2010-07-02

Accuracy of English Speakers Administering Word Recognition Score Tests in Mandarin

Kaylene Barrett Polley
Brigham Young University - Provo

Accuracy of English Speakers Administering

Word Recognition Score Tests in Mandarin

Kaylene Barrett Polley

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

Master of Science

Richard W. Harris, Chair
Shawn L. Nissen
Ron W. Channell

Department of Communication Disorders

Brigham Young University

August 2010

ABSTRACT

Accuracy of English Speakers Administering

Word Recognition Score Tests in Mandarin

Kaylene Barrett Polley

Department of Communication Disorders

Master of Science

The purpose of this study was to examine the accuracy of English-speakers in determining the word recognition score of native Taiwan Mandarin-speakers. Digitally recorded Mandarin word lists were presented to 10 native Mandarin-speakers from Taiwan (five male, five female), from whom oral and written responses were collected. Oral responses were scored by 30 native English-speakers, 15 of which had no experience with Mandarin and 15 with two to three years of college-level Mandarin courses or equivalent knowledge of Mandarin. The judges who had experience with Mandarin were able to score the WRS tests with 97% accuracy (with scores ranging from 10% below to 4% above the actual score of the test). The judges without experience with Mandarin scored the WRS tests with 88.8% accuracy (with scores ranging from 34% below to 26% above the actual score of the test). An analysis of variance found that there was a significant difference between a judge's knowledge of Mandarin and his or her ability to accurately score the oral responses. An inspection of the performance of the judges in respect to the five different Mandarin tones indicated that there are some tone combinations that are more difficult to score correctly than others. While it is apparent that tone combination may play a role in the ability to accurately score WRS words in Mandarin, the implications of this for a clinical setting are uncertain because words with these tone combinations were not heard often. Tone perception training for the judges or simply making clinicians aware of this difficulty in tone identification may be of benefit in overcoming this obstacle.

Keywords: word recognition score, Mandarin, speech audiometry, Taiwan, English, scoring methods, novice, fluent, experienced

ACKNOWLEDGMENTS

Completion of this project would not have happened without the help and support of so many people. I would first like to thank my dear husband, Tom, who willingly married me half way through graduate school. It was his constant love and encouragement that got me through the times when I didn’t think I could keep going. I also would like to give thanks to my parents for cheering me on throughout my entire college experience. Thank you to Dr. Richard Harris, my thesis chair, and Dr. Shawn Nissen and Dr. Ron Channell, my committee members, for their guidance and feedback. Thank you to the BYU Mentoring Environment Grant, which provided, which provided the financial support for this project. A special thank you to all those who participated in my study and to those who helped me find people for my project. I would not have been able to complete my project without you.

Table of Contents

List of Tables

List of Figures

List of Appendixes

Appendix Page

Introduction

The ability to comprehend speech is an integral part of communication between individuals. A breakdown in speech comprehension often motivates a person to seek the help of an audiologist (Brandy, 2002). While pure-tone testing is a useful clinical tool, speech audiometry is needed to give a more accurate representation of an individual's ability to understand speech (American Speech-Language Hearing Association [ASHA], 1988; Brandy, 2002). During routine comprehensive audiometric evaluations, two main tests are used in speech audiometry, called the speech recognition threshold (SRT) and the word recognition score (WRS). The SRT is the lowest intensity level at which an individual can understand speech half of the time (ASHA, 1988; Brandy, 2002; Martin & Clark, 2009; Martin & Stauffer, 1975). The WRS indicates how well a patient is able to correctly understand a list of words at a suprathreshold level (Brandy, 2002).

Many speech audiometry materials that have been designed were developed in the English language. However, these materials produced less valid results when administered to speakers of other languages (Danhauer, Crawford, & Edgerton, 1984; Ramkissoon, Proctor, Lansing, & Bilger, 2002; von Hapsburg, Champlin, & Shetty, 2004). This has created a need for valid speech audiometry materials for speakers of other languages. In response to this need, a number of speech audiometry materials have been developed in other languages, including Russian (Aleksandrovsky, McCullough, & Wilson, 1998; Harris et al., 2007), Spanish (Cokely & Yager, 1993; Comstock & Martin, 1984; Martin & Hart, 1978; McCullough & Wilson, 2001; Spitzer, 1980; Zubick et al., 1983), Greek (Iliadou, Fourakis, Vakalos, Hawks, & Kaprinis, 2006), Cantonese (Conklin, 2007; Kim, 2007; Lau & So, 1988), Korean (Harris, Kim, & Eggett, 2003a, 2003b), Polish (Harris, Nielson, McPherson, Skarzynski, & Eggett, 2004a, 2004b), and

Mandarin (Nissen, Harris, Jennings, Eggett, & Buck, 2005a, 2005b; Nissen, Harris, & Slade, 2007; Wang, Mannell, Newall, Zhang, & Han, 2007).

The difficulty with administering SRT and WRS tests in a language other than English is that the majority of audiologists in the United States do not speak a second language (Ramkissoon et al., 2002). Ramkissoon et al. warns that to ensure the validity of test scores, clinicians need to be proficient in the language of the test being administered. Even if clinicians have knowledge of the language they are testing in, such as Spanish, this knowledge may be limited to a specific dialect and may still restrict the clinicians' ability to accurately score a test in any other dialect of that language (Cokely & Yager, 1993; Spitzer, 1980). Studies have shown that differences in one's experiences with language play a role in how an individual processes language (Sanders, Neville, & Woldorff, 2002). In their study, Sanders et al. looked at the ability of native Japanese- and native Spanish-speakers (both late-learners of English and nearly monolingual speakers) to determine whether specific sounds were found at the beginning or the middle of words in English sentences. Sanders et al. found that the native Japanese- and Spanish-speakers who had learned English used syntactic information to segment the sounds, but not to the same extent as the native English-speakers. Both groups also used stress patterns to help segment sounds, but the extent to which stress patterns were used to help with segmenting sounds depended on the stress pattern characteristics of the individual's native language.

Cokely and Yager (1993) investigated the ability of two groups of native English-speakers to accurately score a WRS test in Spanish. One group of participants had no knowledge of Spanish and the other had two to three years of college Spanish. The scores were compared to written responses given by native Spanish-speaking participants. The mean overall WRS scores calculated from the English-speaking judges were 59.8% (*SD* = 16.0) for those with a knowledge of Spanish and 60.2% (*SD* = 15.7) for those without. While the scores produced by the judges with some knowledge of Spanish were more accurate, an analysis of variance revealed that the difference between the scores for the two groups of English-speaking judges was not significant. In addition, there was excellent interjudge reliability within both groups of participants as judged by correlation measures. When comparing the 15 individual scores of the judges with a knowledge of Spanish with each other, there was a Pearson correlation of .99. When comparing the 15 individual scores of the judges without knowledge of Spanish with each other, there was a Pearson correlation of .98.

The purpose of the present investigation was to determine whether or not the same results found by Cokely and Yager (1993) in their study of scoring WRS tests would occur with a tonal language, such as Mandarin. A series of WRS tests were administered to a group of 10 native Mandarin-speakers from Taiwan, who responded in both oral and written modes. Just as in the Cokely and Yager study, the oral responses were judged as correct or incorrect by both groups of native English-speakers. The scores of the native English-speakers were compared with the written responses to determine whether or not having knowledge of Taiwan Mandarin impacts a listener's ability to accurately score WRS tests in Mandarin.

Review of Literature

The inability (or perceived difficulty) in understanding speech is one of the main reasons why an individual would visit an audiologist. Because the ability to communicate is dependent on comprehending speech, it is important to not only test one's ability to hear pure-tones but also to test how well speech is understood. Speech audiometry, therefore, is used to assess an individual's threshold for speech (ASHA, 1988; Brandy, 2002). According to Ramkissoon

(2001), diagnostic hearing testing for adults and children is incomplete without assessing the ability to understand complex signals.

Speech Audiometry Measurement

Speech Recognition Threshold. One of the two main tests used to examine an individual's communication abilities is the SRT, which is defined as the lowest intensity level at which an individual can comprehend speech 50% of the time (ASHA, 1988; Brandy, 2002; Dennis & Neely, 1991; Egan, 1979; Martin & Clark, 2009; Martin & Stauffer, 1975). In English, the SRT test typically uses spondaic words (two-syllable words with equal stress). Sometimes a carrier phrase is used in the SRT test, such as *Say the word,* followed by the target word (Brandy, 2002; Martin & Clark, 2009). Patients are familiarized with the words used to measure the SRT in order to test their hearing ability and not their vocabulary abilities (ASHA, 1988; Thibodeau, 2007).

There are a number of clinical purposes of the SRT, such as providing an estimate of one's hearing sensitivity to speech. The SRT is also used to check the reliability of pure-tone test results. Brandy (2002) indicated that the SRT and pure-tone average (PTA) should agree within 6-12 dB of each other. If the SRT and PTA differ by 13 dB or more the tests are considered to have poor agreement. Egan (1979) explained that reasons for poor agreement between the SRT and PTA may include poor test administration technique, inaccurate instrumentation calibration, language barriers, or attempts at exaggerating a hearing impairment. The SRT is also used to provide a baseline for determining the presentation levels of suprathreshold word recognition testing (Brandy, 2002; Egan, 1979; Martin & Stauffer, 1975).

Word Recognition Score. The second measure used frequently in speech audiometry testing is the WRS. The WRS indicates how well a patient is able to identify a list of words at a suprathreshold level, where the score obtained from the test is the percentage of words that are identified correctly (Brandy, 2002; Carhart, 1965). The WRS is designed to help the clinician better understand a patient's level of speech communication difficulty. In general, scores of 90% or higher are considered normal (Dennis & Neely, 1991; Egan, 1979). However, as the WRS score decreases, the chance of an individual having a perceptual deficit increases (Hoople, 1952). In English, monosyllabic words are typically used for WRS tests instead of conversational sentences or spondaic words. Carhart (1965) explains that monosyllabic words are less redundant and more unpredictable without being as confusing as nonsense words.

