique word count on each of the questions for both time measures are shown in Table 2.1.

Table 2.1

Least Squares Means and Standard Errors (in Parentheses) for Unique Word Count

	Pre-Test	Post-Test	Combined Pre and Post Average
Task 1	42.81 (2.657)	43.97 (2.657)	43.39 (2.470)
Task 2	35.50 (2.657)	42.00 (2.657)	38.75 (2.470)
Task 3	35.92 (2.701)	38.96 (2.701)	37.44 (2.498)
Average of 3 Tasks	38.08 (2.410)	41.65 (2.410)	

As Table 2.2 provides evidence for, both time and task were significant for unique word count, but the interaction between the two was not. Accordingly, the China study abroad (SA) participants used more unique words during their post-SOPI tasks than during pre-SOPI tasks. The Tukey-Kramer post-hoc analyses reported $p<0.5$ in both cases, which means that, statistically speaking, students used more unique words for Task 1 than the other two tasks.

Table 2.2

Mixed Model ANOVA Results for Unique Word Count

	Num DF	*Den DF*	*F*	*p*
Time	1	101	9.66	0.0024
Task	2	101	9.82	0.0001
Time * Task	2	101	1.92	0.1516

Filler words count. The least squares means and standard errors for filler word count on each of the pre- and post-SOPI tasks are shown in Table 3.1. Table 3.1 reveals that students used more filler words in the post-test than in the pre-test, with the most filler words used in Task 1, fewer in Task 2, and the least in Task 3.

Table 3.1

Least Squares Means and Standard Errors (in Parentheses) for Filler Word Count

	Pre-Test	Post-Test	Combined Pre and Post Average
Task 1	7.69 (1.033)	9.19 (1.033)	8.44 (0.926)
Task 2	5.09 (1.033)	8.50 (1.033)	6.80 (0.926)
Task 3	6.15 (1.057)	7.49 (1.057)	6.82 (0.942)
Average of 3 Tasks	6.31 (0.891)	8.39 (0.891)	

As shown in Table 3.2, both time and task were significant for filler word count, but the interaction between the two was not. The amount of filler words that students used differed significantly across time and tasks. Since the interaction between time and task was not significant, the pattern in terms of changes in filler word count did not differ significantly between questions. The post-hoc analyses showed that students produced more filler words for Task 1 than Tasks 2 and 3 ($p<0.5$ in both cases).

Table 3.2

Mixed Model ANOVA Results for Filler Word Count

	Num DF	*Den DF*	*F*	*p*
Time	1	101	15.12	0.0002
Task	2	101	4.19	0.0179
Time * Task	2	101	1.56	0.2148

Unfilled pauses. Table 4.1 displays the least squares means and standard errors for unfilled pauses on each of the pre- and post-SOPI tasks.

Table 4.1

Least Squares Means and Standard Errors (in Parentheses) for Unfilled Pauses in Milliseconds

	Pre-Test	Post-Test	Combined Pre and Post Average
Task 1	32.83 (1.348)	33.59 (1.348)	33.21 (1.064)
Task 2	31.18 (1.348)	35.38 (1.405)	33.28 (1.083)
Task 3	30.81 (1.406)	36.43 (1.406)	33.62 (1.106)
Average of 3 Tasks	31.61 (0.960)	35.13 (0.970)	

As shown in Table 4.2, only time was significant for this measure; task and the interaction between time and task were not statistically significant. The number of unfilled pauses varied greatly between pre-and post-program SOPIs. The results of both Tables 4.1 and 4.2 indicate that China SA participants had more unfilled pauses for the post-SOPI than the pre-SOPI tests.

Table 4.2

Mixed Model ANOVA Results for Unfilled Pauses

	Num DF	*Den DF*	*F*	*p*
Time	1	101	12.91	0.0005
Task	2	101	0.06	0.9383
Time * Task	2	101	2.20	0.1157

Mean pause length. The least squares means in Table 5.1 and mixed linear ANOVA results in Table 5.2 show that time is significant for mean pause length. The participants paused significantly shorter during the post-test than during the pre-test. Task and Time*Task were not found to be significant.

Table 5.1

Least Squares Means and Standard Errors (in Parentheses) for Mean Pause Length

	Pre-Test	Post-Test	Combined Pre and Post Average
Task 1	0.80 (0.059)	0.75 (0.061)	0.78 (0.048)
Task 2	0.95 (0.061)	0.75 (0.064)	0.85 (0.050)
Task 3	0.95 (0.063)	0.83 (0.064)	0.89 (0.050)
Average of 3 Tasks	0.90 (0.044)	0.78 (0.045)	

Table 5.2

Mixed Model ANOVA Results for Mean Pause Length

	Num DF	*Den DF*	*F*	*p*
Time	1	101	8.02	0.0057
Task	2	101	2.15	0.1227
Time * Task	2	101	1.10	0.3358

Holistic fluency ratings. Table 6.1 shows the least squares means and standard errors for native speaker fluency rating on each of the questions for both pre- and post-testing. .

