THE PENNSYLVANIA STATE UNIVERSITY
SCHREYER HONORS COLLEGE

DEPARTMENT OF PSYCHOLOGY

INTER-RATER RELIABILITY FOR THE JUDGED ACCENTEDNESS OF ENGLISH BILINGUALS

RACHEL MARIE SIGMUND
Spring 2010

A thesis
submitted in partial fulfillment
of the requirements
for a baccalaureate degree
in Psychology
with honors in Psychology

Reviewed and approved* by the following:

David A. Rosenbaum
Distinguished Professor of Psychology
Thesis Supervisor
Honors Adviser

Judith F. Kroll
Distinguished Professor of Psychology, Linguistics, and Women's Studies
Faculty Reader

* Signatures are on file in the Schreyer Honors College.

ABSTRACT

This study examined the extent to which native English speakers and native Chinese-English bilinguals differ in their abilities to perceive accents in the spoken English of other native and non-native English speakers. The main empirical question was whether proficient speakers of a language would agree more when judging the accentedness of another person speaking that same language than would less proficient speakers. This question arose from the hypothesis that speech perception and language proficiency share common representations and that the acquisition of those representations constitutes a core component of language learning.

The participants were two groups of 10 native-English speakers and two groups of 10 native-Chinese speakers of English each. All participants in the first group of native English speakers and in the first group of native Chinese speakers performed in four experiments. In Experiment 1, the participants indicated whether letter series were English words or nonwords. I used this lexical-decision task to assess the language proficiency of the two groups. In Experiment 2, participants responded to mathematical problems while trying to remember subsequent L1 words. I used this operations-span (O-span) task to assess working memory capacities. In Experiment 3, participants responded to the colors of stimuli while locations either were congruent or incongruent with respect to the responses that were made (on the same side as the response key or on the opposite side, respectively). I used this Simon task to assess executive functioning capabilities. Finally, in Experiment 4, one group of the 10 native English speakers and one group of the 10 native Chinese speakers recorded themselves speaking four sentences in English. Later, the other group of 10 native English speakers and the other group of 10 native Chinese speakers listened to these recordings and gave accentedness ratings for heard speaker's rendition of each sentence.

Experiment 1 revealed higher proficiency levels for native-English participants than native-Chinese participants. Experiment 2 showed that native-English participants recalled slightly more words than native-Chinese participants. Experiment 3 revealed a large range of scores for native-Chinese participants, and a smaller range for native-English participants. Native-Chinese and native-English raters gave equally high ratings to native-English speakers, with similar low means of agreement, which may reflect a ceiling effect. Native-Chinese raters gave higher ratings to the spoken English of the native-Chinese speakers than did native-English raters. The most important result of all was that native-Chinese raters had a lower mean agreement among their ratings for native-Chinese speakers than did native-English raters for the same speakers.

This study suggests that inter-rater reliability is a useful metric for exposing differences in expertise. Using inter-rater reliability may provide a new method for probing the abilities of speakers and listeners of different languages and also for teaching new languages.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

First and foremost, I would like to thank David A. Rosenbaum for his unrelenting support and guidance in all aspects of this project. From its brainstorming stages until now, thank you, David, for being there every step of the way. I would also like to thank Judith Kroll for her support and helpful comments and contributions. Without the both of you, David and Judy, this project would not have been feasible.

I would like to thank Joyce Tam for all of her assistance with data collection and scoring. Her patience and kind-hearted nature will never be forgotten. I also thank Mark Minnick for helping me to get my experiment up and running by providing me with any logistical information I needed. Furthermore, I thank Alison Spiro for her assistance in running participants, and offering advice and support in any way possible throughout the year. I could not have done it without you. Lastly, I would like to thank Sangdi Chen and Rachel Kashan for their useful contributions and insight throughout the earlier stages of my research.

Chapter 1

Introduction

It is widely accepted that as an individual becomes highly skilled at a particular task, he or she should also be able to reliably judge whether or not another individual engaging in that same task is skilled at it or not. Surprisingly, however, there has been little empirical research conducted to actually test this theory. This presumption stems from the idea that as an individual initially begins to learn a new skill, he or she most likely is not immediately concerned with that skill's ultimate goals. With time, however, the task's final goals become clearer, as a result of increased familiarization and practice. With a goal in mind, individuals can assess their performance in relation to that goal, and gauge their level of skill at that particular time period accordingly. In addition to assessing his or her own skill level, we believe that the individual in question should also be able to judge the skill level of others engaging in that task, as previously mentioned.

The present study instantiates this well-accepted belief within the context of language proficiency and speech perception. I believe that an individual who is highly proficient in any given language will also be able to judge whether or not someone else is speaking that same language with high proficiency. For example, a native-English speaker is cognizant of the ultimate goals of spoken English; he or she should be able to detect a number of phonetic errors that disrupt the language's true sound. For a second-language learner (L2) of English, however, the ultimate goals of the language may not be apparent at first, although with practice and increased familiarization, the L2 learners of English probably become aware of not only the language's words and spellings, for example, but also become aware of their spoken accent in

English. Learning a second-language is difficult in its own right, and perfecting that L2's spoken accent can be even more difficult.

Thus, to test our belief that with expertise in a skill comes increased recognition of the skill level of others engaging in that same skill, we decided to examine accentedness ratings of monolingual native-English speakers and native-Chinese-English bilinguals for the spoken English of other monolingual native-English and native-Chinese-English speakers. How will the accentedness ratings between the two groups of raters differ, if at all?

I hypothesized that for any two highly proficient speakers of a language (e.g., native-English speakers), they will also have high inter-rater reliability when judging the accentedness of others speaking that same language. Conversely, for any two less proficient speakers of a language (e.g., native-Chinese-English bilinguals), they will have less inter-rater reliability when judging the accentedness of others speaking that same language. If my belief is correct that expertise in a skill leads to the accurate evaluation of others performing that skill, then the native-English raters in the study should have a higher degree of agreement for their ratings of the spoken English of others than do the native-Chinese raters.

Past Research: Second-Language Learning

Much research in the field of second-language learning focuses on factors that affect the degree of foreign accent in one's second language (L2) as opposed to one's first (L1). Researchers have found that the following variables, among many others, are important to consider when assessing one's degree of foreign accent in a second language: one's age of arrival in the new locale; how often one uses L1 versus L2 at home, school, and work; length of residence in the new locale; the age at which one first learned the L2; and if the person had any formal instruction in his or her L2 (Piske, MacKay, & Flege, 2001). Like similar language studies

(e.g., Flege & Fletcher, 1991; Flege, Yeni-Komahian, & Liu, 1999), Piske, MackKay, and Flege (2001) showed that age of arrival in the new locale was found to be one of the most important variables for second-language learning and use, especially with regard to accent in the L2. Italian-English bilinguals who arrived in Canada early on in their lives were rated as having less of an accent in their spoken English as compared to late arriving Italian-English bilinguals.

Additionally, often confounded with one's age of arrival into a new locale is the age at which one first learned his or her L2. Researchers refer to this variable as the "critical period" for L2 language learning, and it has remained a popular area of study within the field of bilingualism (Flege, Munro, & MacKay, 1995; Flege, Birdsong, Bialystok, Mack, Sung, & Tsukada, 2006). Specifically, Flege et al. (1995) showed that a person's age of learning was a powerful indicator of how well he or she could pronounce English sentences, as evidenced by lower foreign accent ratings for individuals with low ages of L2 learning and higher foreign accent ratings for individuals with higher ages of L2 learning. Another study by Flege et al. (2006) revealed evidence supporting the notion that critical period learning is not the only reason why children tend to have less of an L2 accent than adults. Factors such as amount of L2 input, increased motivation for learning the L2, and more positive attitudes toward the culture represented by L2 native speakers were all linked with lower ages of L2 acquisition, and were thought to have comparable significance in regard to one's degree of foreign accent in his or her L2.

Past Research: Inter-Rater Reliability to Measure Performance

The research reviewed above concern L2 accent level and listener perception, but these studies have not focused on inter-rater reliability as a measure for expertise in the language under investigation.