Carhart (1965) states that word recognition testing is helpful with diagnosing central auditory system disorders and that the WRS helps with differential diagnosis, such as distinguishing between an 8^{th} cranial nerve tumor and unilateral Meniere's disease. Thibodeau (2007) explained that the WRS helps distinguish between types of hearing impairment as individuals with a conductive hearing impairment usually are able to recognize single words very well, at least 90% of the time. However, a patient with a sensorineural impairment will have a reduced level of speech recognition. The WRS test helps audiologists to: (a) determine the extent of the speech recognition difficulty, (b) aid in diagnosing the site of the disorder in the auditory system, (c) assist in the determination of the need for and proper selection of amplification systems, and (d) make a prognosis for the outcome of treatment efforts (Dennis & Neely, 1991; Egan, 1979; Martin & Clark, 2009).

History of Speech Audiometry Development. The first use of speech materials in a clinical setting has been credited to Oscar Wolf as early as 1874 (Hood & Poole, 1980). However, the methods of speech audiometry as they are used today most likely originates from word lists that were created during World War II in the Psycho-Acoustic Laboratory (PAL) at

Harvard University (Hood & Poole, 1980). In 1947, Hudgins, Hawkins, Karlin and Stevens developed two spondaic (two-syllable) word lists used for SRT testing. Specifically, the tests were titled the PAL Auditory Tests No. 9 and No. 12 and were used for the measurement of the SRT using spondaic words and sentences, respectively (Hirsh et al., 1952). The PAL Auditory Test No. 14 was also developed, which also measured SRT levels using spondaic words (Garvey, 1953). In 1948, Egan and colleagues developed twenty 50-word phonetically balanced monosyllabic word lists (PAL PB 50) used for word recognition testing.

Reports had been made regarding deficiencies in the Harvard PAL tests. As a result, the PAL lists developed by Hudgins et al. (1947) were modified by Hirsh et al. in 1952 at the Central Institute for the Deaf. Modifications to these tests were made to include words for the tests which were more familiar and homogeneous with respect to intelligibility (Hirsh et al., 1952). The 84 original spondaic words from the PAL Auditory Tests No. 9 and No. 14 were modified to contain only 36 spondees (Hirsh et al., 1952). These changes were referred to as CID Auditory Test W-2 and W-1, respectively. Both the CID W-1 and W-2 tests contain the same 36 spondee words, with the difference being that the W-1 words are presented at a constant level and the W-2 words are presented at descending intensity levels (Hirsh et al., 1952).

Hirsh et al. (1952) also modified the PAL PB-50 lists developed by Egan (1948) because reports had shown that the vocabulary of these lists was too large for many patients. The lists were more rigidly phonetically balanced so that the speech sounds within the list occurred with the same frequency as they do in a sample of English speech. This resulted in a smaller and more familiar test vocabulary. The modified versions of the PAL PB-50 lists were called the CID Auditory Test W-22. However, the W-22 lists have been criticized as being too easy to perceive (Tillman & Carhart, 1966). As a result, Tillman and Carhart (1966) created WRS lists that were phonemically balanced, called the NU-6 lists. Lehiste and Peterson (1959) explained that for a list of words to be phonemically balanced, each initial and final consonant as well as each vowel, needs to have the same frequency of occurrence within the list as it does in the language as a whole.

Speech audiometry provides useful information regarding the educational and rehabilitation needs of a patient, which includes determining the social impact of a person's hearing impairment (Carhart, 1965). Thibodeau explains that speech audiometry testing should be part of the hearing aid selection process in addition to the diagnostic portion of the evaluation. This is because the speech processing abilities of individuals vary, even when their pure-tone audiograms are similar. For example, individuals with sensorineural impairments have difficulty discriminating among various speech sounds regardless of how loud they are presented, while most individuals with conductive impairments benefit from increased presentation intensity levels (Martin & Clark, 2009). As a result, even if two patients demonstrate identical threshold reductions, the lifestyle modifications needed for one patient may be different than the other based on the nature of the individual's hearing impairment (Thibodeau, 2007).

Hall (1983) explained that another important function of speech audiometry is to help in the diagnosis of neural auditory dysfunction. Difficulties in speech perception are associated with neural problems, including the central and peripheral auditory system. Disagreement between pure-tone and speech audiometry tests may indicate the presence of a central auditory processing disorder, pseudohypacusis, or irregular auditory sensitivity, among other things (ASHA, 1988; Carhart, 1951a, 1952; Epstein, 1978; Kerr & Smyth, 1972). However, when there is a discrepancy between pure-tone and speech audiometric test results, it is imperative to rule out the possibility of equipment failure or misunderstanding of the task on the part of the patient before arriving at a diagnosis (ASHA, 1988; Carhart, 1965).

Procedures. The word lists used for WRS tests ought to be homogenous with each other, meaning that they are similar in at least one or more dimensions, such as intelligibility. Psychometric functions, which are the "probability of a correct response as a function of sound level," are used to create equally intelligible words based off of two assumptions: "(a) the level at which the probability of a correct response [P(C)] is .50 (word threshold) and (b) the slope of that function at its midpoint, P(C) = .50" (Bilger, Matthies, Meyer, & Griffiths, 1998, p. 517). For words lists to be equivalent, there should be no measurable difference between the scores obtained on separate tests (Campbell, 1965). Bilger et al. cautioned, however, that a list of words that is equally intelligible for normal-hearing individuals may not be equally intelligible for individuals who are hearing impaired. In addition, Campbell explains that word lists ought to be homogeneous in terms of their level of difficulty.

Different methods of scoring speech audiometry tests have been documented. Nelson and Chaiklin (1970) explained that examiners may use either a talkback or write-down method in scoring the results of SRT and WRS tests. They warn about the possibility of clinicians leaning towards scoring ambiguous items as correct when using a talkback method. Nelson and Chaiklin suggest that if a word sounds even somewhat ambiguous, the clinician ought to ask the patient to repeat or clarify the word, noting that extra caution should be taken when a patient has poor speech articulation. Martin and Clark (2009) explained that the patient's method of response (be it oral, written, or identifying a picture or object) will depend on the type of test being given. Picture or object identification tests are usually only used for small children or adults with special needs who are otherwise unable or unwilling to participate in a traditional test (Martin &

Clark, 2009). Written responses are useful only when the test is able to be scored after its completion, as some tests require an instantaneous judgment from the audiologist. However, written responses eliminate any errors that may be made due to the articulation of the patient (Martin & Clark, 2009).

Speech Audiometry with Non-native Listeners

Over time, the amount of ethnically, racially, and linguistically diverse individuals being seen by an audiologist has been increasing (Beverly-Ducker, 2003). This trend is demonstrated in the 2000 Omnibus Survey Practice Trends in Audiology, which reported that 26% of the average monthly caseload for audiologists consisted of individuals whose race was something other than Caucasian, with a monthly average of 11 clients either signing or speaking a language other than English (ASHA, 2000). In addition, the majority of audiologists are administering these English test materials to minority individuals (Ramkissoon, 2001).

Ferguson, Jongman, Sereno, and Keum (2010) found that English words produced by a native Spanish-speaker were significantly less intelligible than those produced by native English-speakers, regardless of the listener's age, hearing ability or listening environment. Due to the difficulties in perceiving a non-native language, English materials may not be valid when used with native-speakers of other languages. Not taking a listener's native language into consideration may result in false test interpretations and inaccurate conclusions regarding the hearing of these individuals (Carhart, 1951a, 1951b; Danhauer et al., 1984; McCullough & Wilson, 2001; Ramkissoon et al., 2002; von Hapsburg et al., 2004). Carhart (1965) mentioned that it is important to consider the linguistic background of the patient before selecting an appropriate test, so as to avoid the situation of having the words presented sound like nonsense words to the patient.

Modifications may at times be made to accommodate non-native English-speakers. One kind of modification would be to eliminate some of the words from the word list to increase the likelihood of the words being familiar to the patient. For example, while most of the 36 CID Auditory Test W-1 spondaic words are relatively familiar to most individuals, some are not (such as inkwell or whitewash). However, eliminating these words may lead to an inaccurate representation of the patient's speech recognition threshold, resulting in inaccurate diagnostic conclusions (Ramkissoon, 2001; Ramkissoon et al., 2002; Ramkissoon & Khan, 2003). Due to the need to accurately represent the speech recognition abilities of non English-speaking individuals, a number of speech audiometry materials have been developed in various languages, including Mandarin (Nissen et al., 2005a, 2005b; Nissen et al., 2007; Wang et al., 2007). Ramkissoon and Khan (2003) state, however, that a major draw-back to non-English speech tests is their limited use and application. They conclude that this is a result of most clinicians not being proficient in the language of the test he or she would be administering.

Another modification that may be made to speech audiometry materials when the audiologist and patient speak different languages is to use materials in the patient's native language. However, approximately 80% of audiologists in the United States are monolingual English-speakers, and almost 14% of the United States population does not speak English (Ramkissoon & Khan, 2003). Ramkissoon et al. (2002) warn that clinicians need to be proficient in the language of the test that they are administering to ensure the validity of the scores received. A basic understanding of a patient's language is not enough to ensure an accurate interaction with the patient (Ramkissoon & Khan, 2003). Even if the clinician is familiar with a foreign language, his or her knowledge may be limited to a certain dialect and therefore accidentally score pronunciations of other dialects as incorrect (Cokely & Yager, 1993; Spitzer, 1980).