Table 6.1

Least Squares Means and Standard Errors (in Parentheses) for Subjective Native Speaker Fluency Ratings

	Pre-Test	Post-Test	Combined Pre and Post Average
Task 1	5.88 (0.309)	7.20 (0.309)	6.54 (0.291)
Task 2	5.67 (0.309)	6.98 (0.309)	6.33 (0.291)
Task 3	5.51 (0.313)	6.81 (0.314)	6.16 (0.294)
Average of 3 Tasks	5.68 (0.286)	7.00 (0.286)	

Note. 1= no fluency, 10= perfect fluency

As Table 6.2 depicts, both time and task were significant for native speaker subjective fluency ratings, but the interaction between the two was not. In other words, the pattern in terms of changes in fluency did not differ significantly between questions. Student pre- to post-program testing revealed significant gains in fluency ratings. Overall, post-hoc analyses (Tukey-Kramer) showed that students were rated as more fluent for Task 1 than Tasks 2 and 3 ($p<0.5$ in both cases)

Table 6.2

Mixed Model ANOVA Results for Subjective Native Speaker Fluency Ratings

	Num DF	*Den DF*	*F*	*p*
Time	1	101	117.57	<.0001
Task	2	101	3.24	0.0431
Time * Task	2	101	0.00	0.9976

Holistic task completeness ratings. Table 7.1 displays the least squares means and standard errors for native speaker task completeness ratings on each of the pre- and post- SOPI tasks.

Table 7.1

Least Squares Means and Standard Errors (in Parentheses) for Subjective Native Speaker Task Completeness Ratings

	Pre-Test	Post-Test	Combined Pre and Post Average
Task 1	6.38 (0.332)	7.52 (0.332)	6.95 (0.302)
Task 2	5.65 (0.332)	7.17 (0.332)	6.41 (0.302)
Task 3	5.52 (0.338)	6.68 (0.338)	6.10 (0.307)
Average of 3 Tasks	5.85(0.293)	7.12 (0.293)	

Note. 1= no completion 10= perfect completion

As shown in Table 7.2, both time and task were significant for the native speaker subjective task completeness ratings. In other words, students received significantly higher task completeness ratings in the post-testing than pre-testing. The interaction between time and task was not significant as shown in Table 7.2. Put differently, the pattern in terms of changes in task completeness over time did not differ significantly between questions. The Tukey-Kramer post-hoc analyses showed that Task 1 was significantly easier to complete than tasks 2 and 3 ($p<0.5$ in both cases).

Table 7.2

Mixed Model ANOVA Results for Subjective Native Speaker Task Completeness Ratings

	Num DF	*Den DF*	*F*	*p*
Time	1	101	63.75	<.0001
Task	2	101	9.41	0.0002
Time * Task	2	101	0.59	0.5543

Tonal accuracy. As depicted in Table 8.1, the least squares means for tonal errors count by native speakers on each task is lower in post- testing than pre- testing. That is, participants generally had fewer tonal errors in the post- SOPI questions than the pre-SOPI questions.

Table 8.1

Least Squares Means and Standard Errors (in Parentheses) for Tonal Errors Count by Native Speakers

	Pre-Test	Post-Test	Combined Pre and Post Average
Task 1	5.73 (1.003)	4.03 (1.003)	4.88 (0.951)
Task 2	4.32 (1.003)	3.55 (1.003)	3.94 (0.951)
Task 3	4.60 (1.016)	3.75 (1.016)	4.18 (0.959)
Average of 3 Tasks	4.88 (0.934)	3.78 (0.934)	

As shown in Table 8.2, the number of tonal errors during pre- and post- testing was significantly different. However, the tonal accuracy did not significantly vary for the three SOPI tasks. That is to say, the participants made similar amounts of tonal errors in the three tasks. The interaction between time and task was not significant either.

Table 8.2

Mixed Model ANOVA Results for Tonal Errors Count by Native Speakers

	Num DF	*Den DF*	*F*	*p*
Time	1	101	8.71	0.0039
Task	2	101	2.32	0.1030
Time * Task	2	101	0.64	0.5280

OPI ratings

The OPI ratings gave some insights on how some participants' language proficiency has changed. Out of five students who successfully took both pre- and post-OPIs, four students

received post-OPI ratings that were two sub-levels higher than the pre-OPI. One student received the same ratings for both tests. Out of the four gainers, two progressed from Advanced-Low to Advanced-High. The other two students progressed from Intermediate-Mid to Advanced Low.