There has been research regarding one's level of expertise in a musical skill and how this affects assessments of others engaging in that same task. The results of such studies have had mixed findings in regard to whether or not the amount of training a musician had influenced the reliability and accuracy of performance evaluation. For example, Bergee (2003) compared the performance ratings of musical teaching assistants and more experienced musical faculty to test the theory in question. Bergee had these two groups of participants evaluate the end of the semester juries at their university, consisting of graduate and undergraduate music majors and minors. Bergee found that there was no effect of musical expertise on the reliability of musical performance ratings.

In another study, the internal consistency of performance evaluations for piano playing was evaluated as a function of the rater's music expertise (Kinney, 2009). Kinney compared the musical performance evaluations of undergraduate music majors, nonmusic majors, and music faculty. Kinney's results suggested a definite effect of musical expertise and training on the internal consistency of musical performance evaluations. The undergraduate music majors had higher internal consistency among their performance ratings than did the nonmusic majors, and the faculty had even higher internal consistency than the music majors. It is also important to note that Kinney found that all groups of participants had lower internal consistency among their ratings when the musical excerpts they rated were unfamiliar to them, as compared to familiar excerpts. This idea is particularly relevant to the present study because I predict that the native-Chinese speakers will have lower amount of agreement among their ratings of L2 speech, due to a level of unfamiliarity with spoken English's accent.

Chapter 2

Method

Questions in Chapter 1 were raised concerning the extent to which native-English speakers and native-Chinese-English bilinguals differ in their abilities to perceive accents in the spoken English of other native and non-native English speakers. To address this question, the two research areas of second-language learning and inter-rater reliability as a measure of expertise were combined in an experiment employing, in part, past research methods of Flege et al. (2006).

Forty participants from the Pennsylvania State University completed the experiment. Each individual received ten dollars in compensation for his or her participation. The participants were broken up into two groups, each consisting of 10 native monolingual English speakers, and 10 native-Chinese-English bilinguals. In total, there were 20 participants in Group 1, and 20 in Group 2. Each participant completed a Language History Questionnaire to gain information about his or her second language history, if applicable. Each participant, regardless of his or her group, then completed the following three battery tasks, in this order: operation-span, lexical decision, and Simon. Lastly, participants in Group 1 recorded themselves while speaking four sentences in English, and participants in Group 2 listened to these sentences and provided accentedness ratings for each. Altogether, the study took participants from both groups about 45 minutes each to complete.

Battery 1: Operation-Span Task

Participants first completed an operation-span task to test for working memory capabilities. This task requires the participant to actively process information while also storing

other information in working memory to allow it to be available for retrieval (Linck, Kroll, & Hoshino, 2008). Linck et al. (2008) showed that bilinguals tend to have higher working memory abilities and inhibitory control, consistent with the idea that acquiring more than one language calls for the appropriate "tuning out" of one language when the other is in use. Thus, performance on the operation-span task predicts the efficiency of language processing without directly tapping into language-specific processing areas.

Participants were told that a series of correct and incorrect mathematical equations would be presented on the screen, each initiated by the presentation of a fixation point ("+") in the center of the screen. Participants were instructed to respond with the "yes" response key if the equation was true (e.g., 4 + 4= 8), or with the "no" response key if the equation was false (e.g., 9 – 4= 2). Immediately after the participant's response, a word in the participant's native language was presented (e.g., "rock" for native-English subjects; a Chinese symbol for native-Chinese subjects). Participants were told to try to the best of their ability to correctly respond to each math equation, as well as to try and remember each subsequent word. This task consisted of 15 blocks in all, and the number of math equations and words presented each increased by one every three blocks. In total, there were 60 math equations and 60 words presented. After each block, subjects had to recall the words presented from that respective block only. The native-English participants typed the words they recalled in a box that appeared on the computer screen, and native-Chinese participants wrote down the words they remembered on paper. The computer keyboards were not compatible with typing Chinese symbols. Participants were told that when recalling the words from a block, they could not type or write the last word that was presented in each respective block as the first word on their recall list. For example, if the last word presented in a block was "dog," participants could not type or write "dog" first on their recall list for that block. Participants completed two practice blocks. Following practice the experimenter left the room.

Battery 2: Lexical Decision Task

To demonstrate his or her proficiency in English, each participant next completed a lexical decision task. Participants were told that they would see a series of stimuli presented on the screen; the stimuli would consist of English words (e.g., "love") and English nonwords (e.g., "qult"). Each word was initiated by the presentation of a fixation point ("+") in the center of the screen. Participants were instructed to press the "yes" response key when they saw an English word, and the "no" response key for each nonword. In total, 50 English words and 50 English nonwords were presented in a random order. Following the practice trials, the experimenter left the room.

Battery 3: Simon Task

The final battery tasks that each participant completed was the Simon task. The Simon task was used as a measure of each participant's executive functioning capabilities. Participants were told that they would see a either a blue box or a red box on the screen, each initiated by the presentation of a fixation point ("+") in the center of the screen. Participants were instructed to press the "blue" response key when he or she saw a blue box appear on the screen, or the "red" response key when he or she saw a red box appear on the screen. Each trial would either be congruent with its appropriate response key, incongruent with its appropriate response key, or central in its position on the screen. (e.g., a "congruent" trial would consist of a red box appearing on the right side of the screen, since the red response key is located on the right side of the keyboard; an "incongruent" trial would consist of a blue box appearing on the right side of the screen, since its response key is on the left side of the keyboard; a "central" trial would consist of either the red box or the blue box appearing in the center of the screen). It has been well

established from research on stimulus-response compatibility that response times are longer on incongruent trials than congruent trials due to the mismatch between the stimulus location and the response location. This difference is known as the Simon effect (Linck et al., 2008). However, as mentioned earlier in regard to the operation-span task, bilinguals tend to have better executive functioning capabilities due to their experience with multiple language use. Thus, it is thought that bilinguals may have a stronger ability to inhibit an automatic response driven by stimulus location (Linck et al., 2008). There were 42 congruent, 42 incongruent, and 42 central trials. In total, there were 126 trials altogether. Following the practice trials, the experimenter left the room.

Speaker Group Recordings

After completing each of the three battery tasks, participants in Group 1 were asked to record themselves while speaking four sentences in English, as in earlier research by Flege et al. (2006). The experimenter played a pre-recorded mini-dialogue for the subject to listen to, consisting of a question being asked, followed by a response, whereupon the question was asked again. Subjects heard four sample mini-dialogues, and then the experimenter left the room. Subjects were instructed to press the record button on the recording device when they were ready to begin. They recorded themselves while repeating the response they heard for each mini-dialogue. There were four mini-dialogues, so four target sentences were recorded by each participant. An example of this procedure is as follows:

Voice 1: Where did the man go?

Voice 2: He went to work.

Voice 1: Where did the man go?

Subject repeats: [Voice 2]

Piske et al. (2001) referred to these mini-dialogues as the "delayed repetition technique," and they argued that this procedure is best for recording speech because "the delay between the model (e.g., Voice 2) and its repetition, as well as the intervening speech material, probably prevented [participant] direct imitations from sensory memory" (Piske et al., pg. 206, 2001). Appendix A provides the full list of sentences that participants heard and recorded. These sentences were taken from Flege et al. (2001).

Rater Group Accentedness Ratings

Participants in Group 2 listened to the recordings from Group 1 and provided accentedness ratings for each sentence they heard. The ratings occupied a scale of 1-10 (1= very strong accent to 10= no accent at all) (Flege et al., 2001). There were 80 sentences in all (20 speakers × 4 sentences each). The raters were given rating sheets numbered 1-80, each corresponding to a sentence number on the audio track. Each number on the sheet had the scale of 1-10 printed next to it. Participants were told that they would hear 80 sentences produced by 20 different speakers. They were told to circle the corresponding number on the accentedness scale for each sentence they heard. Participants had five seconds after each sentence recording to circle the appropriate rating. Participants were told that if they needed more time, they could stop the audio track, give their rating, and then resume the task. After explaining the rating task, the experimenter left the room. The rating task took about 9 minutes.