Acquisition of Speech Perception. Acquiring a second language in adulthood is not an easy task, where achieving a native-like proficiency is difficult for many second language learners (Ramkissoon, 2001). While many adults may acquire a native-like use of the syntax and lexicon of a language, it is less likely for someone to acquire a native accent after puberty (de Jonge, 1995). Studies have shown that an individual's first language can affect a person's ability to learn or perceive the phonology of a second language. Research has also shown that a listener's ability to perceive phonemes in a certain language relates to the individual's experience with that language (de Jonge, 1995; Eckman, 2004; Keith, Katbamna, Tawfik, & Smolak, 1987; Werker, Gilbert, Humphrey, & Tees, 1981). Individuals are better at identifying sounds when listening to sounds produced in their own language versus any other language (Singh & Black, 1966). In addition, native and non-native speakers of a language perceive some speech contrasts differently (de Jonge, 1995; Garcia de las Bayonas, 2004). Cutler (2004) indicated that an individual's speech perception is slower and less accurate when listening to a non-native language.

Tonal languages, such as Mandarin, have long been a source of difficulty for English-speakers who are learning such languages (White, 1981). The difficulty in learning and perceiving Mandarin tones has been attributed to the inference of the suprasegmental properties of the tones with those of the English language (Wang, Spence, Jongman, & Sereno, 1999). Mandarin tones are used as lexical markers, whereas pitch variations do not carry such information in English (Lee, Vakoch, & Wurm, 1996). The differences in the way stress is perceived also varies between English and Mandarin, which may also cause difficulty for an

English-speaker in perceiving tones correctly (White, 1981). Broselow, Hurtig, and Ringen (1987) and Shen and Lin (1991) studied which tones are more difficult to perceive. It was found that Tone 4 is more difficult to perceive in a non-final position of a word (Broselow et al., 1987). Shen and Lin found that both Tones 2 and 3 have similar fundamental frequency contours, and therefore it is more difficult to discriminate between the two correctly.

Effectiveness of Speech Perception Training for Non-native Listeners. For many years, the opinion of linguistic experts was that an individual's early experience with language had a permanent effect on one's perceptual abilities. Therefore, it was thought to be difficult for anyone to learn to discriminate differences that were not already present in one's native language (Pisoni, Aslin, Perey, & Hennessy, 1982). However, in recent years studies have shown that with training individuals are capable of learning to distinguish between non-native phonemes. Pisoni et al. looked at an adult's ability to discriminate between voice onset time (VOT) contrasts that are not present in his or her native language. Pisoni et al. carried out three different experiments where they adjusted the amount of discrimination training the participants received, as well as whether or not the participants were given immediate feedback on their responses. Pisoni et al. state that overall, the results of their experiments showed that individuals are capable of learning to discriminate between differences in the negative region of the VOT continuum.

Building on the findings of Pisoni et al. (1982), McClaskey, Pisoni, and Carrell (1983) found that individuals trained to detect VOT contrasts in one place of articulation were able to perceive contrasts in a different place of articulation without further training. In the McClaskey et al. study, the participants were trained to detect three categories of VOT in either a labial or alveolar articulation placement. The participants were given immediate feedback on their responses to train them to detect the differences in VOT. Later, the participants were tested using the articulation placement stimuli they were not trained on with no feedback given (McClaskey et al., 1983). McClaskey et al. found that the participants were able to successfully perceive the VOT differences that they were not specifically trained to detect.

In 1991, Logan, Lively, and Pisoni investigated the ability of native Japanese-speakers to be trained to differentiate between the /r/ and /l/ phonemes. The training included the use of minimal pair words that contrasted the /r/ and /l/ phonemes in either the initial or final position of words. Logan et al. used a pretest-posttest format, with a three week training period between the two tests. They found a significant increase in the number of correct responses from the pretest to the posttest (pretest mean = 78.1%, posttest mean = 85.9%). Overall, they found that training could improve a Japanese listener's ability to distinguish between /r/ and /l/. However, they note that the participants' performance on both tests was dependent on where in the word the phoneme was located, as well as the talker. Familiarity with the talker (i.e. they had heard that talker previously) increased the likelihood that the subject would distinguish between the phonemes with greater accuracy (Logan et al., 1991). A number of years later, Lively, Pisoni, Yamada, Tohkura, and Yamada (1994) conducted a similar study. In their study, Lively et al. used a similar training format as Logan et al., including a pretest and posttest to measure the effects of training. However, Lively et al. readministered tests to their participants again after three months and six months to determine how well the participants retained what they had learned during the training period. The pretest-posttest gains were similar to those reported by Logan et al., with a pretest mean of 65% correct to a posttest mean of 77% correct. After three months, each of the participants in the study tested at least 8% above their original pretest score. After six months, accuracy in the participants' test scores was at least 4.5% higher than their pretest score. It was also noted in the Lively et al. study that the same discrepancy of accuracy among talkers occurred like it had in the Logan et al. study.

Wang et al. (1999) looked at the ability of English-speakers who have taken one or two semesters of Mandarin Chinese to learn to perceive Mandarin tones more accurately. The students were given a pretest, trained, and then given a posttest. The participants were also given a retention test six months later. It was found that the students were able to increase their ability to perceive tones correctly by 21%, and retained that increase in ability over a six month period.

Impact of Monolingual Clinicians in Administering WRS Tests. Cokely and Yager (1993) examined the accuracy of native English-speakers with either no knowledge of Spanish or two to three years of college Spanish in scoring word recognition tests in Spanish. WRS tests were administered to native Spanish-speakers. The Spanish-speaking participants not only responded verbally to the stimuli but also wrote down the words they thought they heard. The study indicated that the scores received by the two groups were nearly identical, though the results of the English group with some knowledge of Spanish were more accurate when compared to the written responses. The overall mean WRS scores calculated from the English-speaking judges were 59.8% (SD = 16.0) for those with a knowledge of Spanish and 60.2% (SD = 15.7) for those without. An analysis of variance revealed that there was not a significant difference between the two groups of English-speaking judges, $F(1, 28) = 0.01$, $p > .9$. In addition, there was excellent interjudge reliability within both groups of judges as shown through correlation measures. When comparing the responses of the judges to the other judges within each group, a Pearson correlation showed a mean correlation coefficient of .99 among the responses of the judges with knowledge of Spanish. In addition, there was a mean correlation coefficient of .98 among the responses of the judges without knowledge of Spanish.

A few studies have also been conducted to create Spanish speech audiometry materials that are designed in such a way that an English-speaking clinician could accurately score the test (Comstock & Martin, 1984; Martin & Hart, 1978; Spitzer, 1980). The need in these studies to be familiar with Spanish was overcome by requiring the patients to use a picture-pointing response instead of using an oral response. The stimulus words were recorded using dual recordings, one channel with the words in Spanish and the second channel with the words in English. The participant would hear the word in Spanish while the clinician would hear the word in English. The clinician could then determine whether or not the correct response was given.

The linguistic structure of Spanish and English are relatively similar, both languages have comparable phoneme inventories and do not rely on suprasegmental tone in lexical manner. Therefore, it was uncertain if the same Spanish results of Cokely and Yager's (1993) study would be found if the same study were to be done with a language such as Mandarin that utilizes a subset of phonemes not found in English and relies on suprasegmental tones to express meaning. There are five tones in the Mandarin language. The four main tones in Mandarin are numbered and described as (1) high level, (2) high rising, (3) low rising, and (4) high falling to low (Katzner, 1995). The fifth tone is a neutral tone that occurs when syllables are unstressed; it also does not appear in the initial position of a word (Wu, Yu, Zhang, & Tian, 2006).

Since there is no data concerning the use of non-native Mandarin-speaking clinicians in administering speech audiometry tests in Taiwan Mandarin, the purpose of the present study is to (a) identify native Mandarin-speakers to whom to administer WRS tests, (b) identify native English-speakers with either no or partial knowledge of Mandarin to score the tests in Mandarin, (c) compare the results of the two groups of English-speakers with written responses from the Mandarin participants to determine if there are significant differences between any or all of the groups. This information will be useful in determining the efficacy of non Mandarin-speaking clinicians in administering Mandarin speech audiometry tests. The present study will only use Taiwan Mandarin so as to reduce the effect of dialect.

Method

The present study was divided into two separate phases, following the methodology of Cokely and Yager (1993). In the first phase, native Mandarin-speaking listeners both repeated and wrote down bisyllabic Mandarin words that were presented to them at intensity levels between 2-8 dB HL. Their oral responses were recorded for use during the second phase, in which English-speaking individuals listened to the audio recordings to judge whether or not they thought the words were repeated correctly.

Phase 1

Participants. The participants in this phase of the study were 10 native Mandarin-speakers from Taiwan. The main dialect in Mainland China is called Mandarin, also known as pǔtōnghuà, or common speech. Mandarin is also the official language of Taiwan, called guóyǔ. However, guóyǔ differs slightly from pǔtōnghuà in both vocabulary and grammar (Li & Thompson, 1987). All Mandarin participants reported that they continue to speak Taiwan Mandarin on at least an every-other-day basis. All participants had normal hearing, defined as pure-tone thresholds ≤ 15 dB HL at octave and mid-octave frequencies from 125 to 8000 Hz. Participants were also required to have static acoustic admittance between 0.3 and 1.4 mmhos with peak pressure between -10 and +50 daPa (ASHA, 1990; Roup, Wiley, Safady, & Stoppenbach, 1998). All of the participants signed an informed consent (See Appendix A). Each participant completed two 1-hour listening sessions.

