2008-03-14

The Effect of Regional Dialect on the Validity and Reliability of Word Recognition Scores

Jamie Ann Garlick
Brigham Young University - Provo

THE EFFECT OF REGIONAL DIALECT ON THE VALIDITY AND RELIABILITY

OF WORD RECOGNITION SCORES

by

Jamie A. Garlick

A thesis submitted to the faculty of

Brigham Young University

in partial fulfillment of the requirements for the degree of

Master of Science

Department of Communication Disorders

Brigham Young University

April 2008

BRIGHAM YOUNG UNIVERSITY

GRADUATE COMMITTEE APPROVAL

of a thesis submitted by

Jamie A. Garlick

This thesis has been read by each member of the following graduate committee and by majority vote has been found to be satisfactory.

Date	Shawn L. Nissen, Chair

Date	Richard W. Harris

Date	Ron W. Channell

BRIGHAM YOUNG UNIVERSITY

As chair of the candidate's graduate committee, I have read the thesis of Jamie Ann Garlick in its final form and have found that (1) its format, citations, and bibliographical style are consistent and acceptable and fulfill university and department style requirements; (2) its illustrative materials including figures, tables, and charts are in place; and (3) the final manuscript is satisfactory to the graduate committee and is ready for submission to the university library.

_____   _____
Date   Shawn L. Nissen
                                        Chair, Graduate Committee

Accepted for the Department

_____   _____
Date   Ron W. Channell
                                        Graduate Coordinator

Accepted for the College

_____   _____
Date   K. Richard Young
                                        Dean, David O. McKay School of Education

ABSTRACT

THE EFFECT OF REGIONAL DIALECT ON THE VALIDITY AND RELIABILITY

OF WORD RECOGNITION SCORES

Jamie A. Garlick

Department of Communication Disorders

Master of Science

The purpose of this study was to examine the effect of talker and listener dialect on the validity and reliability of word recognition scores from two sets of Mandarin speech audiometry materials. Four lists of bisyllabic words in Mainland Mandarin and Taiwan Mandarin dialects were administered to 16 participants of each dialect with normal hearing across two test sessions. The performance on materials presented in the native dialect was compared to performance on non-native dialect assessment to determine validity and reliability of test materials. Statistical analysis indicated significant differences between word recognition scores across test sessions, talker and listener dialect, and among lists. However it is unclear if such differences constitute clinically significant differences.

ACKNOWLEDGMENTS

I would like to acknowledge my mother for her endless encouragement and example in obtaining higher education. I am grateful to Nathan Richardson for his help and good natured sense of humor during the data collection and writing process. Accolades to my husband for his enduring patience and support. And finally, I express great appreciation to my graduate committee for their collaborative guidance and support in completing this project.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Introduction

To discover a participant's true abilities during speech audiometry testing, the test needs to be presented in the participant's native language. However, due to a lack of linguistically-diverse speech audiometry materials, many audiologists in the United States use English materials to test non-English speaking populations (Ramkissoon, 2001). Thus, there is clearly a need for audiometric materials that have been developed in a variety of languages. However, it remains unclear if such materials not only need to match a speaker's native language, but also their regional dialect.

Previous research has shown conflicting evidence on the effect of speech audiometry materials presented in a regional dialect other than that of the participants. Some researchers found little or no difference in the results of speech audiometry testing when using materials from a non-regional dialect (Martin & Hart, 1978; Schneider, 1992). Other researchers have concluded that the speaker's dialect may affect their performance on speech audiometry evaluations (Ramkissoon, 2001; von Hapsburg & Pena, 2002; Weisleder & Hodgson, 1989). These latter studies showed that speech audiometry scores declined when materials were not presented in the native regional dialect. These findings support the claim that the linguistic nature of the test materials can impact the results of audiometric testing for individuals with a native dialect other than that presented in testing materials. If participants are unfamiliar with the phonological forms of the words that are presented to them, they will likely not perform to their capacity (Danhauer, Crawford, & Edgerton, 1984; von Hapsburg & Pena, 2002).

The aim of this study was to examine the validity and reliability of using previously developed Mandarin speech audiometry materials to evaluate the word recognition abilities of regional and non-regional speakers of the presented dialects.