Chapter 3

Results

Overall, native-Chinese speakers were assigned lower ratings than were native-English speakers. Native-Chinese raters and native-English raters gave equally high ratings to native English speakers, but native-Chinese raters gave higher ratings to the spoken English of native-Chinese speakers than did native-English raters. Native-English raters and native-Chinese raters had the same low mean agreement among their ratings for native-English speakers (r= .20 and r=.17, respectively). This outcome may reflect a ceiling effect. Most importantly of all, native-Chinese raters had a lower mean agreement among their ratings for native-Chinese speakers (r= .73) than did native-English raters for the same native-Chinese speakers (r=.86). The latter result is consistent with the hypothesis that the native-English raters were better able to pick up phonetic departures from native-English speech than were the native-Chinese raters.

Language History Questionnaire Data

The proficiency measures from the language history questionnaire are summarized in Table 3-1 for the Chinese-English bilingual participants. Although participants considered themselves relatively proficient in their L2, they rated their proficiency in L2 substantially lower (8.0) than their L1 (9.2). These participants were more dominant in their L1, and reported using their L2 only in their English classes and comparable restricted environments requiring spoken English. During regular daily communication, these participants used their L1.

Table 3-1
Language experiences and self-assessed proficiency ratings for the Chinese-English bilingual participants (n=20)

Number of Years Speaking L1 / L2			
English (L2) 11.2		Chinese (L1) 20.7	
Skill	Self-assessed ratings[a]		
	English (L2)	Chinese (L1)	
Reading	8.3	9.0	
Spelling	7.9	8.8	
Writing	7.7	8.7	
Speaking	7.8	9.7	
Speech comprehension	8.2	9.9	
Mean rating	8.0	9.2	

[a] Based on a scale of 1-10.

Battery 1: Operation-Span Task

The Operation-Span task was calculated as the number of English or Chinese words that were recalled correctly for trials in which participants made a correct judgment on the mathematical equation. The highest possible score was 60. In the study, the native-Chinese speakers had an average Operation-Span recall number of about 40, and the native-Chinese raters had an average recall number of about 43. The native-English speakers had an average recall number of about 50, whereas the native-English raters had an average recall of about 44. Figure 3-1 summarizes these results, with the line in each bar representing plus or minus one standard error. These results show that both the native-Chinese and native-English raters scored similarly, even though overall, the English participants from both groups scored higher on average. Such a result reveals that in this study, the Chinese-English bilingual participants showed no clear

working memory advantage over the monolingual English participants. See Appendix B for the programming script that was used to generate the results.

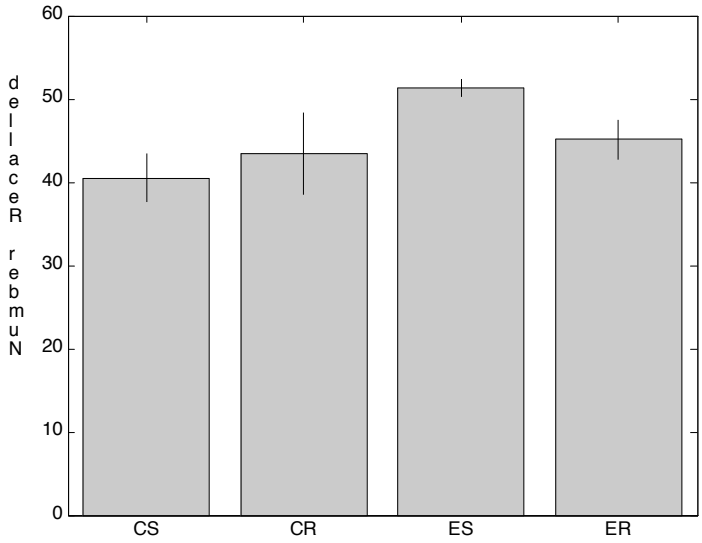

Figure 3-1 Operation-Span Task Results for all Participants

Battery 2: Lexical Decision Task

The native-Chinese participants of both groups scored lower in their English proficiency than did native-English participants of both groups. Native-Chinese participants from both groups had, on average, longer reaction times and a higher number of errors than did native-English participants from both groups. These normalized reaction times and normalized error rates are shown in Figure 3-1. The red and blue circles symbolize the native-Chinese and native-English speakers, respectively; whereas the native-Chinese and native-English raters are symbolized by the red and blue squares, respectively. The fact that the native-English participants

scored higher in their English proficiency than the native-Chinese participants is not surprising. This result was expected, but it helps assure me that my intended hypothesis was being tested accurately, since the English participants did in fact prove to be more experienced and proficient that the Chinese participants. These varying levels of expertise in the English language, especially among the raters, were a vital variable in this experiment, since we want to address how differing levels of expertise, in particular, will affect accentedness ratings. See Appendix C for the programming script that was used to generate the results.

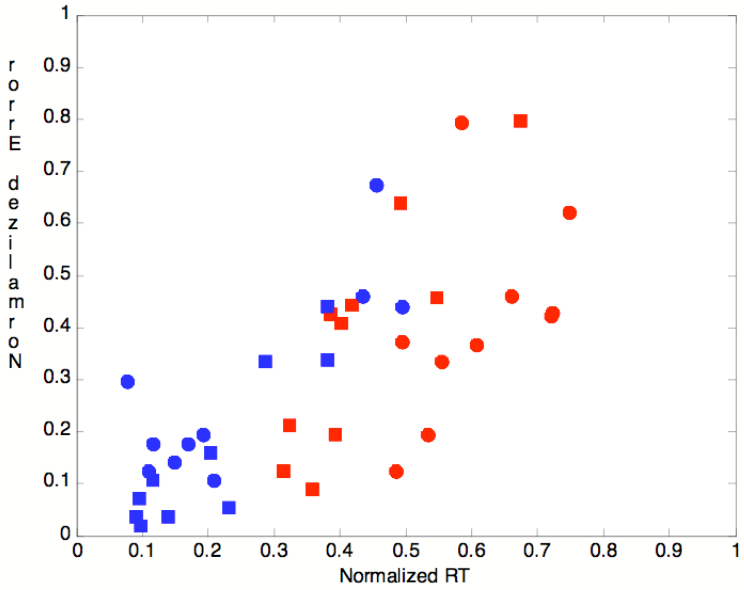

Figure 3-2 Lexical Decision Results for all Participants

Battery 3: Simon Task

To test the idea that bilinguals have higher inhibitory control, the magnitude of the Simon effect was computed for each group of participants. The results revealed that there was a definite Simon effect for the Chinese speakers, but not for the Chinese raters. In fact, the Chinese raters

scored lower on the Simon task than all of the other participant groups. The Chinese speakers had an average Simon score of 48, whereas the Chinese raters had an average Simon score of only 28. As for the English participants, the English speakers had an average Simon score of 35, and the English raters had an average score of 38. As one can see, the English participants did not have nearly as wide of a range between Simon scores than did the Chinese speaker and Chinese rater groups. Figure 3-3 summarizes these results, with the line in each bar representing plus or minus one standard error. See Appendix D for the programming script that was used to generate the results.