Procedure and Materials. Eight lists of 50 bisyllabic words were presented to the participants (four male and four female lists). The words lists were obtained from Brigham Young University and contained words which were judged to be familiar words to speakers of Taiwan Mandarin (Nissen, Harris, & Dukes, 2008).

All testing of the Taiwan Mandarin participants was performed in a double-walled sound booth which meets ANSI S3.1 standards for maximum permissible ambient noise levels for the ears not covered condition using one-third octave-bands (American National Standards Institute [ANSI], 1999). A Grason-Stadler model 1761 audiometer was used with TDH-50P Telephonics supra-aural earphones. Prior to testing each Taiwan Mandarin participant, the external inputs to the audiometer were calibrated to 0 VU using a 1000 Hz calibration tone designed for use with the recorded Taiwan Mandarin materials. The audiometer was calibrated prior to, weekly during, and at the conclusion of data collection, in accordance with ANSI S3.6 specifications (ANSI, 2004). No changes in calibration were necessary throughout the course of data collection.

The order of presentation of the eight lists was randomized for each subject. The words were presented to the Taiwan Mandarin-speaking participants at an intensity level so as to result in approximately 50-70% correct recognition of the stimulus words. The Taiwan Mandarin-speaking participants responded to the stimuli in both written and oral forms. Oral responses were recorded using a Larson-Davis model 2541 microphone, positioned at 0^{o} azimuth and covered by a 7.62 cm windscreen at a distance of approximately 15 cm from the participant's mouth. The microphone was connected to a Larson-Davis model 900 microphone preamp, which was coupled to a Larson-Davis model 2221 power supply. The signal was digitized using a Benchmark ADC1 analog-to-digital converter and then stored on a hard drive for later editing by the SADiE Disk Editor software, Series 5 (Studio Audio & Video Ltd., 2007). Written responses were scored by each individual Taiwan Mandarin-speaking participant. Recorded oral responses were later scored as correct or incorrect by both groups of English-speakers.

Phase 2

Participants. The participants in the second phase were 30 native English-speakers, split into two groups. One group had no knowledge of or experience with Mandarin (novice) and the other group had two to three years of college-level Mandarin courses or an equivalent knowledge of Taiwan Mandarin (experienced), many of whom who had approximately 15-20 months of in country experience. All participants had normal hearing, defined as pure-tone thresholds ≤ 15 dB HL at octave and mid-octave frequencies from 125 to 8000 Hz. Participants were also required to have static acoustic admittance between 0.3 and 1.4 mmhos with peak pressure between -10 and +50 daPa (ASHA, 1990; Roup et al., 1998). All of the participants signed an informed consent (See Appendix A).

Procedures and Materials. Prior to the listening session, instructions were read to each judge (See Appendix B). Each judge was provided with the written instructions to follow along with as well. Each judge was placed in a double-walled sound suite that meets ANSI S3.1 standards for maximum permissible ambient noise levels for the ears not covered condition using one-third octave-bands (ANSI, 1999). Sixteen of the recordings by the Taiwan Mandarin-speaking participants in Phase 1 were randomly chosen to be presented to the English-speaking judges. The recordings were presented to the judges using a Benchmark DAC1 digital-to-analog converter connected to a Grason-Stadler model 1761 audiometer. The recordings were then routed to the judge via a single TDH-50P headphone. The English-speaking judges heard a recording of the stimulus word followed by the response given by the Taiwan Mandarin-speaker at a level of 60 dB HL. Each judge had a list of the Mandarin words being presented (with both the Mandarin characters and Romanization of the words listed) in front of them in an Excel spreadsheet. The judges were instructed to listen to the stimulus word and then score the Taiwan Mandarin-speaker's response as either correct or incorrect.

Results

Phase 1

In Phase 1 of the study, eight WRS tests were administered to participants from Taiwan at low intensity levels so as to receive percentage correct scores between 50-70%. Sixteen of the recorded lists were randomly chosen to be presented to the English-speaking judges. Of those 16 lists, the actual scores received by the Taiwan Mandarin-speaking participants ranged from 50-74% with a mean score of 62% (*SD* = 7.11). The 16 lists were comprised of four different female word lists and three different male word lists. Table 1 presents the frequency that the different combinations of the Mandarin tones appeared within those 16 lists. Also presented in Table 1 is the proportion of each tone combination that was correctly identified by all of the study participants.

Phase 2

The overall mean accuracy of scoring the Taiwan Mandarin WRS lists by the experienced and novice judges was 97.0% and 88.8% respectively. Out of all of the responses given by the novice judges, 54% were a correct response scored as correct (hit), 7% were a correct response scored as incorrect (miss), 3% were an incorrect response scored as correct (false alarm), and 34% were incorrect responses scored as incorrect (correct rejection). Out of all of the responses given by the experienced judges, 59% were hits, 2% were misses, 0.8% were false alarms, and 37% were correct rejections. Among the experienced judges, little variability

Table 1

Frequency of Mandarin Tone Combinations, and Proportions of Tone Combinations Correctly Identified by Study Participants

Tone Combination	Count[a]	Times Heard[b]	All[c]	Novice[d]	Experienced[e]
1 – 1	14	59	0.94	0.90	0.98
1 – 2	11	37	0.91	0.85	0.98
1 – 3	6	27	0.95	0.92	0.99
1 – 4	26	79	0.95	0.91	0.98
1 – 5	6	16	0.95	0.92	0.99
2 – 1	9	25	0.92	0.86	0.98
2 – 2	10	40	0.95	0.92	0.98
2 – 3	6	23	0.86	0.83	0.89
2 – 4	22	93	0.92	0.88	0.95
2 – 5	7	26	0.91	0.86	0.98
3 – 1	6	24	0.93	0.88	0.99
3 – 2	6	23	0.95	0.91	0.99
3 – 3	8	29	0.96	0.94	0.99
3 – 4	10	40	0.92	0.87	0.97
3 – 5	4	13	0.87	0.84	0.91
4 – 1	4	10	0.85	0.83	0.87
4 – 2	20	81	0.94	0.90	0.99
4 – 3	8	20	0.94	0.90	0.98
4 – 4	37	132	0.90	0.86	0.95
4 – 5	2	3	1.00	0.98	1.02

[a]The number of times each tone combination appeared within the seven word lists. [b]The number of times each tone combination was heard by each judge over the 16 recordings. [c]Proportion of correct responses obtained by all judges combined. [d]Proportion of correct responses obtained by the novice judges. [e]Proportion of correct responses obtained by the experienced judges.

was found between the accuracy of their responses within the group. These scores had a mean absolute difference of 1.91% from the actual score of the lists. The scores ranged from 4 percentage points higher than the actual score to 10 percentage points below ($M = 97.0$, $SD = 2.39$). However, a greater range of variability was found among the responses of the novice judges. Novice judge scores had a mean absolute difference score of 7.09% deviation from the actual score of a given list. The scores for the novice participants ranged from 26 percentage points higher than the actual score to 34 percentage points below ($M = 88.8$, $SD = 9.10$). To determine if there was a significant difference regarding the level of Mandarin experience between the two groups of judges, the data were analyzed using an ANOVA**.** This analysis revealed that there was a significant difference between the two groups of judges with respect to experience with Mandarin, $F(1, 28) = 89.94$, $p < .0001$.

The accuracy of the scores for both groups of participants obtained from lists with a female talker ranged from 90.4% to 94.0% and from 91.7% to 94.5% for lists with a male talker. A significant difference was found with respect to talker gender of the lists for both groups of participants, $F(6, 174) = 7.66$, $p < .0001$. A significant effect was also found in the scores obtained based on the talker gender of the recorded lists, $t(174) = 2.59$, $p = .01$. However, even though a statistical significance was found, clinically, the differences among the lists based on talker gender were not significant.

Since Mandarin is a tonal language, the effect of tone on a judge’s ability to score the lists was investigated. Table 2 presents the percentages of hits, misses, false alarms, and correct rejections for both groups of judges that were obtained for each of the different tone combinations. Just considering the accuracy of the scores for the tone of the first syllable for all

Table 2

Percentages of Hits, Misses, False Alarms, and Correct Rejection with Respect to Mandarin Tone Combinations

Tone Combination	Hit N	Hit E	Miss N	Miss E	False Alarm N	False Alarm E	Correct Rejection N	Correct Rejection E
1 – 1	8.52	8.42	6.48	3.97	5.14	3.88	6.00	5.97
1 – 2	4.86	5.05	6.37	4.76	5.37	0.00	3.80	4.04
1 – 3	3.26	3.27	2.96	1.98	0.70	0.00	3.92	3.70
1 – 4	9.25	9.25	7.79	3.97	7.94	10.68	11.52	11.20
1 – 5	2.27	2.25	1.87	1.19	0.47	0.00	1.77	1.68
2 – 1	2.80	3.05	4.61	2.38	3.26	0.97	3.29	3.34
2 – 2	7.11	6.93	5.05	4.76	0.47	0.00	2.13	2.02
2 – 3	3.03	2.89	2.96	6.75	6.31	15.53	2.25	2.33
2 – 4	8.68	8.42	6.92	9.92	21.26	31.07	16.31	16.45
2 – 5	3.57	3.67	4.06	2.38	4.44	2.91	2.44	2.63
3 – 1	2.48	2.67	3.62	1.19	2.10	0.00	3.78	3.70
3 – 2	3.28	3.30	2.85	1.19	0.47	0.00	2.49	2.36
3 – 3	3.23	3.13	1.54	0.00	3.04	2.91	4.77	4.64
3 – 4	5.41	5.59	7.35	7.14	2.10	0.00	4.14	4.04
3 – 5	0.93	1.02	3.18	6.75	0.70	0.00	2.74	2.36
4 – 1	1.06	1.04	2.31	5.95	1.17	4.85	1.33	1.23
4 – 2	10.74	10.66	8.67	5.56	8.41	1.94	9.66	9.72
4 – 3	2.50	2.64	3.51	1.98	0.00	0.00	2.54	2.36
4 – 4	16.59	16.33	17.78	28.17	26.64	25.24	15.03	15.91
4 – 5	0.44	0.42	0.11	0.00	0.00	0.00	0.36	0.34

Note. N = Novice; E = Experienced

of the judges ranged from 91.1% to 94.0%. The accuracy of the scores based on the tone of the second syllable for both groups of judges ranged from 90.8% to 93.7%.