Specifically, this study will evaluate the use of Taiwan Mandarin and Mainland Mandarin materials with participants from Taiwan and Mainland China. In addition, this study will evaluate if both types of materials can be effectively administered by an audiologist or interpreter that does not speak the regional dialect of the listener.

Literature Review

Hearing loss is among the most prevalent global health concerns today. In 2004, the World Health Organization estimated the prevalence of worldwide hearing loss to be approximately 250 million people (Smith, 2004). The global prevalence of hearing loss is a universal communication problem. When the intelligibility of a conversation involving one or more persons with hearing loss is compromised, all participants of the conversation are influenced.

If a person has a moderate or severe hearing loss, their ability to participate in social interaction is significantly limited. At its most debilitating, hearing loss can interfere with social participation, physical independence, and economic self-sufficiency (Weinstein, 1996). Beyond the frustrating effect these factors play in the lives of those with the impairment, acquired hearing loss often causes profitable workers to lose their employment because of an inability to maintain their past performance levels. Due to the significant impact that hearing loss can have on an individual's quality of life, it is important to diagnose a possible hearing loss in a timely, reliable, and valid manner.

Methods of Evaluating Hearing Acuity

A comprehensive hearing evaluation often involves several types of diagnostic tests to evaluate an individual's hearing acuity. Diagnostic methods most commonly used are pure-tone testing, tympanometry, and speech audiometry. In pure-tone testing, hearing is measured by presenting a series of sinusoidal tones of different frequencies at different intensity levels, commonly ranging from 125 to 8000 Hz. Because the speech frequencies generally fall between 500 and 2000 Hz, the thresholds obtained from tested pure-tone frequencies of 500, 1000, and 2000 Hz are often averaged to determine the

pure tone average (PTA). This type of testing is used to establish an individual's hearing threshold or the lowest level at which a person can hear the presented stimulus at least 50% of the time. A threshold of 20 dB HL or lower indicates normal hearing sensitivity; a threshold of 30 dB HL is indicative of a mild hearing loss, 45 dB HL signifies a moderate loss, 60 dB HL a severe hearing loss, and a threshold of 90 dB HL or above implies a profound hearing loss (Brandy, 2002).

Before digital recording became widely available, it was common practice to use monitored live voice to present stimuli during speech audiometry testing. However, each person's voice is perceived somewhat differently due to acoustic and physiological factors, such as voice quality, loudness, speech rate, breath support, and articulation (Brandy, 2002). Thus listeners are presented with a varied set of test materials across multiple hearing evaluations, thereby resulting in a low degree of test-retest reliability (Resnick, 1962). Speech audiometry testing became a more standardized practice with the introduction of recorded materials, thus increasing the reliability of the materials.

While using recorded analog materials is clearly advantageous over using monitored live voice for standardization, the introduction of digital materials has further improved the validity and reliability of speech audiometry testing. Advantages of using digital rather than analog materials include increased dynamic range and channel separation, the reduction in harmonic distortion, an improved signal-to-noise ratio, the absence of the wow and flutter commonly associated with tape recordings, and a dramatic decrease in usage degradation (Ingram, Bunta, & Ingram, 2004; Nissen, Harris, Jennings, Eggett, & Buck, 2005a). In addition, through the use of custom software individual

speaker tracks or words can be efficiently randomized prior to presentation (Nissen et al., 2005b).

Types of Speech Audiometry

Speech audiometry is another important component in estimating one's hearing abilities (Weisleder & Hodgson, 1989). Ramkissoon (2001) outlines three reasons why speech audiometry is an essential part of hearing assessment: (a) the speech stimuli used in speech audiometry are representative of daily speech sounds, (b) speech comprehension is necessary for communication and participation in society, and (c) words used in hearing assessment are familiar to most participants and therefore promote high face validity.

Two commonly used types of speech audiometry are word recognition (WR) and speech reception threshold (SRT) testing. Testing for WR and SRT are performed by presenting words to the participant at differing presentation levels and recording their ability to accurately identify the stimuli. The scores obtained from these measures are indicative of an individual's ability to perceive speech, given that the measures are collected in a manner that ensures they are reliable and valid.

The lowest intensity level at which a participant can identify spoken words with 50% accuracy is measured through SRT. Threshold levels determined through PTA should be consistent with SRT scores. The SRT is generally measured by presenting bisyllabic words which are spondaic in stress pattern, whereby both syllables are given relatively equal emphasis during production.