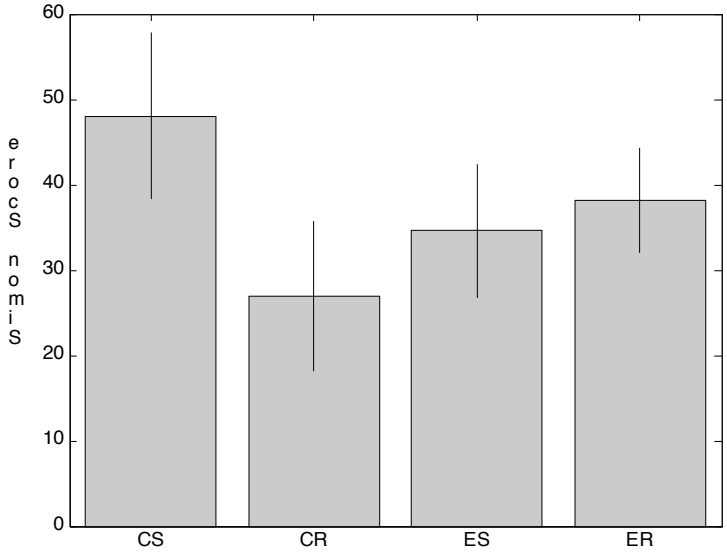

Figure 3-3 Simon Task Results for all Participants

Mean Accentedness Ratings & Degree of Agreement among Raters

The mean accentedness ratings for both the native-Chinese and native-English speaker's spoken English were found by averaging together the ratings given to each of these groups by the native-Chinese and native-English raters. On average, the Chinese speakers were given higher

mean accentedness ratings by native-Chinese raters (mean rating = about a 6) than by the native-English raters (mean rating = about a 4). This result indicates that the native-English raters were able to detect more phonetic errors in the spoken English of the Chinese speakers than the native-Chinese raters were. This result is most likely attributable to the lower English proficiency of the Chinese raters. However, the Chinese and English raters gave equally high accentedness ratings to the spoken English of the native-English speakers (both rater groups had mean ratings = about a 9). This result indicates that both groups of raters were equally likely to detect who was highly skilled at speaking English, whereas the English raters proved to be better than the Chinese raters at detecting who was not as skilled at speaking English.

To measure the degree of agreement among raters for each speaker group, the mean lexical decision task results for each rater-speaker group were crossed with one another, and likewise, the accentedness ratings for each rater-speaker group were crossed with one another to generate the degree of agreement correlations based on language proficiency and respective ratings. The results revealed that both the Chinese raters and native-English raters had low mean agreement among their accentedness ratings for the spoken English of the native-Chinese and native-English speakers ($r= 0.17$ and $r= .2$, respectively). This result may reflect a ceiling effect, that is, both groups found the quality of the native-English speech to be fine. The English speakers were, overall, given very high ratings, as judged from the 9 mean average ratings from both groups of raters. This creates an artificial clustering of ratings at the high end of the scale. Thus, the scale, in itself, may be limiting the true degree of accentedness among the raters from coming about. Most importantly, however, are the degrees of agreement among the native-Chinese and native-English rater's accentedness ratings for the spoken English of the native-Chinese speakers. Here I found that the native-English raters had a higher degree of agreement among their ratings than did the native-Chinese raters for the spoken English of native-Chinese speakers ($r= .86$ and $r= .73$, respectively). This indicates that although the Chinese raters were

able to detect some phonetic errors in the spoken English of the native-Chinese speakers, they were not as accurate or reliable in doing so as the native-English raters.

The mean accentedness ratings ±1 SE given by the raters (first letter in each x-axis label) to the speakers (second letter in each x-axis label) are shown in Figure 3-4. Also, the mean correlation among the ratings within each rater-speaker group is presented as text within each bar. See Appendix E, F, G, and H for more information on how these results were generated.

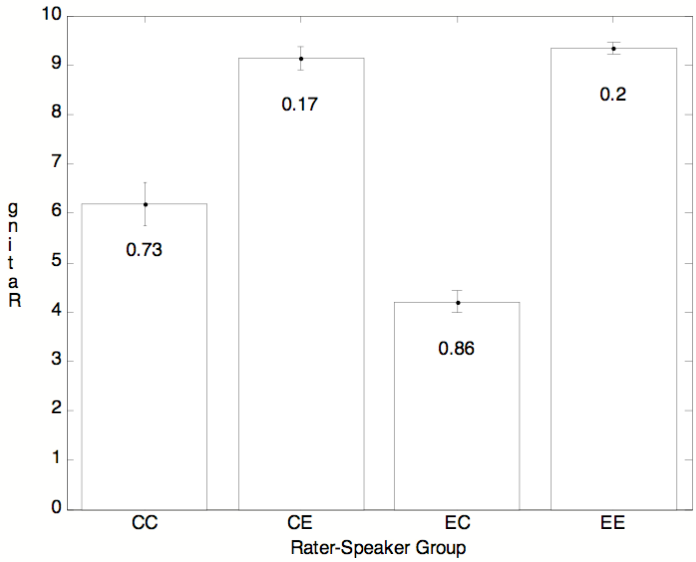

Figure 3-4 Mean Accentedness Ratings & Degree of Agreement among Raters

Chapter 4

Discussion

Summary of Findings

The Operation-Span results did not show that the Chinese-English bilingual participants had an advantage over the monolingual English participants in regard to working memory capabilities. On the other hand, the Simon task did show that the Chinese speakers, but not the Chinese raters, had a significant advantage over the monolingual English participants in regard to inhibitory control and executive functioning abilities. This outcome suggests that in some cases, but not all, learning two languages promotes the effectiveness of one's cognitive functioning. The lexical decision results showed that the native-Chinese participants had lower proficiency levels in English than did the native-English participants, as evidenced by the native-Chinese subject's longer reaction times and higher error numbers in the task. Moreover, the native-Chinese speakers were assigned lower accentedness ratings than were native-English speakers. Native-Chinese raters and native-English raters gave equally high ratings to native English speakers, but native-Chinese raters gave higher ratings to the spoken English of native-Chinese speakers than did native-English raters. Native-English raters and native-Chinese raters had the same low mean agreement among their ratings for native-English speakers (r= .20 and r=.17, respectively). This outcome may reflect a ceiling effect. Most important of all, native-Chinese raters had a lower mean agreement among their ratings for native-Chinese speakers (r= .73) than did native-English raters for the same native-Chinese speakers (r=.86). This result is consistent with the hypothesis that the native-English raters were better able to pick up phonetic departures from native-English speech than were the native-Chinese raters.

Conclusions

The present study showed that there is an effect of language proficiency on the agreement among users of that language on the accentedness ratings they give to other speakers speaking that language. Even though native-Chinese participants were able to reliably pick up some phonetic departures in native-Chinese speaker's spoken English ($r = .73$), their reliability, as measured by their inter-rater correlations, were not as high as that of native-English raters ($r = .86$). This result can be attributed to the fact that the native-English raters are more in tune with spoken English's ultimate goals relating to accent, proper stress levels within speech, phonetics, etc. For native-Chinese-English bilinguals, however, such details in their L2 may still be unfamiliar, thus resulting in the lower inter-rater reliability among their accentedness ratings.

These findings introduce a new method for probing the abilities of speakers and listeners of different languages. It also suggests a new method for training. An L2 learner can listen to speech within that L2, provide ratings for the speech, and then compare his or her ratings to those of experts in that language. If the L2 learner's ratings are far off from the experts' correlations, then he or she has gotten feedback that s/he is missing something that the experts perceive. On the other hand, if his or her ratings correlate well with those of the expert ratings, then that L2 learner can infer that s/he is picking up on the features that the experts are perceiving and judging comparably. Finally, this study also suggests, more broadly, that inter-rater reliability is a useful metric for exposing differences in expertise.

References

Bergee, M.J. (2003). Faculty interjudge reliability of music performance evaluation. *Journal of Research in Music Education, 51,* 137-150.

Doerksen, P.F. (1999). Aural-diagnostic and prescriptive skills of preservice and expert instrumental music teachers. *Journal of Research in Music Education, 47,* 78-88.

Flege, J.E., Birdsong, D., Bialystok, E., Mack, M., Sung, H., & Tsukada, K. (2006). Degree of foreign accent in English sentences produced by Korean children and adults. *Journal of Phonetics, 34,* 153-175.

Flege, J.E. & Fletcher, K.L. (1991). Talker and listener effects on degree of perceived foreign accent. *Journal of the Acoustical Society of America, 91,* 370-389.

Flege, J.E. & Frieda, E.M. (1997). Amount of native-language (L1) use affects the pronunciation of an L2. *Journal of Phonetics, 25,* 169-286.

Flege, J.E., Munro, M.J., & MacKay, I.R.A. (1995). Factors affecting the strength of perceived foreign accent in a second language. *Journal of the Acoustical Society of America, 97,* 3125-3134.

Flege, J.E., Yeni-Komshian, G.H., & Liu, S. (1999). Age constraints on second-language acquisition. *Journal of Memory and Language, 41,* 78-104.

Kinney, D.W. (2009). Internal consistency of performance evaluations as a function of music expertise and excerpt familiarity. *Journal of Research in Music Education, 56,* 322-337.