The accuracy in scoring the WRS lists with regard to the different tone combinations ranged from 84.7% to 100% for both groups of judges. Table 1 presents the percentage of correctly identified tone combinations for the novice and experienced judges, and for both groups combined. The results showed that the most difficult tone combination for both groups of participants to hear were words which had 4th tone on the first syllable and 1st tone on the second syllable (4 – 1), with 84.7% scored correctly, whereas the easiest tone combination was 4 – 5, with 100% scored correctly (see Figure 1). There was much variability among the tone interactions for the novice judges, with accuracy ranging from 83% to 98%. The tone combinations of 4 – 1 and 2 – 3 were the most difficult for the novice judges to score correctly, with 83% scored correctly for both combinations. The combination of 4 – 5 was the easiest for the novice judges to score, with 98% scored correctly (see Figure 2). There was much less variability among the accuracy of the scores obtained from the experienced judges. However, three tone combinations appeared to be more difficult for the experienced judges than any of the others. These tone combinations are 4 – 1 (87% scored correctly), 2 – 3 (89% scored correctly), and 3 – 5 (91% scored correctly). The easiest tone combination for the experienced judges to hear was 4 – 5, with 100% scored correctly (see Figure 3). A significant effect was found for both groups of judges based on the tone of the first syllable, $F(3, 87) = 10.61$, $p < .0001$, as well as for the tone of the second syllable, $F(4, 116) = 6.61$, $p < .0001$. However, these differences are not significant from a clinical standpoint. A significant interaction was also found between the tones of the first and second syllables of the words for both groups of judges, $F(12, 348) = 12.63$, $p < .0001$.

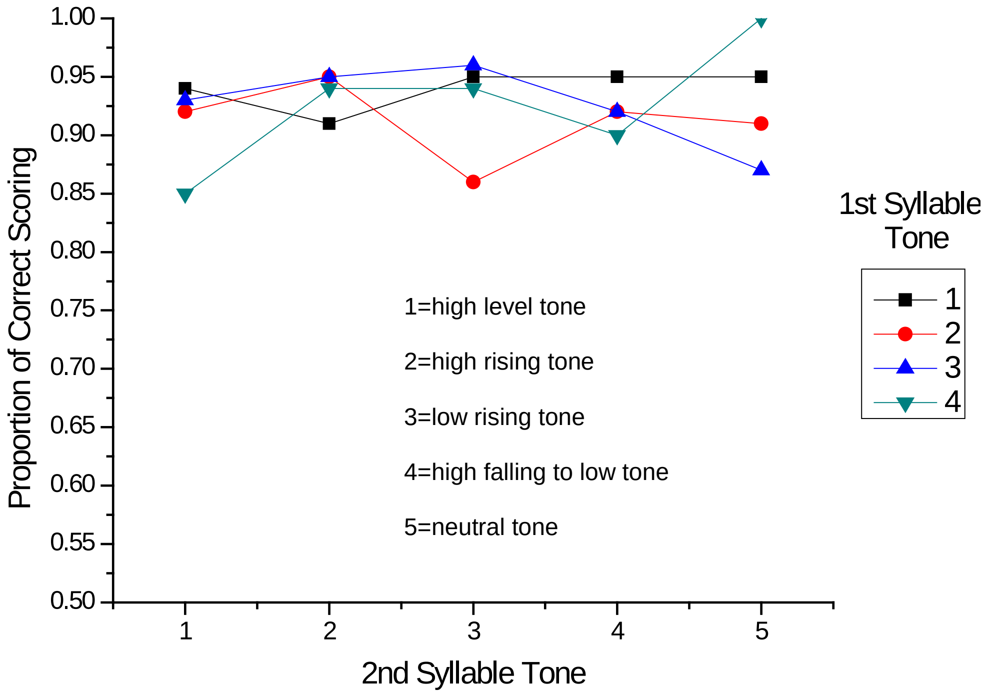

Figure 1.

Mean proportion of correct scoring for bisyllabic Taiwan Mandarin words scored by all participants (novice and experienced) as a function of syllabic tone.

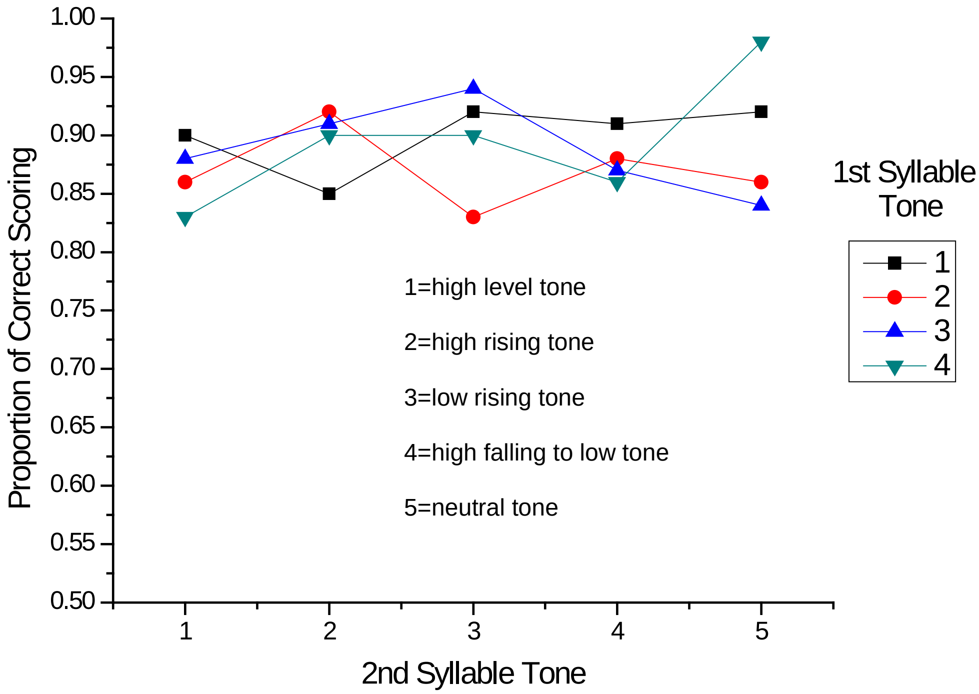

Figure 2.

Mean proportion of correct scoring for bisyllabic Taiwan Mandarin words scored by novice participants as a function of syllabic tone.

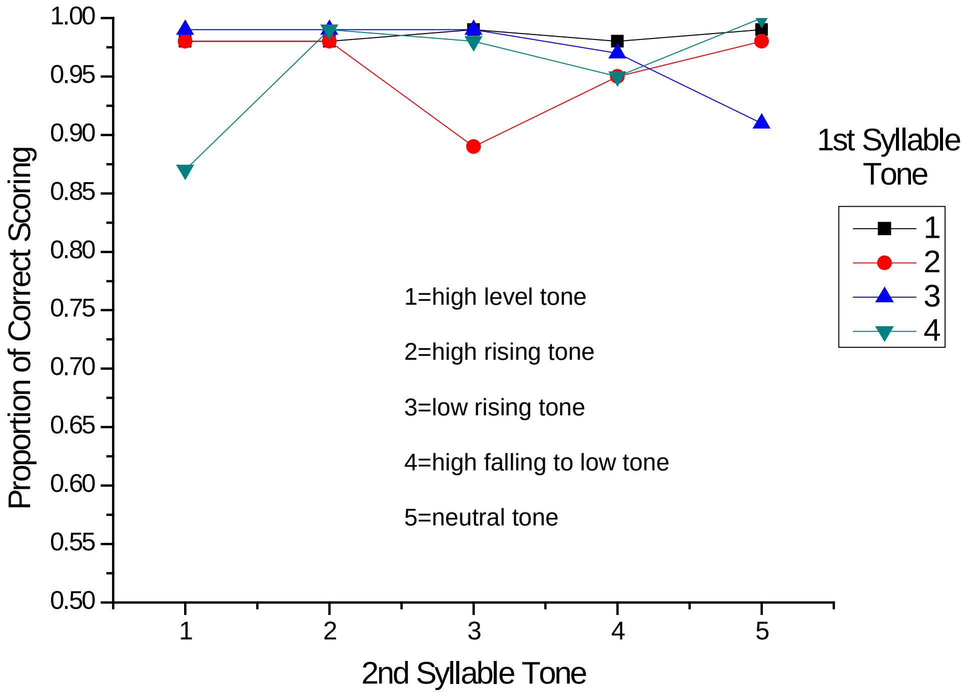

Figure 3.

Mean proportion of correct scoring for bisyllabic Taiwan Mandarin words scored by experienced participants as a function of syllabic tone.

Discussion

The purpose of the present study was to investigate the ability of English-speakers to accurately score WRS tests in Taiwan Mandarin. Two groups of native English-speaking judges, one with no experience with Mandarin and a second with at least a two to three year college-level or equivalent knowledge of Taiwan Mandarin, listened to and scored previously recorded WRS tests given to native Taiwan Mandarin-speaking participants. The results of this study found that there was a significant difference between the ability of the two groups of judges to accurately score WRS tests in Taiwan Mandarin. The novice judges scored the WRS tests with 88.8% accuracy, whereas the experienced judges scored the WRS tests with 97.0% accuracy. These results differ from those found by Cokely and Yager (1993), where no significant difference was found between the accuracy of the WRS scores given by judges with and without knowledge of Spanish. This may be due to differences found between tonal and non tonal languages, and Mandarin, as a tonal language, has long been a source of difficulty for English-speakers (White, 1981).