According to Brandy (2002), there are three purposes that SRT tests fulfill. The first purpose is to validate threshold levels obtained from pure-tone testing. Secondly, SRT testing estimates speech hearing sensitivity. Thirdly, this type of hearing testing can

provide a baseline to determine appropriate presentation levels for subsequent WR

testing. Listeners are familiarized with the list of SRT words at a suprathreshold intensity

level prior to testing to aid in establishing a consistent response pattern from the listener.

The ear with the better PTA is generally tested first. The listener hears the word and then

repeats the word orally (Epstein, 1978).

An individual's ability to understand speech when it is presented at a comfortable

loudness level that surpasses a possible loss in hearing sensitivity is measured through

WR testing. The main objective of WR testing is to estimate a person's ability to hear and

comprehend continuous dialogue (Epstein, 1978). In order to simulate continuous

dialogue and alert the listener that the next word is about to be presented, some

audiologists use carrier phrases, such as *say the word* before each word is presented.

However, there is little evidence that the presence or absence of a carrier phrases effects

threshold measurements (Brandy, 2002). In WR testing listeners are not familiarized with

the words before repeating the words that they hear. A full list of 50 words or a half-list

of 25 words is presented at the same suprathreshold level. Word recognition scores are

calculated separately for each ear, and are reported as a percentage of the number of

words correctly identified for each list or half-list. This information is useful for the

audiologist to determine how well the listener can recognize words at a comfortable

listening level.

The Reliability and Validity of Speech Audiometry Materials

Measurement of reliability in speech audiometry materials. The reliability of WR

test materials can be established through an evaluation of the relationship between the

predicted and actual results obtained from using the materials. Test materials should be

reliable in yielding similar results across contexts, independent of clinic or test

administrator (Ostergard, 1983). A possible method of establishing reliability is through a test-retest procedure. In this type of procedure, participants are tested multiple times with the same assessment battery, after which the scores for each testing session are compared. It is recommended that test materials have a test-retest reliability of at least .90 (Skarakis-Doyle & Mallet, 1991). It is also important to design materials that also exhibit internal reliability.

Factors of construct validity. To increase the validity of WR testing some audiologists prefer to use phonetically balanced (PB) word lists, with the frequency of occurrence of consonants across the list being approximately proportional to the frequency of their usage in the language from which they are developed (Brandy, 2002). However, several previous research studies have questioned the validity of using PB word lists (Giolas & Epstein, 1963; Martin, Champlin, & Perez, 2000). Giolas and Epstein (1963) found that the WR lists that were not PB were more consistent with the subjects' comprehension of conversational speech. The study by Martin et al. found that WR scores obtained from randomly selected words were not significantly different from the scores obtained from PB word lists. Both studies provide evidence that phonetic balance may be a less important factor to consider when creating word lists.

For WR materials, word lists should be relatively psychometrically equivalent or homogenous with respect to audibility and be familiar to the listener being evaluated. Using words that are unfamiliar to the listener may increase the probability that the WR results will be influenced by the listener's vocabulary knowledge rather than their auditory threshold. Similar concerns exist for materials that are not from a listener's native language or regional dialect.

There is also some disagreement in the field of clinical audiology in regard to the ideal WR list length. Schwarzt, Bess, and Larson (1977), Kruel et al. (1968), Jirsa, Hodgson, and Goetzinger (1975), as well as Schumaler and Rintelmann (1974) argue that a list of 50 test items is needed to obtain a valid measurement of an individual's WR ability. These researchers concluded that a minimum of 50 words was needed to provide enough phonetic variation to represent a speaker's language as a whole. However, Elpern (1961), Egan (1948), Deutsch and Kruger (1971), and Resnick (1962) found that half-lists of 25 words can accurately measure an individual's hearing acuity, as well as the added benefit of reducing testing time and client fatigue.