Linck, J.A., Hoshino, N., & Kroll, J.F. (2008). Cross-language lexical processes and inhibitory control. *The Mental Lexicon, 3,* 349-374.

Piske, T., MacKay, I.R.A., & Flege, J.E. (2001). Factors affecting degree of foreign accent in an L2: a review. *Journal of Phonetics, 29,* 191-215.

Schwartz, A.I. & Kroll, J.F. (2006). Bilingual lexical activation in sentence context. *Journal of Memory and Language, 55,* 197-212.

Appendix A. Questions and answers used to elicit target sentence recordings.*

Questions	Answers
1. Where did the man go?	He went to work.
2. What did he drink?	He drank a glass of water.
3. What did the girl eat?	She ate a sandwich.
4. What did you read?	I read a good book.
5. How are you today?	*I'm fine, thank you.*
6. What time is it?	*It's now ten o'clock.*
7. How much does it cost?	*It costs five dollars.*
8. Where did the children go?	*They went to school.*

*The questions and answers used to elicit the production of four English sentences. The four italicized sentences were produced and rated for foreign accent. The other sentences were listened to for practice.

Appendix B. Program Script used for Operation-Span Task Results

```
% rachel_O_spans_03
% 03-18-10
% Participant Recalled Intrusions Errors RT_Total
% Chinese Speakers
CS= [
   1 42 11 4 2012.5
   2 48 6 4 2010.6
   3 42 11 5 2027.1
   4 19 17 21 2058.2
   5 35 12 7 2232.1
   7 40 13 4 2271.9
   8 40 6 8 2151.9
   11 50 1 7 1750.9
   12 38 10 12 2378.3
   13 51 3 2 1758.5
   ];
%English Speakers
ES=[
14 49 1 5 2204.6
15 49 3 5 2413.1
16 48 1 10 1861
17 55 1 3 2026.4
18 56 0 2 1912.6
19 55 2 2 2187.7
20 48 2 5 1792.6
21 50 0 5 2373.9
22 50 0 2 2014.4
23 53 1 4 1811.1
   ];
% Chinese Raters
CR=[
   3.00 3 26 21 2620.2
   5.00 48 1 12 2169.7
   6.00 51 2 6 1653.5
   7.00 47 6 2 1876.1
   8.00 35 13 7 1880.8
   10.00 53 1 6 2324.1
   11.00 42 8 6 1851.6
   14.00 56 2 1 1507.3
   15.00 50 0 8 1600.4
   18.00 49 5 5 1687.1
   ];
```

```
% English Raters
ER=[
1 39 1 12 2361.5
2 28 2 21 2533.3
4 51 1 6 2325.2
9 44 1 8 1895.1
12 45 6 5 2111
13 53 3 3 1888.1
16 44 3 11 2490.2
17 50 2 3 3023.5
19 49 4 3 1651.2
20 48 0 7 1623.7
];

mean_recalled=[];
se_recalled=[];
for g=1:4
   if g==1
      d=CS;
   elseif g==2
      d=CR;
   elseif g==3
      d=ES;
   else
      d=ER;
   end
   mean_recalled=[mean_recalled mean(d(:,2))];
   se_recalled=[se_recalled std(d(:,2))/sqrt(length(d))];
end
close all
figure(11)
hold on
box on
bar([1:4],mean_recalled)
colormap([.75 .75 .75])
ylabel('Number Recalled')
set(gca,'xtick',[1:4])
set(gca,'xticklabel',['CS' ;'CR'; 'ES' ;'ER']);

for i=1:4
   plot([i i],[mean_recalled(i)-se_recalled(i) mean_recalled(i)+se_recalled(i)],'k-')
end
```

Appendix C. Program Script used for Lexical-Decision Task Results

```
% Program to calculate mean RTs and accuracy for word targets and nonword targets
% in a lexical decision task conducted with E-Prime and whose data are contained in an
Excel
% spreadsheet.

% The program makes three passes through the data.
% In the first pass, mean RTs and accuracy are computed for the raw data.
% In the second pass, mean RTs, standard deviation of RTs, and accuracy are
% computed for RTs between a finite lo-threshold and a finite hi_threshold.
% In the third pass, mean RTs and accuracy are computed for RTs
% with acceptable z scores relative to the mean RTs from the second pass (and
% using the standard deviations compupted in the second pass).
% The outputs for each pass show, for each subject (including for each
% subject with no data) the mean RT for words correct, the number of words correct,
% the proportion words correct, the mean RT for nonwords correct, the number of
nonwords correct,
% the proportion nonwords correct.

clc
clear all
close all
commandwindow
format bank

lo_threshold(1)=-inf;
hi_threshold(1)=inf;
lo_threshold(2)=300;
hi_threshold(2)=3000;
lo_threshold(3)=lo_threshold(2);
hi_threshold(3)=hi_threshold(2);
z=2.5; % acceptable z score for pass 3
ns=13; % number of subjects
[a b c]=xlsread('Chinese_LDT_Group1_Condensed.xlsx');

for pass=1:3
   WR=[]; % Word results
   NWR=[]; % Nonword results
   for s=1:ns
      words=0;
      words_correct=0;
      RTs_for_words_correct=[];
      nonwords=0;
      nonwords_correct=0;
      RTs_for_nonwords_correct=[];
      for r=1:length(c)
```

```
                if c{r,2}==s
                    if strcmp(c{r,3},'WORD')
                        words=words+1;
                        if strcmp(c{r,5},'c') && c{r,6}>lo_threshold(pass) &&
c{r,6}<hi_threshold(pass)
                            if pass<3
                                words_correct=words_correct+1;
                                RTs_for_words_correct=[RTs_for_words_correct c{r,6}];
                            else
                                if c{r,6}>mWR-z*sWR && c{r,6}<mWR+z*sWR
                                    words_correct=words_correct+1;
                                    RTs_for_words_correct=[RTs_for_words_correct c{r,6}];
                                end
                            end
                        elseif strcmp(c{r,3},'NONWORD')
                            nonwords=nonwords+1;
                            if strcmp(c{r,5},'m') && c{r,6}>lo_threshold(pass) &&
c{r,6}<hi_threshold(pass)
                                if pass<3
                                    nonwords_correct=nonwords_correct+1;
                                    RTs_for_nonwords_correct=[RTs_for_nonwords_correct c{r,6}];
                                else
                                    if c{r,6}>mNWR-z*sNWR && c{r,6}<mNWR+z*sNWR
                                        nonwords_correct=nonwords_correct+1;
                                        RTs_for_nonwords_correct=[RTs_for_nonwords_correct c{r,6}];
                                    end
                                end
                            end
                        end
                    end
                    WR=[WR;s mean(RTs_for_words_correct) words_correct words_correct/words];
                    NWR=[NWR;s mean(RTs_for_nonwords_correct) nonwords_correct
nonwords_correct/nonwords];
                end
                if pass==1
                    disp('Raw Data')
                elseif pass==2
                    disp('Data Between Thresholds')
                    mWR=nanmean(WR(:,2));
                    sWR=nanstd(WR(:,2));
                    mNWR=nanmean(NWR(:,2));
                    sNWR=nanstd(NWR(:,2));
                elseif pass==3
                    disp('Data Between Thresholds and Within Acceptable z Scores')
                end
                R=[WR NWR(:,2:end)]
            end
```

Appendix D. Program Script used for Simon Task Results

Chinese Speakers =

Columns 1 through 8

1.0000	42.0000	37.0000	38.0000	496.0263	431.3333	519.3684	88.0351
2.0000	42.0000	40.0000	39.0000	364.9189	352.2105	384.4737	32.2632
3.0000	41.0000	42.0000	36.0000	406.6750	371.8205	428.1471	56.3265
4.0000	37.0000	42.0000	39.0000	428.2703	405.1795	487.2564	82.0769
5.0000	41.0000	41.0000	38.0000	460.4872	415.2750	469.7027	54.4277
7.0000	42.0000	40.0000	41.0000	373.3902	386.2308	392.7750	6.5442
8.0000	40.0000	39.0000	39.0000	334.5263	336.5135	355.2703	18.7568
11.0000	40.0000	40.0000	39.0000	429.5385	411.2500	458.7027	47.4527
12.0000	41.0000	40.0000	41.0000	392.3902	386.1842	390.4146	4.2304
13.0000	42.0000	42.0000	41.0000	439.9268	404.7073	495.2000	90.4927