While there was considerable variability in the accuracy of scoring the various tone combinations among the novice judges, certain tone combinations were found to be more challenging for the experienced judges to score correctly. Specifically, bisyllabic words with the tone combinations of 4 – 1, 2 – 3, and 3 – 5 were the most difficult to score correctly, as presented above in Figure 3. It is possible that the experienced judges' difficulty in hearing some of the different tone combinations came from only hearing the words in isolation, instead of in the context of a sentence. It is uncertain what exactly either of the groups of English judges were listening for when determining the correctness of a response (such as tone or phoneme differences). The answers to these questions, however, are beyond the scope of the present investigation. However, many different studies have shown that the main source of difficulty in learning Mandarin tones has been attributed to interference with the suprasegmental properties of the English language (Wang et al., 1999). Tonal languages, such as Mandarin, use pitch variations as lexical markers, where this is not the case in nontonal languages, such as English (Lee et al., 1996). White (1981) also explains that differences in stress between English and Mandarin may cause confusion when trying to perceive tones. In English, high pitch corresponds with stress. As a result, an English speaker may hear the high pitched Tones 1 and 4 as stressed. Conversely, the stressed Tone 3 in Mandarin, with a lower fundamental frequency, would not be detected as stressed by an English-speaker (White, 1981). Sanders et al. (2002) found that non-native speakers of a language, regardless of familiarity with the language, are not able to use syntactic or stress pattern information to segment sounds as accurately as a native speaker of that language. However, they did find that non-native speakers familiar with a given language benefit from the presence of lexical information to segment sounds to a similar extent as native-speakers. As the tones are lexical information, this corresponds favorably with the present study, as the experienced judges, who are late-learners of Mandarin, were able to identify tone combinations more accurately than the novice judges.

A study done by Shen (1989) found that the rate of errors in the production of Tones 4 and 1 by learners of Mandarin was significantly greater than those of the other three tones, with the number of errors being the greatest for Tone 4. The reason for the errors in the production of Tones 4 and 1 was because of an error in tonal register, where the register of the voice of the participant was lower than it needed to be to produce the tone (Shen, 1989). However, Miracle (1989) did not find any significant differences in the number of errors in the production of any of the tones made by students of Mandarin. Broselow et al. (1987) found that in terms of tone perception, Tone 4 was the most difficult to perceive correctly when it was in a non-final position of a word. This difficulty in perceiving Tone 4 in a non-final position compares favorably with the results of the present study as shown by tone combination 4 – 1 being one of the more difficult combinations to identify correctly. In terms of discriminating between tones, Shen and Lin (1991) found that, due to similar fundamental frequency contours, discrimination between Tones 2 and 3 is quite difficult. It was noted that this difficulty in discrimination occurs for both native Mandarin-speakers as well as learners of Mandarin. While Tones 2 and 3 may be easier to learn than Tones 1 and 4, the pair of Tones 2 and 3 is said to be the most confusing pair for an English learner of Mandarin (Wang et al., 1999). The difficulty in identifying the pair of Tones 2 and 3 also compares to the present study, as tone combination 2 – 3 was one of the three most difficult tone combinations to identify accurately for both novice and experienced speakers of Mandarin. The difficulty in identifying the tone combination 2 – 3 may also be due to the occurrence of tone sandhi, which is a phenomenon where Tone 3 is turned into Tone 2 when immediately followed by Tone 3 (Cheng, 1987). It is possible, though uncertain, if tone sandhi played a role in the difficulty of the experienced judges in correctly identifying words with the tone combination of 2 – 3.

Wang et al. (1999) conducted a study where perceptual tone training was given to English-speakers who had taken one or two semesters of Mandarin Chinese at Cornell University. They found that the participants were able to increase their ability to correctly perceive each of the tones by 21%, and maintained that same increase in ability when presented with new stimuli and talkers. The participants were also able to retain this improvement of 21% over a six month period (Wang et al., 1999).

In terms of clinical application, it is uncertain how significant of a role the interaction between tones would take. While the tone combination of 4 – 1 was scored the least accurately of all of the possible tone combinations, words with this combination were only heard by each judge 10 times out of a total of 800 words. Tone combination 2 – 3 was heard by each judge 23 times, and the tone combination of 3 – 5 was heard only 13 times. It is interesting to note that words with the tone combination that was seen as the easiest for both groups of judges to score accurately, 4 – 5, were only heard by each judge three times.

In conclusion, the purpose of the present study was to determine if the Spanish results found by Cokely and Yager (1993) would hold true for a language whose sound inventory was less similar to English than Spanish, such as Mandarin. It was found that there was a significant difference between an individual's knowledge of Mandarin and one's ability to accurately score a WRS test in Taiwan Mandarin. This differs from the Spanish results of Cokely and Yager, which found that an individual's knowledge of Spanish did not make a significant difference in one's ability to accurately score WRS tests in Spanish. Possible future research could include determining if perceptual tone training would increase a judge's ability to accurately score WRS tests in Taiwan Mandarin. It could also be determined if it would make a difference in scoring ability to simply bring to the judge's attention that there are three different tone combinations that appear to be more difficult to score accurately. Possible future research could also include investigating what types of cues the judges pay the most attention to while scoring the words, such as tone or phoneme differences (and if that differs for novice and experienced judges). This could include looking at the phonemic inventories of the two languages to see how the phonemic differences between English and Mandarin played a role in the ability of the judges to accurately score the WRS tests. Another possible future study could investigate that amount that a non

Mandarin-speaking audiologist could gain with a one-hour tutorial about Mandarin tones. Future research could also include comparing the results of this study to the ability of native Taiwan Mandarin speakers to scoring WRS tests, especially to investigate if there were any differences between the accuracy of scores obtained from the native Taiwan Mandarin speakers and the experienced judges. Since Mandarin is a tonal language, future research may also include investigating the ability of an individual who speaks a tonal language other than Mandarin to accurately score a WRS test in Mandarin. Future research could be conducted to determine if English-speaking judges would be able to accurately score WRS tests in a nontonal language whose sound inventories also differ greatly from English, such as Japanese or Korean.

References

Aleksandrovsky, I. V., McCullough, J. K., & Wilson, R. H. (1998). Development of suprathreshold word recognition test for Russian-speaking patients. *Journal of the American Academy of Audiology, 9*(6), 417-425.

American National Standards Institute. (1999). *Maximum permissible ambient noise levels for audiometric test rooms* (ANSI S3.1-1999). New York: ANSI.

American National Standards Institute. (2004). S*pecification for audiometers* (ANSI S3.6-2004). New York: ANSI.

American Speech-Language Hearing Association. (1988). Guidelines for determining threshold level for speech. *ASHA, 30*(3), 85-89.

American Speech-Language Hearing Association. (1990). Guidelines for screening for hearing impairments and middle-ear disorders. *ASHA, 32*(Supplement 2), 17-24.

American Speech-Language Hearing Association. (2000). 2000 *Omnibus survey: Practice trends in audiology*. Rockville, MD: ASHA.

Bell, T. S., & Wilson, R. H. (2001). Sentence recognition materials based on frequency of word use and lexical confusability. *Journal of the American Academy of Audiology, 12*(10), 514-522.

Best, C. T. (1989). The emergence of native-language phonological influences in infants: A perceptual assimilation model. In J. C. Goodman & H. C. Nusbaum (Eds.), *The development of speech perception: The transition from speech sounds to spoken words* (pp. 167-224). USA: Massachusetts Institute of Technology.

Beverly-Ducker, K. (2003, April). Multicultural issues in audiology. *Perspectives on Hearing and Hearing Disorders in Childhood, 13*, 12-15. doi:10.1044/hhdc13.1.12

Bilger, R. C., Matthies, M. L., Meyer, T. A., & Griffiths, S. K. (1998). Psychometric equivalence of recorded spondaic words as test items. *Journal of Speech, Language, and Hearing Research, 41*(3), 516-526.

Bradlow, A. R., & Pisoni, D. B. (1999). Recognition of spoken words by native and non-native listeners: Talker-, listener-, and item-related factors. *Journal of the Acoustical Society of America, 106*(4), 2074-2085. doi:10.1121/1.427952

Brandy, W. T. (2002). Speech Audiometry. In J. Katz (Ed.), *Handbook of clinical audiology.* (5th ed., pp. 96-110). Baltimore, MD: Williams & Wilkins.

Broselow, E., Hurtig, R. R., & Ringen, C. (1987). The perception of second language prosody. In G. Ioup & S. H. Weinberger (Eds.), *Interlanguage phonology: The acquisition of a second language sound system* (pp. 350-361). Cambridge, MA: Newbury House Publishers.

Campbell, R. A. (1965). Discrimination test word difficulty. *Journal of Speech and Hearing Research, 8*(1), 13-22.

Carhart, R. (1951a). Basic principles of speech audiometry. *Acta Oto-Laryngologica, 40*(1-2), 62-71. doi:10.3109/00016485109138908

Carhart, R. (1951b). Instruments and materials for speech audiometry. *Acta Oto-Laryngologica, 40*(5-6), 313-323. doi:10.3109/00016485109123646

Carhart, R. (1952). Speech audiometry in clinical evaluation. *Acta Oto-Laryngologica, 41*(1-2), 18-42. doi:10.3109/00016485209124357

Carhart, R. (1965). Problems in the measurement of speech discrimination. *Archives of Otolaryngology, 82*(3), 253-260.