Linguistic validity of speech audiometry testing in non-regional dialects. Previous studies assessing the impact of dialect on speech audiometry testing have reported contradictory results (Schneider, 1992; Weisleder & Hodgson, 1989). Schneider examined the effect of dialectal variations on speech audiometric assessments. Her goal was to determine the impact of dialect on SRT scores. Schneider took a close look at phonetic differences between regional dialects of Spanish in Mexico and Puerto Rico. Twelve Spanish-dominant typically-developing 6 and 7 year-old children with normal hearing thresholds at 13 dB HL or better at all frequencies in the right ear (left ear thresholds were not reported) participated in the study. These children were born in Puerto Rico and were living in Connecticut at the time of the study. The same twelve words were produced by three different talkers, a Castilian male talker and female talkers from the Caribbean and Mexico. The three sets of materials were presented to the participants in a randomized order and were selected based on their phonetically distinct vowel sequence patterns. Participants were familiarized with the word lists, and were

instructed to respond to the stimuli by pointing to one of four presented pictures that represented the auditory stimuli. Statistical analysis showed no statistically significant differences among the different lists. The data were assumed to be reliable due to their high correlation with the pure tone average of each participant, although concession was given for using the average of the three lowest thresholds of test frequencies rather than the standardized average of thresholds at 500, 1000, and 2000 Hz. This deviation from the standard method was acknowledged to bias the PTA values closer to the SRT values. Schneider (1992) concluded that no statistically significant differences among the lists were found; providing some evidence that the impact of regional dialect on the accuracy of SRT testing may be minimal for Spanish-speaking children.

Weisleder and Hodgson (1989) also performed an experiment with the purpose to evaluate the possible effect of regional dialect on the accuracy of speech audiometry testing in Spanish. The authors delineated some of the pitfalls of past attempts to develop dialect-specific speech audiometry materials, such as the use of word forms not native to the language, lack of standardized recordings, and the use of arbitrary predetermined presentation levels. The authors determined inter-list equivalence, word difficulty, talker intelligibility, and psychometric slope function of the words used in a standardized Spanish language test using 16 native Spanish-speaking college students from Mexico, Panama, Venezuela, Spain, Honduras, and Columbia. Materials consisted of 50 different Spanish words for each of four word recognition lists professionally recorded by a native Spanish-speaking Mexican male. Each list of words was presented in randomized order at 8, 16, 24, and 32 dB HL for each participant. Participants were not familiarized with the words previous to the presentation. The participants were requested to repeat the stimuli

presented in the Spanish-equivalent carrier phrase, y*ou will say* ___. All participants judged the talker's speech as highly intelligible. However, analysis of the collected perception data found statistically significant differences in scores due to dialectal variations between the talker and the participants. The results showed that Mexican participants performed better at lower intensity levels than participants of other origins. The most commonly missed words contained the /s/ and /t/ phonemes, which the talker often substituted for /z/ and /k/, respectively. The test items produced by the Mexican talker provided reduced phonetic cues to the non-Mexican participants, which may have negatively impacted their scores (Weisleder & Hodgson, 1989). The authors further concluded that the results of their study provided support for the importance of testing participants in their native dialects to obtain accurate speech audiometry scores.

Nature of Regional Dialects in Mandarin

Much like Spanish, Mandarin is a language that has many different regional variants or dialects. As a language group Mandarin dialects are tonal, with four standard tones: high-level, high-rising, low-dipping, and high-falling. These suprasegmental tones carry lexical information in speech (Zhou & Marslen-Wilson, 1995). Morphologically, there are no inflections; rather, understanding semantic meaning is a result of syntactic markers and logical order (Campbell, 1991). Individual Mandarin characters often have multiple meanings and are combined with other characters in spoken and written language to avoid lexical and semantic ambiguity. Zhou & Marslen-Wilson (1995) claim that the majority of Mandarin words are bisyllabic in nature and are comprehended as whole, single words rather than as two combined words.

Many individuals from the northern regions of the People's Republic of China speak a standard dialect of Mandarin Chinese which is similar to the Beijing dialect. For

the purposes of this paper, this dialect will be referred to as Mainland Mandarin. Speakers from the Republic of China or Taiwan speak a dialect of Mandarin referred to as Taiwan Mandarin. These dialects were chosen as the focus of this study because the dialects have relatively high mutual intelligibility, yet also have distinct regional characteristics. Each dialect contains marked differences in syntactic, lexical, phonetic, and orthographic features (Li, 1984). It is estimated that 70% of lexical items are shared between the dialects (Cheng, 1985). In addition, high-quality digital speech audiometry materials that are relatively psychometrically equivalent have been recently developed and are readily-available (Nissen et al., 2005a, 2005b, 2007).