Columns 9 through 11

0.0317	0.0397	0.1111
0	0.0397	0.1032
0.0317	0.0238	0.1032
0.0556	0.0079	0.0873
0.0397	0.0079	0.0794
0.0159	0.0079	0.0476
0.0317	0.0317	0.1111
0.0397	0.0159	0.0794
0.0317	0	0.0476
0.0079	0	0.0317

English Speakers =

Columns 1 through 8

14.0000	41.0000	38.0000	41.0000	476.3250	482.7778	496.9750	14.1972
15.0000	40.0000	42.0000	37.0000	336.8947	323.7561	367.3889	43.6328
16.0000	40.0000	39.0000	40.0000	426.0513	433.7105	458.2051	24.4946
17.0000	39.0000	39.0000	39.0000	336.0000	358.3947	353.1892	-5.2055
18.0000	42.0000	41.0000	41.0000	396.2250	427.4750	436.9268	9.4518
19.0000	39.0000	40.0000	36.0000	430.5263	394.7949	471.8000	77.0051
20.0000	41.0000	38.0000	40.0000	347.1351	346.7632	368.8649	22.1017
21.0000	42.0000	39.0000	41.0000	459.0250	445.4211	485.6585	40.2375
22.0000	40.0000	40.0000	40.0000	406.6316	382.1579	438.1351	55.9772
23.0000	42.0000	42.0000	38.0000	363.5000	342.4750	406.3889	63.9139

Columns 9 through 11

0.0397	0.0079	0.0794
0.0397	0.0159	0.0873
0.0397	0.0159	0.0794
0.0238	0.0476	0.1270
0.0159	0	0.0397
0.0556	0.0317	0.1111
0.0238	0.0317	0.1111
0.0238	0.0079	0.0556
0.0159	0.0317	0.1032
0.0238	0.0079	0.0635

Chinese Raters =

Columns 1 through 8

3.0000	38.0000	42.0000	36.0000	409.2500	382.3684	430.3824	48.0139
5.0000	41.0000	41.0000	40.0000	365.7000	352.9737	389.1026	36.1289
6.0000	41.0000	39.0000	41.0000	346.2051	341.2895	369.7949	28.5054
7.0000	41.0000	38.0000	40.0000	366.2051	367.3056	407.7949	40.4893
8.0000	38.0000	42.0000	36.0000	407.2432	398.5500	442.7143	44.1643
10.0000	41.0000	41.0000	40.0000	463.9744	492.3421	482.7500	-9.5921
11.0000	41.0000	42.0000	37.0000	431.8205	423.7692	498.8889	75.1197
14.0000	40.0000	42.0000	42.0000	451.0513	450.3171	462.1707	11.8537
15.0000	42.0000	39.0000	41.0000	322.3902	334.4737	353.9268	19.4531
18.0000	42.0000	41.0000	39.0000	549.4390	605.0789	580.5385	-24.5405

Columns 9 through 11

0.0317	0.0476	0.1429
0.0079	0.0238	0.0714
0.0238	0.0159	0.0794
0.0317	0.0238	0.0952
0.0556	0.0238	0.1111
0.0159	0.0159	0.0714
0.0159	0.0317	0.0952
0.0159	0	0.0397
0.0317	0	0.0476
0.0238	0.0079	0.0635

English Raters =

Columns 1 through 8

1.0000	42.0000	41.0000	41.0000	582.8500	546.8718	575.7317	28.8599
2.0000	39.0000	39.0000	42.0000	486.1538	502.5946	543.5250	40.9304
4.0000	42.0000	40.0000	40.0000	391.5714	351.1892	426.2250	75.0358
9.0000	42.0000	40.0000	40.0000	360.5714	352.3846	371.0526	18.6680

12.0000	39.0000	39.0000	37.0000	355.9444	323.1892	388.7222	65.5330
13.0000	39.0000	40.0000	38.0000	370.1316	336.2821	364.4706	28.1885
16.0000	39.0000	38.0000	40.0000	394.6579	395.9730	439.3750	43.4020
17.0000	41.0000	40.0000	40.0000	353.7297	361.2821	376.1500	14.8679
19.0000	40.0000	41.0000	38.0000	339.4211	328.5854	375.3143	46.7289
20.0000	42.0000	41.0000	38.0000	350.5250	343.7632	363.2222	19.4591

Columns 9 through 11

0.0079	0.0079	0.0476
0.0317	0.0159	0.0794
0.0238	0.0079	0.0556
0.0238	0.0079	0.0556
0.0476	0.0397	0.1349
0.0317	0.0397	0.1190
0.0556	0.0159	0.0873
0.0238	0.0159	0.0794
0.0317	0.0238	0.0952
0.0079	0.0317	0.0952

Appendix E. Program Script used for Rating Results

```
start
% read in data
[a b c]=xlsread('Inter_rater_data_04.xls');
% Inter_rater_data_04.xls

[c_rows c_cols]=size(c);
% Note that chinese speaker 1 does not equal Chinese rater, etc.
chinese_rater=[3 5 6 7 8 10 11 14 15 18];
english_rater=[1 2 4 9 12 13 16 17 19 20];
chinese_speaker=[9 11 1 14 15 8 20 2 18 7];
english_speaker=[6 19 12 3 10 13 5 4 17 16];
RSm_and_se =[];
for pass=1:4
   if pass == 1
      rater=chinese_rater;
      speaker=chinese_speaker;
   elseif pass ==2
      rater=chinese_rater;
      speaker=english_speaker;
   elseif pass == 3
      rater=english_rater;
      speaker=chinese_speaker;
   elseif pass ==4
      rater=english_rater;
      speaker=english_speaker;
   end
   RS=[];
   for ra=1:length(rater)
      rs_ratings=[];
      for sp=1:length(speaker)
         rsrs=0;
         n_rsrs=0;
         for row=3:c_rows
            for col=5:c_cols
               if c{1,col}==rater(ra) && c{row,3}==speaker(sp)
                  rs_ratings=[rs_ratings c{row,col}];
                  rsrs=rsrs + c{row,col};
                  n_rsrs=n_rsrs+1;
               end
            end
         end
         sentences_collapsed(sp,ra)=rsrs/n_rsrs;
      end
      RS=[RS nanmean(rs_ratings)];
   end
```

```
    RSm_and_se =[RSm_and_se;nanmean(RS) nanstd(RS)/sqrt(length(rater))];
    sentences_collapsed
    nanmean(nanmean(sentences_collapsed))
    correlations_for_sentences_collapsed;
    if pass==1
       V_CS_CR=correl_sentences_collapsed
       xlswrite('V_CS_CR.xls',V_CS_CR)
    elseif pass==2

       V_ES_CR=correl_sentences_collapsed
       xlswrite('V_ES_CR.xls',V_ES_CR)
    elseif pass==3
       V_CS_ER=correl_sentences_collapsed
       xlswrite('V_CS_ER.xls',V_CS_ER)
    elseif pass==4
       V_ES_ER=correl_sentences_collapsed
       xlswrite('V_ES_ER.xls',V_ES_ER)
    end
    if pass<4
       clear sentences_collapsed
    end
end % for pass=1:4
RSm_and_se
figure(1)
yoffset=1;
hold on
box on
bar(RSm_and_se(:,1))
colormap([1 1 1])
set(gca,'xtick',[1:4])
set(gca,'xticklabel',['CC';'CE';'EC';'EE'])
for p=1:4
    errorbar(p,RSm_and_se(p,1),RSm_and_se(p,2),'k')
    text(p,RSm_and_se(p,1)-
yoffset,num2str(correl_text(p),2),'horizontalalignment','center');
end
ylabel('Rating')
xlabel('Rater-Speaker Group')
correl_text
```