Discussions

The mixed linear model ANOVA results show that fluency, task completeness, unique words, and filler words were significant for both time and tasks. The measures of tonal accuracy, word count, unfilled pauses, and mean pause length were only significant for time. These results are summarized in the following Table, Table 9. There were no significant Time*Task interactions for tonal accuracy, word count, unfilled pauses, and mean pause length. This means that the patterns in changes from pre to post did not vary significantly by tasks in these measures.

Table 9

Time only and both Time and Task Significant Measure of Fluency

Time Significant	Time and Task Significant
Tonal Accuracy	Fluency Ratings
Word Count	Task Completeness Ratings
Unfilled Pauses	Unique Word Count
Mean Pause Length	Filler Words

It is noteworthy that China SA students' SOPI recordings showed statistically significant differences across time for all the fluency, accuracy, and completeness measures of this study. This suggests that

1. The students produced more words, especially unique words, at the end of the program than the beginning.
2. The participants also paused for shorter durations of time in their speech at the end of the program than in the beginning.
3. The native judges perceived that students sounded more fluent in post-program tests than in pre-program tests.
4. The native judges perceived that the SA participants accomplished each task more completely at the end of the SA than in the beginning.
5. The data also reveal that the SA was a benefit to the students' tonal accuracy. The students made fewer tonal errors at the end of the program.

These five patterns are all in line with intuition that SA would help students improve fluency. What is counter-intuitive is that the students had a larger number of unfilled pauses and used more filler words during the post-SOPI than the pre-SOPI. These results deviate from those by Freed (1995b), who found a significantly smaller number of unfilled pauses and filled pauses in SA participants' speech than in the speech samples of those students who received instruction at home. It is assumed that this discrepancy is due to several possible reasons: 1) increased word production related to pauses, 2) patterns in the use of filler words in different cultures, and 3) personal characteristics of the learners. First of all, increased number of pauses may be related to increased production of words. This study adopted a different definition of unfilled pauses (.2 seconds) than Freed's (.4 seconds), following the same technical procedures as De Jong and Wempe (2009) and Segalowitz (2010). While there are varied opinions among scholars regarding which definition of silent pause is more accurate, Riggenbach (1991), like Freed

(1995b), argued that speech with pauses shorter than .4 seconds is assumed normal and should not be considered dysfluent. Therefore, it is likely that very brief pauses between word productions were captured and recorded as unfilled pauses in the current study. It means that the more words a participant produced, the more unfilled pauses occurred as well.

Another reason for difference may be different perception of pauses in cultures. While pause is not necessarily encouraged to English speakers in the U.S., some cultures like Japan assumes that the use of pause may be an expression of one's thoughtfulness and humility. Last of all, students who tend to pause more in their speech may have adopted this habit in speaking in a foreign language. In summary, the issue of increased filled and unfilled pauses in this study data may not only be a language issue, but also methodological or cultural.

Native Judges' Definition of fluency

After the native judges finished rating the SOPI items, they were asked to give a definition of fluency in their own words. They were asked to do so in either English or Chinese so that they could best express their thoughts. Their initial response were translated from Chinese to English and English mistakes were corrected when they used English. The three native judges defined fluency in the following ways and the results are shown in an illustrative summary:

- The standard of fluency is accurate pronunciation, proper word usage, and speech that is easily understood by native speakers. Fluency can be improved by reading in the target language frequently, conversing in the language, and memorizing some good sentences or writings (Personal communication, August 2, 2011). —Rater 1

- From my point of view, people who speak fluently not only speak fast, but also speak with right intonation, pausing at the right places, and using proper words. —Rater 2
- I considered a lot of elements when it came to fluency. Tone, choice of vocabulary, pace of speech, and choice of transitions as well as classification and speakers' confidence in speech. —Rater 3

As shown above, the native judges perceived fluency as a broad concept that includes pronunciation, tones, word usage, intonation, pace of speech, comprehensibility, and confidence. The only measure of fluency that matched native judges' definition was pauses. Even though these three natives' opinions cannot be generalized, they provide an insight to how natives perceive fluency. These judges perceived fluency as general competency in language skills which include linguistic accuracy. Their definitions therefore provide rationale for investigating other speaking measures such as tonal accuracy, in this study.