Cheng, L. L. S. (1987). On the prosodic hierarchy and tone sandhi in Mandarin. In P. Avery (Ed.), *Toronto Working Papers in Linguistics 7* (pp. 24-52). University of Toronto: Linguistic Graduate Course Union.

Cokely, J. A., & Yager, C. R. (1993). Scoring Spanish word-recognition measures. *Ear and Hearing, 14*(6), 395-400. doi:10.1097/00003446-199312000-00004

Comstock, C. L., & Martin, F. N. (1984). A children's Spanish word discrimination test for non-Spanish-speaking clinicians. *Ear and Hearing, 5*(3), 166-170. doi:10.1097/00003446-198405000-00008

Conklin, B. K. (2007). *Psychometrically equivalent Cantonese bisyllabic word recognition materials spoken by male and female talkers.* Unpublished master's thesis, Brigham Young University, Provo, UT.

Crandell, C. C., & Smaldino, J. J. (1996). Speech perception in noise by children for whom English is a second language. *American Journal of Audiology, 5*(3), 47-51.

Danhauer, J. L., Crawford, S., & Edgerton, B. J. (1984). English, Spanish, and bilingual speakers' performance on a nonsense syllable test (NST) of speech sound discrimination. *Journal of Speech and Hearing Disorders, 49*(2), 164-168.

De Jonge, C. E. (1995). Interlanguage phonology: Perception and production. *Dissertation Abstracts International, 56A*(07), 2658. (UMI No. 9539993)

Dennis, J. M., & Neely, J. G. (1991). Basic hearing tests. *Otolaryngologic Clinics of North America, 24*(2), 253-276.

Dirks, D. D. (2004). Neighborhood activation in spoken word recognition. *Seminars in Hearing, 25*(1), 17-24. doi:10.1055/s-2004-823044

Dirks, D. D., Takayanagi, S., & Moshfegh, A. (2001). Effects of lexical factors on word recognition among normal-hearing and hearing-impaired listeners. *Journal of the American Academy of Audiology, 12*(5), 233-244.

Egan, J. P. (1948). Articulation testing methods. *The Laryngoscope, 58*(9), 955-991. doi:10.1288/00005537-194809000-00002

Egan, J. J. (1979). Basic aspects of speech audiometry. *Ear, Nose, and Throat Journal, 58*(5), 190-193.

Eckman, F. R. (2004). From phonemic differences to constraint rankings. *Studies in Second Language Acquisition, 26*(4), 513-549. doi:10.1017/S027226310404001X

Epstein, A. (1978). Speech audiometry. *Otolaryngologic Clinics of North America, 11*(3), 667-676.

Ferguson, S. H., Jongman, A., Sereno, J. A., & Keum, K. A. (2010). Intelligibility of foreign-accented speech for older adults with and without hearing loss. *Journal of the American Academy of Audiology, 21*(3), 153-162.

Flege, J. E. (1995). Second language speech learning: Theory, findings, and problems. In W. Strange (Ed.), *Speech perception and linguistic experience: Issues in cross-language research* (pp. 233-272). Baltimore, MD: York Press.

Garcia de las Bayonas, M. (2004). The acquisition of vowels in Spanish and English as a second language. *Dissertation Abstracts International, 65A*(07), 2587. (UMI No. 3141596)

Garvey, W. D. (1953). The intelligibility of speeded speech. *Journal of Experimental Psychology, 45*(2), 102-108. doi:10.1037/h0054381

Guion, S. G., Flege, J. E., Akahane-Yamada, R., & Pruitt, J. C. (2000). An investigation of current models of second language speech perception: The case of Japanese adults' perception of English consonants. *Journal of the Acoustical Society of America, 107*(5), 2711-2724. doi:10.1121/1.428657

Hall, J. W. (1983). Diagnostic applications of speech audiometry. *Seminars in Hearing, 4*(3), 179-204. doi:10.1055/s-0028-1091425

Harris, R. W., Kim, E., & Eggett, D. L. (2003a). Psychometrically equivalent Korean bisyllabic words spoken by male and female talkers. *Korean Journal of Communication Disorders, 8*(1), 217-243.

Harris, R. W., Kim, E., & Eggett, D. L. (2003b). Psychometrically equivalent Korean monosyllabic speech discrimination materials spoken by male and female talkers. *Korean Journal of Communication Disorders, 8*(1), 244-270.

Harris, R. W., Nielson, W. S., McPherson, D. L., Skarzynski, H., & Eggett, D. L. (2004a). Psychometrically equivalent Polish bisyllabic words spoken by male and female talkers. *Audiofonologia, 25,* 1-15.

Harris, R. W., Nielson, W. S., McPherson, D. L., Skarzynski, H., & Eggett, D. L. (2004b). Psychometrically equivalent Polish monosyllabic word recognition materials spoken by male and female talkers. *Audiofonologia, 25,* 16-31.

Harris, R. W., Nissen, S. L., Pola, M. G., McPherson, D. L., Tavartkiladze, G. A., & Eggett, D. L. (2007). Psychometrically equivalent Russian speech audiometry materials by male and female talkers. *International Journal of Audiology, 46*(1), 47-66. doi:10.1080/14992020601058117

Hirsh, I. J., Davis, H., Silverman, S. R., Reynolds, E. G., Eldert, E., & Benson, R. W. (1952). Development of materials for speech audiometry. *Journal of Speech and Hearing Disorders, 17*(3), 321-337.

Hood, J. D., & Poole, J. P. (1980). Influence of the speaker and other factors affecting speech intelligibility. *Audiology, 19*(5), 434-455. doi:10.3109/00206098009070077

Hoople, G. D. (1952). Practical demonstration of speech audiometry. *Transactions – American Academy of Ophthalmology and Otolaryngology, 56*(1), 76-77.

Iliadou, V., Fourakis, M., Vakalos, A., Hawks, J. W., & Kaprinis, G. (2006). Bi-syllabic, modern Greek word lists for use in word recognition tests. *International Journal of Audiology, 45*(2), 74-82. doi:10.1080/14992020500376529

Katzner, K. (1995). *The languages of the world.* Great Britain: The Guernsey Press Co. LTD.

Keith, R. W., Katbamna, B., Tawfik, S., & Smolak, L. H. (1987). The effect of linguistic background on staggered spondaic word and dichotic consonant vowel scores. *British Journal of Audiology, 21*(1), 21-26. doi:10.3109/03005368709077770

Kerr, A. G., & Smyth, G. D. (1972). Routine speech discrimination tests. *The Journal of Laryngology and Otology, 86*(1), 33-41.

Kim, M. N. (2007). *Psychometrically equivalent trisyllabic words for speech reception threshold testing in Cantonese.* Unpublished master's thesis, Brigham Young University, Provo, UT.

Kuhl, P. K. (1993). Early linguistic experience and phonetic perception: Implications for theories of developmental speech perception. *Journal of Phonetics, 21*(1/2), 125-139.

Kuhl, P. K., & Iverson, P. (1995). Linguistic experience and the "perceptual magnet effect." In W. Strange (Ed.), *Speech perception and linguistic experience: Issues in cross-language research* (pp. 121-154). Baltimore, MD: York Press.

Kuhl, P. K., Williams, K. A., Lacerda, F., Stevens, K. N., & Lindblom, B. (1992). Linguistic experience alters phonetic perception in infants by 6 months of age. *Science, 255*(5044), 606-608. doi:10.1126/science.1736364

Lau, C. C., & So, K. W. (1988). Material for Cantonese speech audiometry constructed by appropriate phonetic principles. *British Journal of Audiology, 22*(4), 297-304. doi:10.3109/03005368809076467

Lee, Y., Vakoch, D. A., & Wurm, L. E. (1996). Tone perception in Cantonese and Mandarin: A cross-linguistic comparison. *Journal of Psycholinguistic Research, 25*(5), 527-542. doi:10.1007/BF01758181

Lehiste, I., & Peterson, G. E. (1959). Linguistic considerations in the study of speech intelligibility. *The Journal of the Acoustical Society of America, 31*(3), 280-286. doi:10.1121/1.1907713

Li, C. N., & Thompson, S. A. (1987). Chinese. In B. Combrie (Ed.), *The world's major languages* (811-833). Great Britain: Croom Helm Ltd.

Lively, S. E., Pisoni, D. B., Yamada, R. A., Tohkura, Y., & Yamada, T. (1994). Training Japanese listeners to identify English /r/ and /l/. III. Long-term retention of new phonetic categories. *The Journal of the Acoustical Society of America, 96*(4), 2076-2087. doi:10.1121/1.410149

Logan, J. S., Lively, S. E., & Pisoni, D. B. (1991). Training Japanese listeners to identify English /r/ and /l/: A first report. *The Journal of the Acoustical Society of America, 89*(2), 874-886. doi:10.1121/1.1894649

Martin, F. N., & Clark, J. G. (2009). *Introduction to Audiology 10th Edition*. Boston, MA: Allyn & Bacon.

Martin, F. N., & Hart, D. B. (1978). Measurement of speech thresholds of Spanish-speaking children by non-Spanish speaking clinicians. *Journal of Speech and Hearing Disorders, 43*(2), 255-262.

Martin, F. N., & Stauffer, M. L. (1975). A modification of the Tillman-Olsen method for obtaining the speech reception threshold. *Journal of Speech and Hearing Disorders, 40*(1), 25-28.

McCullough, J. A., & Wilson, R. H. (2001). Performance on a Spanish picture identification task using a multimedia format. *Journal of American Academic Audiology, 12*(5), 254-260.