Appendix F. Program Script used to Generate Matrix Comparisons

```
% rachel_summary_matrix_comparisons_09
% 03-10-10
% read in relevant matrices from Excel files to correlate the
clc
clear all

commandwindow
V_CS_CR=xlsread('V_CS_CR.xls');
V_ES_CR=xlsread('V_ES_CR.xls');
V_CS_ER=xlsread('V_CS_ER.xls');
V_ES_ER=xlsread('V_ES_ER.xls');
MEL_CS=xlsread('MEL_CS.xls');
MEL_ES=xlsread('MEL_ES.xls');
MEL_CR=xlsread('MEL_CR.xls');
MEL_ER=xlsread('MEL_ER.xls');
disp(' 1  corrcoef(V_CS_CR,MEL_CS)')
disp(' 2  corrcoef(V_CS_CR,MEL_ES)')
disp(' 3  corrcoef(V_CS_CR,MEL_CR)')
disp(' 4  corrcoef(V_CS_CR,MEL_ER)')
disp(' 5  corrcoef(V_CS_ER,MEL_CS)')
disp(' 6  corrcoef(V_CS_ER,MEL_ES)')
disp(' 7  corrcoef(V_CS_ER,MEL_CR)')
disp(' 8  corrcoef(V_CS_ER,MEL_ER)')
disp(' 9  corrcoef(V_ES_CR,MEL_CS)')
disp('10  corrcoef(V_ES_CR,MEL_ES)')
disp('11  corrcoef(V_ES_CR,MEL_CR)')
disp('12  corrcoef(V_ES_CR,MEL_ER)')
disp('13  corrcoef(V_ES_ER,MEL_CS)')
disp('14  corrcoef(V_ES_ER,MEL_ES)')
disp('15  corrcoef(V_ES_ER,MEL_CR)')
disp('16  corrcoef(V_ES_ER,MEL_ER)')
j=0;
CC=[];
c=corrcoef(V_CS_CR,MEL_CS);
j=j+1;
CC=[CC; j c(1,2)];
c=corrcoef(V_CS_CR,MEL_ES);
j=j+1;
CC=[CC; j c(1,2)];
c=corrcoef(V_CS_CR,MEL_CR);
j=j+1;
CC=[CC; j c(1,2)];
c=corrcoef(V_CS_CR,MEL_ER);
j=j+1;
CC=[CC; j c(1,2)];
```

```
c=corrcoef(V_CS_ER,MEL_CS);
j=j+1;
CC=[CC; j c(1,2)];
c=corrcoef(V_CS_ER,MEL_ES);
j=j+1;
CC=[CC; j c(1,2)];
c=corrcoef(V_CS_ER,MEL_CR);
j=j+1;
CC=[CC; j c(1,2)];

c=corrcoef(V_CS_ER,MEL_ER);
j=j+1;
CC=[CC; j c(1,2)];
c=corrcoef(V_ES_CR,MEL_CS);
j=j+1;
CC=[CC; j c(1,2)];
c=corrcoef(V_ES_CR,MEL_ES);
j=j+1;
CC=[CC; j c(1,2)];
c=corrcoef(V_ES_CR,MEL_CR);
j=j+1;
CC=[CC; j c(1,2)];
c=corrcoef(V_ES_CR,MEL_ER);
j=j+1;
CC=[CC; j c(1,2)];
c=corrcoef(V_ES_ER,MEL_CS);
j=j+1;
CC=[CC; j c(1,2)];
c=corrcoef(V_ES_ER,MEL_ES);
j=j+1;
CC=[CC; j c(1,2)];
c=corrcoef(V_ES_ER,MEL_CR);
j=j+1;
CC=[CC; j c(1,2)];
c=corrcoef(V_ES_ER,MEL_ER);
j=j+1;
CC=[CC; j c(1,2)];
CC
```

Appendix G. Sets of Speaker-Rater Correlations

1.) Correlations between the ratings given to all Chinese speakers (CS) by every pair of English raters (ER)

	A	B	C	D	E	F	G	H	I	J
1	1	0.85834	0.73621	0.88577	0.82570	0.97304	0.80159	0.81546	0.91430	0.86493
2	0.85834	1	0.88933	0.86316	0.90567	0.83802	0.85221	0.80806	0.89746	0.96283
3	0.73621	0.88933	1	0.83020	0.84679	0.71767	0.86315	0.84625	0.77332	0.87980
4	0.88577	0.86316	0.83020	1	0.92763	0.80820	0.96089	0.88476	0.86927	0.89015
5	0.82570	0.90567	0.84679	0.92763	1	0.79123	0.94202	0.92354	0.90596	0.93327
6	0.97304	0.83802	0.71767	0.80820	0.79123	1	0.72848	0.82086	0.93410	0.84461
7	0.80159	0.85221	0.86315	0.96089	0.94202	0.72848	1	0.91423	0.79505	0.85926
8	0.81546	0.80806	0.84625	0.88476	0.92354	0.82086	0.91423	1	0.87299	0.87042
9	0.91430	0.89746	0.77332	0.86927	0.90596	0.93410	0.79505	0.87299	1	0.94081
10	0.86493	0.96283	0.87980	0.89015	0.93327	0.84461	0.85926	0.87042	0.94081	1

2.) Correlations between the ratings given to all Chinese speakers (CS) by every pair of Chinese raters (CR)

	A	B	C	D	E	F	G	H	I	J
1	1	0.76026	0.90407	0.88698	0.56498	0.54652	0.74580	0.80736	0.84676	0.59858
2	0.76026	1	0.79939	0.70320	0.63492	0.63457	0.86907	0.87495	0.87402	0.80371
3	0.90407	0.79939	1	0.74878	0.75719	0.49168	0.80781	0.89966	0.79517	0.62589
4	0.88698	0.70320	0.74878	1	0.44197	0.74537	0.74471	0.75460	0.91046	0.70923
5	0.56498	0.63492	0.75719	0.44197	1	0.23246	0.59196	0.80751	0.64771	0.44689
6	0.54652	0.63457	0.49168	0.74537	0.23246	1	0.69763	0.58651	0.76576	0.80660
7	0.74580	0.86907	0.80781	0.74471	0.59196	0.69763	1	0.76713	0.80692	0.80385
8	0.80736	0.87495	0.89966	0.75460	0.80751	0.58651	0.76713	1	0.88580	0.76513
9	0.84676	0.87402	0.79517	0.91046	0.64771	0.76576	0.80692	0.88580	1	0.84577
10	0.59858	0.80371	0.62589	0.70923	0.44689	0.80660	0.80385	0.76513	0.84577	1

3.) Correlations between the ratings given to all English speakers (ES) by every pair of English raters (ER)

	A	B	C	D	E	F	G	H	I	J
1	1	-0.10276	-0.07147	0.64757	0.50724	0.88440	0.16071	0.36626	0	0.19069
2	-0.10276	1	0.44537	-0.44113	0.27126	-0.16904	0.22013	0.25682	0.27394	-0.04227
3	-0.07147	0.44537	1	-0.05047	0.32243	-0.20984	0.54510	0.86761	0.14338	0.28557
4	0.64757	-0.44113	-0.05047	1	0.13414	0.59053	-0.00934	0.34367	-0.37116	0.62374
5	0.50724	0.27126	0.32243	0.13414	1	0.52315	0.83182	0.39923	-0.28588	0.24525
6	0.88440	-0.16904	-0.20984	0.59053	0.52315	1	0.05550	0.22717	-0.17839	0.09369
7	0.16071	0.22013	0.54510	-0.00934	0.83182	0.05550	1	0.42379	-0.23593	0.28092
8	0.36626	0.25682	0.86761	0.34367	0.39923	0.22717	0.42379	1	0.04251	0.34922
9	0	0.27394	0.14338	-0.37116	-0.28588	-0.17839	-0.23593	0.04251	1	-0.20286
10	0.19069	-0.04227	0.28557	0.62374	0.24525	0.09369	0.28092	0.34922	-0.20286	1

4.) Correlations between the ratings given to all English speakers (ES) by every pair of Chinese raters (CR)