Chapter 5: Summary and Conclusion

Implications

This study attempted to capture some of the changes that occurred during China SA for students from an American university. The results of this study provide evidence that participants of this program have made significant fluency gains: Native judges perceived learners' speech to be more fluent and more complete, and to contain fewer tonal errors at the end of the SA program. Students also produced more words and unique words per minute in their utterances by the end of the program. Their mean pause length was shorter as well. Students produced more number of filled and unfilled pauses in their speech at the end of the program than the beginning. This may be due to methodological, cultural, or personal reasons. Overall, all the fluency measures used in this study showed significant differences between pre-and post-program tests.

The OPI ratings also display students' gain in overall language abilities. Out of five students who successfully took both pre- and post- OPIs, four students gained two sub-levels above their initial ratings even though one student failed to gain. These results are impressive overall, knowing many studies such as Freed's (1990) and Milleret's (1991) reported that the OPI tests failed to capture short-term language gains.

These results coincide with initial assumptions about the study. Since SA students spend a significant amount of time each week in language classes and are surrounded by target language speakers and culture, the researcher assumed that students will gain in fluency measures, if not all, some. Objective fluency measures and subjective native judges' ratings both confirm that this assumption is validated.

Since the acquisition of Chinese tones is assumed to be very challenging by researchers who study this topic, the effect of SA on tonal accuracy was unknown. The results show that SA

also had a beneficial influence on Chinese tonal accuracy. Along with tonal accuracy, students were perceived to have significantly improved their abilities to complete language tasks.

In the Literature Review section of this paper, the researcher mentioned the general lack of literature in SA and Chinese language acquisition. This study provides valuable data on fluency development during SA in China.

Limitations of the Study

This study can be improved by overcoming some shortcomings. First of all, one could decrease the possibility of rater bias by improving the method of native ratings. As three native raters assessed fluency, completeness, and tonal accuracy of China SA participants' SOPI items, they were given two piles of audio recordings—pre- and post-SOPIs. Their ratings might have been affected by their expectation that students must have improved linguistically after the SA. It is also possible that some students' improvement might have affected raters' overall judgment of other students who did not necessarily gain linguistically. For a more accurate rating of students' performance, it would be wise to mix up pre- and post-SOPI recordings and have raters assess them in a random order. In this way, the bias based on expectations for students' level of fluency before and after SA can be eradicated.

This study could have better captured the participants' language gain through understanding their curriculum and incorporating assessment tools that better reflect SA students' language gain during SA. The questions asked during the OPIs and the three SOPI tasks may not have much to do with what students have been learning in their classes. Thus, the criticism regarding the OPI being an insufficient measurement tool for language can be addressed by using more focused assessment tools. Kinginger notes that, "[d]espite claims to the effect that the OPI measures context-independent language ability, in other words, each

individual test is an assemblage of particular topical or functional tasks... [It]... may or may not correspond to students' history of language use..." (Kinginger, 2008, p.47). In order for the OPI to measure language ability based on learners' history of language use, "... more finely-tuned analysis" will be needed. "...those which will reveal, with specificity, development in students' lexical breadth, syntactic complexity, and cohesion and coherence in language use." (Freed 1990, p. 475).

Suggestions for Future Research

Despite the shortcomings of this study, some aspects may suggest ideas for future research. First of all, future researchers may be interested in further discovering fluency and accuracy gains in the Chinese language. Tonal accuracy was the only measure of linguistic accuracy in this study. It would be interesting to study other aspects of linguistic accuracy gains, such as phonological accuracy (Chinese sounds) and intonation. Other than pronunciation, grammatical accuracy, word usage, or sociolinguistic accuracy (cultural appropriateness) could also be studied as accuracy measures.

With a scant amount of research on reading gain in Chinese and the complicated Chinese orthography being a constant challenge to even advanced learners of Chinese, reading gain during SA would be a very interesting additional topic to investigate. Dewey (2004) asserts that students of Asian languages can benefit from SA for their reading comprehension abilities. A possible replication of his study in Chinese to see if there is any difference in reading gains in the Japanese and Chinese languages may be feasible. Dewey's (2004) study compared SA students with IM and AH students. It would be interesting to replicate this study design and compare students who go on SA in China and those who pursue their study of Chinese in domestic institutions or in immersive settings.

Future researchers may build on the findings of this study through related data regarding factors that might contribute to language development, such as personality, language use, and social networks. For example, future researchers may consider investigating the relationship between personality and fluency gains through administering a personality survey before and after a SA program. Similarly, studying the extent of the participants' language use and social networks through measures such as a language log and a social network survey can allow researchers to compare these data with the students' speaking gains (see, for example, Martinsen et al., 2011; Freed, Segalowitz, & Dewey, 2004).