McClaskey, C. L., Pisoni, D. B., & Carrell, T. D. (1983). Transfer of training of a new linguistic contrast in voicing. *Perception and Psychophysics, 34*(4), 323-330.

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

Nelson, D. A., & Chaiklin, J. B. (1970). Writedown versus talkback scoring and scoring in speech discrimination testing. *Journal of Speech and Hearing Research, 13*(3), 645-654.

Nissen, S. L., Harris, R. W., & Dukes, A. (2008). Word recognition materials for native speakers of Taiwan Mandarin. *The American Journal of Audiology, 17*(1), 68-79. doi:10.1044/1059-0889(2008/008)

Nissen, S. L., Harris, R. W., & Slade, K. B. (2007). Development of speech recognition threshold materials for speakers of Taiwan Mandarin. *International Journal of Audiology, 46*(8), 449-458. doi:10.1080/14992020701361296

Nissen, S. L., Harris, R. W., Jennings, L. J., Eggett, D. L., & Buck, H. (2005a). Psychometrically equivalent Mandarin bisyllabic speech discrimination materials spoken by male and female talkers. *International Journal of Audiology, 44*(7), 379-390. doi:10.1080/14992020500147615

Nissen, S. L., Harris, R. W., Jennings, L. J., Eggett, D. L., & Buck, H. (2005b). Psychometrically equivalent trisyllabic words for speech reception threshold testing in Mandarin. *International Journal of Audiology, 44*(7), 391-399. doi:10.1080/14992020500147672

Pisoni, D. B., Aslin, R. N., Perey, A. J., & Hennessy, B. L. (1982). Some effects of laboratory training on identification and discrimination of voicing contrasts in stop consonants. *Journal of Experimental Psychology: Human Perception and Performance, 8*(2), 297-314. doi:10.1037/0096-1523.8.2.297

Ramkissoon, I. (2001). Speech recognition thresholds for mulitilingual populations. *Communication Disorders Quarterly, 22*(3), 158-162. doi:10.1177/152574010102200305

Ramkissoon, I., Proctor, A., Lansing, C. R., & Bilger, R. C. (2002). Digit speech recognition thresholds (SRT) for non-native speakers of English. *American Journal of Audiology, 11*(1), 23-28. doi:10.1044/1059-0889(2002/005)

Ramkissoon, I., & Khan, F. (2003, February 18). Serving multilingual clients with hearing loss: How linguistic diversity affects audiologic management. *The ASHA Leader*.

Roup, C. M., Wiley, T. L., Safady, S. H., & Stoppenbach, D. T. (1998). Tympanometric screening norms for adults. *American Journal of Audiology, 7*(2), 55-60. doi:10.1044/1059-0889(1998/014)

Sanders, L. D., Neville, H. J., & Woldorff, M. G. (2002). Speech segmentation by native and non-native speakers: The use of lexical, syntactic, and stress-pattern cues. *Journal of Speech, Language, and Hearing Research, 45*(3), 519-530. doi:10.1044/1092-4388(2002/041)

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

Shen, X. S., & Lin, M. (1991). A perceptual study of Mandarin tones 2 and 3. *Language and Speech, 34*(2), 145-156.

Singh, S., & Black, J. W. (1966). Study of twenty-six intervocalic consonants as spoken and recognized by four language groups. *The Journal of the Acoustical Society of America, 39*(2), 372-387. doi:10.1121/1.1909899

Spitzer, J. B. (1980). The development of a picture speech reception threshold test in Spanish for use with urban U.S. residents of Hispanic background. *Journal of Communication Disorders, 13*(2), 147-151. doi:10.1016/0021-9924(80)90031-3

Studio Audio, & Video Ltd. (2007). SADiE disk editor software (Version 5.6).

Thibodeau, L. M. (2007). Speech audiometry. In R. J. Roeser, M. Valente, & H. Hosford (Eds.), *Audiology Diagnosis* (pp. 288-313). New York: Thieme.

Tillman, T. W., & Carhart, R. (1966). *An expanded test for speech discrimination utilizing CNC monosyllabic words: Northwestern University auditory test no. 6.* San Antonio, TX: USAF School of Aerospace Medicine, Brooks Air Force Base.

Vitevitch, M. S., & Rodriguez, E. (2005). Neighborhood density effects in spoken word recognition in Spanish. *Journal of Multilingual Communication Disorders, 3*(1), 64-73. doi:10.1080/14769670400027332

Von Hapsburg, D., Champlin, C. A., & Shetty, S. R. (2004). Reception thresholds for sentences in bilingual (Spanish/English) and monolingual (English) listeners. *Journal of the American Academy of Audiology, 15*(1), 88-98. doi:10.3766/jaaa.15.1.9

Wang, S., Mannell, R., Newall, P., Zhang, H., & Han, D. (2007). Development and evaluation of Mandarin disyllabic materials for speech audiometry in China. *International Journal of Audiology, 46*(12), 719-731. doi:10.1080/14992020701558511

Wang, Y., Spence, M. M., Jongman, A., & Sereno, J. A. (1999). Training American listeners to perceive Mandarin tones. *The Journal of the Acoustical Society of America, 106*(6), 3649-3658. doi:10.1121/1.428217

Weisleder, P., & Hodgson, W. R. (1989). Evaluation of four Spanish word-recognition-ability lists. *Ear and Hearing, 10*(6), 387-393. doi:10.1097/00003446-198912000-00012

Werker, J. F., Gilbert, J. H. V., Humphrey, K., & Tees, R. C. (1981). Developmental aspects of cross-language speech perception. *Child Development, 52*(1), 349-355. doi:10.2307/1129249

White, C. M. (1981). Tonal pronunciation errors and interference from English intonation. *Journal of Chinese Language Teachers Association,* 16*(2)*, 27-56.

Wu, S., Yu, Y., Zhang, Y., & Tian, W. (2006). *Chinese link.* Upper Saddle River, NJ: Prentice Hall.

Zubick, H. H., Irizarry, L. M., Rosen, L., Feudo, P., Kelly, J. H., & Strome, M. (1983). Development of speech-audiometric materials for native Spanish-speaking adults. *Audiology, 22*(1), 88-102. doi:10.3109/00206098309072772

Appendix A

Informed Consent

RESEARCH PARTICIPATION FORM

Participant:__________________________________Age:___________

You are asked to participate in a research study sponsored by the Department of Communication Disorders at Brigham Young University, Provo, Utah. The faculty director of this research is Richard W. Harris, Ph.D. Students in the Communication Disorders program may assist in data collection.

This research project is designed to evaluate a word list recorded using improved digital techniques. You will be presented with this list of words at varying levels of intensity. Many will be very soft, but none will be uncomfortably loud to you. You may also be presented with this list of words in the presence of a background noise. The level of this noise will be audible but never uncomfortably loud to you. This testing will require you to listen carefully and repeat what is heard through earphones or loudspeakers. Before listening to the word lists, you will be administered a routine hearing test to determine that your hearing is normal and that you are qualified for this study.

It will take approximately two hours to complete the test. Testing will be broken up into 2 or 3 one hour blocks. Each subject will be required to be present for the entire time, unless prior arrangements are made with the tester. You are free to make inquiries at any time during testing and expect those inquiries to be answered.

As the testing will be carried out in standard clinical conditions, there are no known risks involved. Standard clinical test protocol will be followed to ensure that you will not be exposed to any unduly loud signals.

Names of all subjects will be kept confidential to the investigators involved in the study. Participation in the study is a voluntary service and no payment of monetary reward of any kind is possible or implied.

You are free to withdraw from the study at any time without any penalty, including penalty to future care you may desire to receive from this clinic.

If you complete your participation in this research project you will be paid the amount of $_________ for your participation.

If you have any questions regarding this research project you may contact Dr. Richard W. Harris, 131 TLRB, Brigham Young University, Provo, UT 84602; phone (801) 422-6460 or Dr. Shawn L. Nissen, 138 TLRB, Brigham Young University, Provo, UT 84602, phone (801) 422-5056. If you have any questions regarding your rights as a participant in a research project you may contact Dr. Lane Fischer, Chair of the Institutional Review Board, 340-L MCKB, Brigham Young University, Provo, UT 84602; phone (801) 422-8293, email: lane_fischer@byu.edu.

YES: I agree to participate in the Brigham Young University research study mentioned above. I confirm that I have read the preceding information and disclosure. I hereby give my informed consent for participation as described.

____________________________________ ____________________
Signature of Participant Date

____________________________________ ____________________
Signature of Witness Date

Appendix B

Instructions Sheet

You will hear a recording of some word pairs spoken in Mandarin. The first word was presented to a listener who was requested to repeat the word, and the second word is the listener's response. You are to determine whether the first and second words in each pair are the same.

You will be sitting in a sound booth with a laptop computer that has a spreadsheet of the words. Each word is shown in both the Mandarin character and in a Romanization format. If the first and second word of each pair are the same enter a "+" in the cell next to the word on the spreadsheet. If the words are different, the word was not repeated, or if the listener said "I don't know", score the word as incorrect using a minus "-" sign. If you need to stop or take a break at any time during the experiment please say "Stop" and when you are ready I will resume the experiment where you left off. You will listen to 16 lists of word-pairs, each taking approximately four minutes to complete.

In Mandarin, there are four *tones*, which are indicated in the diacritic markings on the Romanization. These four *tones* are labeled as: 1) high level tone, 2) high rising tone, 3) low rising tone, and 4) high falling to low tone. When you are listening to the pairs of words the Mandarin tone must be taken into consideration. For the word to be scored as correct the word and tone must both be repeated correctly.

Examples of these tones are:

High level tone	yī
High rising tone	yí
Low rising tone	yǐ
High falling to low tone	yì

Do you have any questions?