	A	B	C	D	E	F	G	H	I	J
1	1	-0.20913	0.04251	0.44115	-0.10302	0.49569	-0.33553	0.59530	0.34107	-0.21539
2	-0.20913	1	0.50457	-0.18334	0.15534	0.44213	0.82292	0.12448	0.36212	0.74158
3	0.04251	0.50457	1	0.10927	0.25721	-0.15431	0.36820	0.24547	0.47483	-0.10485
4	0.44115	-0.18334	0.10927	1	-0.12977	0.01835	-0.33275	0.83638	0.28895	-0.25850
5	-0.10302	0.15534	0.25721	-0.12977	1	-0.06539	-0.20468	-0.14886	-0.12863	-0.29314
6	0.49569	0.44213	-0.15431	0.01835	-0.06539	1	0.26794	0.40191	0.39160	0.54687
7	-0.33553	0.82292	0.36820	-0.33275	-0.20468	0.26794	1	0	0.08426	0.80403
8	0.59530	0.12448	0.24547	0.83638	-0.14886	0.40191	0	1	0.51958	0
9	0.34107	0.36212	0.47483	0.28895	-0.12863	0.39160	0.08426	0.51958	1	0.04657
10	-0.21539	0.74158	-0.10485	-0.25850	-0.29314	0.54687	0.80403	0	0.04657	1

Appendix H. Mean Lexical Decision Efficiencies

1.) Mean LD Efficiencies for every pair of English Speakers (ES)

	A	B	C	D	E	F	G	H	I	J
1	0.58388	0.42134	0.41129	0.34117	0.34178	0.51321	0.36379	0.54663	0.37055	0.35102
2	0.42134	0.25881	0.24876	0.17863	0.17925	0.35068	0.20125	0.38409	0.20802	0.18848
3	0.41129	0.24876	0.23871	0.16858	0.16920	0.34063	0.19120	0.37404	0.19797	0.17843
4	0.34117	0.17863	0.16858	0.09846	0.09907	0.27051	0.12108	0.30392	0.12784	0.10831
5	0.34178	0.17925	0.16920	0.09907	0.09969	0.27112	0.12169	0.30454	0.12846	0.10892
6	0.51321	0.35068	0.34063	0.27051	0.27112	0.44255	0.29312	0.47597	0.29989	0.28035
7	0.36379	0.20125	0.19120	0.12108	0.12169	0.29312	0.14369	0.32654	0.15046	0.13092
8	0.54663	0.38409	0.37404	0.30392	0.30454	0.47597	0.32654	0.50938	0.33330	0.31377
9	0.37055	0.20802	0.19797	0.12784	0.12846	0.29989	0.15046	0.33330	0.15722	0.13769
10	0.35102	0.18848	0.17843	0.10831	0.10892	0.28035	0.13092	0.31377	0.13769	0.11815

2.) Mean LD Efficiencies for every pair of Chinese Raters (CR)

	A	B	C	D	E	F	G	H	I	J
1	0.56805	0.70106	0.53418	0.60785	0.59336	0.70404	0.77704	0.68649	0.63906	0.76957
2	0.70106	0.83407	0.66719	0.74086	0.72637	0.83705	0.91006	0.81950	0.77207	0.90259
3	0.53418	0.66719	0.50031	0.57398	0.55949	0.67017	0.74318	0.65262	0.60519	0.73571
4	0.60785	0.74086	0.57398	0.64765	0.63316	0.74384	0.81685	0.72629	0.67886	0.80938
5	0.59336	0.72637	0.55949	0.63316	0.61868	0.72935	0.80236	0.71180	0.66437	0.79489
6	0.70404	0.83705	0.67017	0.74384	0.72935	0.84003	0.91304	0.82248	0.77505	0.90557
7	0.77704	0.91006	0.74318	0.81685	0.80236	0.91304	0.98604	0.89548	0.84806	0.97857
8	0.68649	0.81950	0.65262	0.72629	0.71180	0.82248	0.89548	0.80492	0.75750	0.88801
9	0.63906	0.77207	0.60519	0.67886	0.66437	0.77505	0.84806	0.75750	0.71007	0.84059
10	0.76957	0.90259	0.73571	0.80938	0.79489	0.90557	0.97857	0.88801	0.84059	0.97110

3.) Mean LD Efficiencies for every pair of Chinese Speakers (CS)

	A	B	C	D	E	F	G	H	I	J
1	0.36934	0.48894	0.70654	0.54144	0.47226	0.35356	0.47125	0.40367	0.37884	0.58745
2	0.48894	0.60854	0.82614	0.66104	0.59186	0.47316	0.59085	0.52326	0.49844	0.70705
3	0.70654	0.82614	1.04374	0.87864	0.80946	0.69076	0.80845	0.74087	0.71604	0.92465
4	0.54144	0.66104	0.87864	0.71354	0.64436	0.52566	0.64335	0.57577	0.55094	0.75955
5	0.47226	0.59186	0.80946	0.64436	0.57518	0.45648	0.57417	0.50659	0.48176	0.69037
6	0.35356	0.47316	0.69076	0.52566	0.45648	0.33778	0.45547	0.38789	0.36306	0.57167
7	0.47125	0.59085	0.80845	0.64335	0.57417	0.45547	0.57315	0.50557	0.48075	0.68936
8	0.40367	0.52326	0.74087	0.57577	0.50659	0.38789	0.50557	0.43799	0.41316	0.62178
9	0.37884	0.49844	0.71604	0.55094	0.48176	0.36306	0.48075	0.41316	0.38834	0.59695
10	0.58745	0.70705	0.92465	0.75955	0.69037	0.57167	0.68936	0.62178	0.59695	0.80556

4.) Mean LD Efficiencies for every pair of English Raters (ER)

	A	B	C	D	E	F	G	H	I	J
1	0.27353	0.46819	0.25333	0.23956	0.24217	0.29037	0.45345	0.54360	0.25929	0.21877
2	0.46819	0.66285	0.44799	0.43422	0.43683	0.48503	0.64811	0.73826	0.45395	0.41343
3	0.25333	0.44799	0.23312	0.21936	0.22196	0.27017	0.43325	0.52340	0.23909	0.19856
4	0.23956	0.43422	0.21936	0.20559	0.20820	0.25640	0.41948	0.50963	0.22532	0.18479
5	0.24217	0.43683	0.22196	0.20820	0.21080	0.25901	0.42209	0.51224	0.22792	0.18740
6	0.29037	0.48503	0.27017	0.25640	0.25901	0.30721	0.47029	0.56044	0.27613	0.23560
7	0.45345	0.64811	0.43325	0.41948	0.42209	0.47029	0.63337	0.72352	0.43921	0.39868
8	0.54360	0.73826	0.52340	0.50963	0.51224	0.56044	0.72352	0.81367	0.52936	0.48884
9	0.25929	0.45395	0.23909	0.22532	0.22792	0.27613	0.43921	0.52936	0.24505	0.20452
10	0.21877	0.41343	0.19856	0.18479	0.18740	0.23560	0.39868	0.48884	0.20452	0.164

Academic Vita

Rachel M. Sigmund

Education: Bachelor of Science Degree in Psychology, Penn State University, Spring 2010
Concentration in Business
Honors in Psychology
Thesis Title: Inter-Rater Reliability for the Judged Accentedness of English Bilinguals
Thesis Supervisor: David Rosenbaum

Experience: Vision, Memory, and Computational Neuroscience Laboratory Research Assistant
Supervisor: Dr. Michael Wenger
September 2007- March 2008

Independent Research, Psychology Honors Program
Studied body esteem according to gender, final project title: Attitudes Towards Body Image According to Gender
Supervisor: Dr. Theresa Vescio
September 2008- December 2008

Awards: Dean's List every semester

National Honors Society

Activities: Writing Tutor at the Penn State Learning Center

Seasonal Naturalist at Trailside Nature and Science Center, Mountainside, NJ

Kappa Delta, National Sorority, Beta Theta Chapter of Penn State University

Pennsylvania State IFC/Panhellenic Dance Marathon (THON) Committee Member

Volunteer at Friends of Linden Animal Shelter, Linden, NJ

Study Abroad in Rome, Italy, Spring 2009