Qualitative studies on SA participants' speaking gain could be used to corroborate the quantitative findings on fluency and other speaking gains in this study. Wilkinson (2005) argued that SA researchers should look at the "depth dimension" of SA, rather than taking a pre- and post-test approach, closely examining how individual students view their SA experience. Stewart (2010), who advocates the use of e-journals to assess SA experience, argues that the "SA is often the 'black hole' semester when educators lose contact with students for a ...period just when their language and personal identity as L2 speaker are undergoing the greatest change" (Stewart, 2010, p. 141). Language contact logs, diaries/journals, oral interviews, and on-site observations can help us understand SA participants' view of their fluency and development of speaking skills. These methods can help us understand more thoroughly how native target language speakers view fluency. Case studies on SA students' development of fluency may be helpful since case studies often have the advantage of capturing idiosyncrasies and linguistic, attitudinal, and other personal changes that may be imperceptible via standardized or quantitative measures (Kinginger ,2004, Douglass, 2006, Pellegrino, 2005, 2006). Qualitative research on speaking

gains can show a lot more than pre- and post-program test scores—,they can show how SA participants and native speakers' perceptions of fluent speech develop and possibly change.

This study reveals the three native judges' definitions of fluency. Since the native judges may not be aware of how the participants usually speak, the judges may impose their own definition of fluency as they rate speech samples. It would be helpful for future researchers to collect the participants' L1 speech samples as baseline data for comparison. It would also be beneficial if future researchers could provide example L2 speech samples and their ratings to define rating scales more distinctively for the NS raters.

As more and more students go on SA programs, it is critical that the participants understand how they can learn a target language successfully. Those who hope to make speaking gains during a SA program may consider how natives perceive fluency and consciously try to improve in various fluency measures. For example, a foreign language learner can put in a conscious effort to reduce the usage of filler words and to incorporate proper word usage, shorter pause length, etc. To increase speech accuracy, receiving feedback from native speaker(s) might be helpful for learners to better understand how to develop more fluent and native-like speech.

Chinese SA programs may consider informing program participants regarding what native speakers' general perceptions of fluency are. They may do so by providing various speech samples and supporting each factor of fluency with specific examples. Chinese SA programs can also help students to increase in tonal accuracy throughout SA. Pronouncing tones accurately can be challenging even to advanced speakers of Chinese, who develop fossilized errors as beginning Chinese learners struggling to comprehend relative difference of tones. As Cheng (2011) suggested, SA programs do so by incorporating tones as a consistent part of their curricula.

Researchers who are interested in studying speaking gains during SA, especially in Asian languages, may gain insights from this study as how to construct a study design to capture participants' growth in speaking abilities.

References

Allen, H. and Herron, C. (2003). A mixed methodology investigation of the linguistic and affective outcomes of summer study abroad. *Foreign Language Annals, 36*(3), 370-85.

American Council on the Teaching of Foreign Languages. About the ACTFL OPI. Retrieved from http://www.actfl.org/i4a/pages/index.cfm?pageid=3348

Bailey, K., and Ochsner, R. (1983). A methodological review of the diary studies: Windmill tilting or social science? In K. Bailey, M. Long, and S. Peck (Eds.), *Second language acquisition studies* (pp.188-98). Rowley, MA: Newbury House.

Bennett, M. (1986). A developmental approach to training for intercultural sensitivity. *International Journal of Intercultural Relations, 10*, 179-96.

Brecht, R. D. and Davidson, D. (1991). *Language acquisition gains in study abroad: Program assessment and modification.* Paper presented at the NFLC Conference of Language Testing. Washington D.C.

Brecht, R. D., Davidson, D. and Ginsburg, R. (1995). Predictors of foreign language gain during study abroad. In B. F. Freed (Ed.), *Second language acquisition in a study abroad context* (pp. 317-34). Philadelphia, PA: John Benjamins..

Carroll, J. (1967). Foreign language proficiency levels attained by language majors near graduation from college. *Foreign Language Annals, 1*(1), 131-151.

Central Intelligence Agency, The World Fact Book- China. Retrieved from https://www.cia.gov/library/publications/the-world-factbook/geos/ch.html

Cheng, X. (2011). 程潇。浅析外国留学生习得汉语声调的难点及教学策略. Retrieved from http://www.lwlm.com/yuyanxue/201102/542663.htm

Christensen, M. and Warnick, J. P. (2006). *Performed culture: An approach to East Asian language pedagogy.* Columbus, OH: National East Asian language resource center, Ohio State University.

Davidson, D. & Ginsburg, R. (1995). On the value of formal instruction in study abroad: Student reactions in context. In B. F. Freed (Ed.), *Second language acquisition in a study abroad context* (pp. 317-334). Philadelphia, PA: John Benjamins Publishing Company.

Davidson, D. (2010). Study abroad: When, how long, and with what results? new data from the Russian front. *Foreign Language Annals, 43*(1), 6-26.

De Jong, N.H. & Wempe, T. (2009). Praat script to detect syllable nuclei and measure speech rate automatically. *Behavior Research Methods, 41*(2),385 - 390.

DeKeyser, R. (1991). Foreign language development during a semester abroad. In B. F. Freed (ed.), *Foreign language acquisition research and the classroom* (pp. 104-31). Lexington, MA: D.C. Heath.

DeKeyser, R. M. (1997). Beyond explicit rule learning: Automatizing second language morphosyntax. *Studies in Second Language Acquisition, 19*, 195-221.

Dewey, D. P. (2004). A comparison of reading development by learners of Japanese in intensive domestic immersion and study abroad contexts. *Studies in Second Language Acquisition, 26*(2), 303-27.

Dewey, D. P. (2006). Reading comprehension and vocabulary development in orthographically complex languages during study abroad. In S. Wilkinson (Ed.), *Insights from study abroad for language programs* (pp. 72-83). Boston: Thompson Heinle.

Douglass, K. (2006). From the learner's perspective: A case study on motives and study abroad. In S. Wilkinson (Ed.), *Insights from study abroad for language programs* (pp. 116-33). Boston: Thompson Heinle.

Dubiner, D. Freed, B, & Segalowtiz, N. (2007). Native speakers' perceptions of fluency acquired by study abroad students and their implications for the classroom at home. In S. Wilkinson, *Insights from study abroad for language programs* (pp.3-21). Boston: Thompson Heinle.

DuFon, M. & Churchill, E. (Eds.) (2006). *Language learners in study abroad contexts*. Clevedon, UK: Multilingual Matters.

Foltz, D. (1991). *A study of the effectiveness of studying Spanish overseas*. Paper presented at the Pennsylvania State Modern Language Association Annual Meeting, Pittsburgh, PA.

Fraser, C. (2002). Study abroad: An attempt to measure the gains. *German as a Foreign Language Journal, 1*, 45-65.

Freed, B. F. (1990). Language learning in a study abroad context: The effects of interactive and noninteractive out-of-class contact on grammatical achievement and oral proficiency. In J. Atlas (Ed.), *Linguistics, language teaching, and language acquisition: The interdependence of theory, practice, and research* (pp. 459-77). Washington, DC: Georgetown University Press.

Freed, B. F. (Ed.) (1991). *Foreign language acquisition research and the classroom*. Lexington, MA: Heath and Company.

Freed, B. F. (Ed.) (1995a). Language learning and study abroad. In B. F. Freed (Ed.), *Second language acquisition in a study abroad context* (pp. 3-34). Philadelphia: John Benjamins.

Freed, B. F. (Ed.) (1995b). What makes us think that students who study abroad become fluent?, In B. F. Freed (Ed), *Second language acquisition in a study abroad context* (pp. 123-48). Philadelphia, PA: John Benjamins.

Freed, B. F., Segalowitz, N. and Dewey, D. (2004). Contexts of learning and second language fluency in French: Comparing regular classrooms, study abroad, and intensive domestic programs. *Studies In Second Language Acquisition, 26*(2),275-301

Freed, B., So, S., and Lazar, N. (2003). Language learning abroad: How do gains in written fluency compare with gains in oral fluency in French as a second language? *ADFL Bulletin, 34*(3), 34-40.

Ginsburg, R. B. and Miller, L. (2000). What do they do? Activities of students during study abroad. In R. D. Lambert and E. Shohamy (Eds.), *Language policy and pedagogy: Essays in honor of A. Ronald Walton* (pp. 237-61). Philadelphia, PA: John Benjamins.

Hayden, J.J. (1998). The influence of a semester abroad on reading fluency: A descriptive study. *Journal of the Chinese Language Teachers Association, 33*(3), 12-14.

Huebner, T. (1995). The effects of overseas language programs: Report on a case study of an intensive Japanese course. In B. F. Freed (Ed.), *Second language acquisition in a study abroad context* (pp. 171-193). Philadelphia: John Benjamins.

Isabelli-García, C. L. (2006). Study abroad social networks, motivation, and attitudes: Implications for SLA. In M. A. DuFon and E. Churchill (Eds.), *Language learners in study abroad contexts* (pp.231-58). Clevedon, UK: Multilingual Matters.

Juan-Garau, M. J. and Pérez-Vidal, C. (2007). The effect of context and contact on oral performance in students who go on a stay abroad. *Vigo International Journal of Applied Linguistics (VIAL), 4*, 117-35.

Kinginger, C. (2004). Alice doesn't live here anymore: Foreign language learning as identity (re)construction. In A. Pavlenko & A. Blackledge (Eds.), *Negotiation of identities in multilingual contexts* (pp. 219-242). Clevedon, UK: Multilingual Matters.

Kinginger, C. (2007). Language learning in study abroad: Case histories of Americans in France. [online]. Retrieved from http://calper.la.psu.edu/LL_in_Study_Abroad_ summary.pdf

Kinginger, C. (2008). Language learning in study abroad: Case studies of Americans in France. *Modern Language Journal Monograph, 1*. Oxford: Blackwell.

Kinginger, C. (2009). Language learning and study abroad: A critical reading of research. Basingstoke, UK: Palgrave Macmillan.

Liskin-Gasparro, J.E. (1984). The ACTFL proficiency guidelines: A historical perspective. In T. Higgs (Ed.), "Teaching for proficiency, the organizing principle" (ACTFL Foreign Language Education Series, pp. 11-42). Lincolnwood, IL: National Textbook Co.

Magnan, S. (1986). Assessing speaking proficiency in undergraduate curriculum: Data from French. *Foreign Language Annals, 19*(5), 429-38.

Martinsen, R. A., Baker, W., Bown, J., Johnson, C. (2011). The benefits of living in foreign language housing: The effect of language use and second-language type on oral proficiency gains. *The Modern Language Journal, 95*(2), 274-290.

McGinnis, S. (1997). Tonal distinction errors by beginning Chinese language students: A comparative study of American and Japanese native speakers. In S. McGinnis (Ed.), *Chinese pedagogy: An emerging field* (pp. 81-92), Columbus, OH: The Ohio State University Foreign Language Publications.

Meara, P. (1995). Student attitudes to learning modern languages in the 1980s: Data from Nuffield Modern Language Inquiry, 1986. *University of Southampton Center for Language Education Occasional Papers, 36*. ERIC Document Reproduction Service ED 389 197.

Meredith, R. A. (1990). The oral proficiency interview in real life: Sharpening the scale. *The Modern Language Journal, 74*(3), 288-296.

Milleret, M. (1991). Assessing the gain in oral proficiency from summer foreign study. *ADFL Bulletin, 22*(3), 39-43.

Miracle, W. C. (1989). Tone production of American students of Chinese: A preliminary acoustic study. *Journal of the Chinese Language Teachers Association, 24*(3), 49-65.

Möehle, D. & M. Raupach. (1983). *Planen in der Fremdsprach*. Frankfurt: Peter Lang.

National Center for Education Statistics. Indicator 40 (U.S. students studying abroad). Retrieved from http://nces.ed.gov/programs/coe/2010/section5/indicator40.asp

Ochsner, R. (1979). A poetics of second language acquisition. *Language Learning, 29*(1), 53-80.

O'Brien, I., Segalowitz, N., Freed, B, & Collentine, J. (2007). Phonological memory predicts second language oral fluency gains in adults. *Studies in Second Language Acquisition, 29*, 557-581.

O'Connor, N. (1988). Oral proficiency testing of junior year abroad: Implications for the undergraduate curriculum. Paper presented at the 1988 Annual Meeting of the MLA.

Omaggio, Alice C. (1986). *Teaching Language in Context: proficiency-oriented instruction*. Boston: Heinle & Heinle Publishers.

Peirce, B. N. (1995). Social identity, investment, and language learning. *TESOL Quarterly, 29*(1) 10-32.

Pellegrino Aveni, V. (2005). *Study abroad and second language use: Constructing the self.* New York, NY: Cambridge University Press.

Pellegrino Aveni, V. (2006). Speak for your self: Second language use and self-construction during study abroad. In S. Wilkinson (Ed.), *Insights from study abroad for language programs* (pp. 99-115). Boston: Thomson Heinle.

Polanyi, L. (1995). Language learning and living abroad: Stories from the field. In B. F. Freed (Ed.), *Second language acquisition in a study abroad context* (pp. 271-91). Philadelphia, PA: John Benjamins.

Praat: doing phonetics by computer [Computer Software]. Amsterdam, Netherlands: University of Amsterdam.

Rifkin, B. (2005). A ceiling effect in traditional classroom foreign language instruction: Data from Russian. *The Modern Language Journal, 89*(1)*,* 3-18.

Riggenbach, H. (1991). Towards an understanding of fluency: A microanalysis of nonnative speaker conversations. *Discourse Processes, 14*, 423–441.

Schumann, F. (1980). Diary of a language learner: A further analysis. In R. Scarcella and S. Krashen (Eds.), *Research in Second Language Acquisition* (pp. 51-7). Cambridge, MA: Newbury House.

Schumann, F. & Schumann, J. (1997). Diary of a language learner: Introspective study of SLA. In H. Brown, C. Yorio and R. Crymes (Eds.), *On TESOL '77* (pp. 241-9). Washington, DC: TESOL.

Scott, J. (2000). *Social network analysis: A handbook* (2nd Ed.). London: Sage Publications.

Segalowitz, N. and Freed, B. F. (2004). Contexts, contact and cognition in oral fluency acquisition: Learning Spanish in at home and in study abroad contexts. *Studies in Second Language Acquisition, 26*(2), 173-99.

Segalowitz, N. (2010). *Cognitive bases of second language fluency*. New York, NY: Routledge.

Shen, S. X. (1989). Toward a register approach in teaching Mandarin tones. *Journal of the Chinese Language Teachers Association, 24*(3): 27-47.

Stewart, J. A. (2010). Using e-journals to assess students' language awareness and social identity during study abroad. *Foreign Language Annals, 43*(1), 138-59.

Veguez, R. (1984). *The oral proficiency interview and the junior year abroad: Some unexpected results.* Paper presented at the Northeast Conference on the Teaching of Foreign Language. New York, April 1984.

Wang, M. (2006). Indonesian Chinese students' acquisition of Chinese tones. *Journal of College of Chinese Language and Culture*. Retrieved from http://en.cnki.com.cn/Article_en/CJFDTotal-JNHW200602003.htm

Wilkinson, S. (2002). The omnipresent immersion during study abroad: Some participants' perspectives. *Frontiers: The Interdisciplinary Journal of Study Abroad, 4*, 12-38.

Wilkinson, S. (2005). Articulating study abroad: The depth dimension. In C. M. Barrette and K. Paesani (Eds.), *Language program articulation: Developing a theoretical foundation* (pp. 44-58). Boston, MA: Thomson Heinle.

White, C. M. (1991). Tonal production and interference from English intonation. *Journal of the Chinese Language Teachers Association, 16*(2): 27-56.

Whitworth, K. F. (2006). Access to language learning during study abroad: The roles of identity and subject positioning. (Unpublished doctoral dissertation). Pennsylvania State University, State College, PA.

Willis, F., G. Doble, U. Sankarayya & A. Smithers. (1977). *Residence abroad and the student of modern languages. A preliminary study.* Bradford, UK: Modern Language Center, University of Bradford.

Yager, K. (1998). Learning Spanish in Mexico: The effect of informal contact and student attitudes on language gain. *Hispania, 81*(4), 898-913.

Zhtoolkit: Tools for studying Chinese. (Software) Available from http://www.zhtoolkit.com/apps/wordlist/create-list.cgi

Appendix- Three SOPI Items Instructions

Instructions for SOPI Tasks to be read to the student by the researcher assistant:

This brief measure of your language abilities contains just three questions. For each question, you will be given a scenario and asked to respond in the language you have been studying while abroad. Remember there are no wrong or right answers. This is an opportunity to show your ability to speak (Spanish, Russian, French, Chinese, Arabic etc.). Speaking less in order to make fewer mistakes will not help your performance. Try to always say as much as you can, speak as well as you can and show what you can do. Thanks for your help!!!!!!!!!!!!!

Item 1

Imagine that you are staying with native speaker friends in the country where you have studied abroad. These friends are having a birthday party. None of them have ever been to the U.S. They ask you to describe what birthday parties are like there. Please describe for them what birthday parties are like in the U.S. You may use the picture here to help you or you may draw on your own experiences. You will have 1 minute 30 seconds to respond.

Item 2

You are having a conversation with a professor in the country you will visit during your study abroad experience. During that conversation the topic turns to foreign language study in U.S. high schools. The professor asks you what you think the effects would be if every high school student in the U.S. were required to study a foreign language throughout high school. Tell the professor what you think the possible consequences might be if such a requirement were established. You will have 1 minute 30 seconds to respond.

Item 3 of 3

Imagine that you are discussing health care policies with a group of visitors to BYU from the country where you are studied abroad. One of the group members who has lived in Canada for several years complains that health care is too expensive in the U.S. She feels that it should be provided free of charge to all U.S. citizens, as in countries such as Canada. Another member of the group asks how you feel about this issue. Answer her question by indicating how you stand on free universal healthcare for U.S. citizens and explain why you hold this view. You'll have two minutes to respond.

Thanks for your time and participation!

Note: There was a picture prompt of a birthday party scene for SOPI Task 1, taken from the SOPI Manual for Chinese.