Previous research has shown conflicting evidence on the effect of testing an individual's hearing acuity with speech audiometry materials created in a mutually intelligible, yet non-regional dialect (Schneider, 1992; Weisleder & Hodgson, 1989). Thus, the purpose of this study is three-fold in nature: (a) to determine the test-retest reliability of existing sets of word recognition materials for two regional dialects of Mandarin Chinese, (b) to examine the validity of using materials from a non-regional yet mutually intelligible dialect to evaluate an individual's word recognition abilities in a quiet setting, and (c) to investigate whether a native speaker of one Mandarin dialect is able to accurately administer and score word recognition testing for speakers of a different regional dialect.

Method

Participants

Participants in this study included 32 Mandarin listeners; 16 participants were

native speakers of Mainland Mandarin and 16 spoke Taiwan Mandarin (see Table 1). The

participants ranged in age from 18 to 50 years. Each participant passed a hearing

screening including (a) otoacoustic emissions, (b) static acoustic admittance between 0.3

and 1.4 mmhos with peak pressure between −100 and +50 daPa in the test ear (American

Speech-Language-Hearing Association, 1990; Roup, Wiley, Safady, & Stoppenback,

1998), (c) acoustic reflexes present at levels ≤95 dB HL at 1000 Hz in the test ear, and (d)

thresholds of 15 dB HL or better for octave and mid-octave frequencies from 125 to 8000

Hz. All participants lived in the United States at the time of the study and spoke their

native dialect of Mandarin on a regular basis. All participants signed an informed consent

form before participating in the study (see Appendix).

Materials

Recorded materials of male speakers of each dialect of Mandarin (Mainland and

Taiwan) were presented to each participant, consisting of four full lists of 50 bisyllabic

words each. The materials are from compact discs produced by Brigham Young

University (Harris & Nissen, 2004, 2007). The words in each list were selected based on

high frequency of usage in the language, to ensure that listeners would be familiar with

them. Words that had similar pronunciation but different meanings, represented by

different characters, were avoided. In addition, words were eliminated from the original

lists if they were considered by native Mandarin Chinese judges to be culturally

insensitive, unfamiliar, or representative of inappropriate content. After listener

Table 1

Subject Information and Audiograms

Ss[a]	Sex	Age	Hometown	Time in US[b]	Better ear	125	250	500	750	1K	1.5K	2K	3K	4K	6K	8K
M1	f	37	Beijing	5:0	L	5	5	0	0	0	5	0	0	–5	5	5
M2	f	25	Shanghai	2:0	L	10	10	10	10	5	0	5	–5	0	0	15
M3	m	50	Huhhot	0:11	L	0	0	10	5	5	10	5	0	0	10	5
M4	m	26	Guangzhou	3:0	R	10	5	5	0	0	5	0	5	0	10	10
M5	f	20	Changsha	5:0	R	5	5	5	0	0	15	15	10	0	–5	–5
M6	m	22	Guangzhou	1:11	L	5	0	5	10	5	5	5	0	5	–5	–5
M7	m	23	Shanghai	4:10	L	5	5	5	5	5	10	15	10	0	5	5
M8	f	27	Huangshan	0:3	L	15	10	5	10	5	5	5	5	10	–5	5
M9	m	27	Nanjing	3:6	R	5	5	5	5	5	5	10	5	5	5	10
M10	f	30	Zhuji	1:0	R	10	10	10	5	5	5	10	5	0	0	5
M11	m	19	Shanghai	3:0	R	10	5	0	5	10	5	5	–5	–5	–5	5
M12	f	35	Tianjin	2:10	R	5	0	0	0	5	10	10	5	5	0	0
M13	f	32	Yantai	3:2	R	10	5	5	5	5	5	10	0	0	0	–5
M14	f	27	Zi Bo	3:0	L	0	0	0	–5	5	0	0	0	–5	0	0
M15	m	23	Guiyang	0:6	L	15	10	5	10	15	0	5	0	0	10	10
M16	f	28	Xi'an	3:0	L	5	0	0	10	10	10	5	10	10	0	0
T1	f	29	Keelung	5:0	L	–5	5	5	5	0	5	5	–5	–5	0	0
T2	f	27	Taitung	0:10	L	5	5	5	0	5	15	10	10	10	10	10
T3	f	49	Taipei	18:0	L	15	15	10	5	–5	0	0	5	15	10	15
T4	f	21	Pingtung	2:0	L	–5	–5	0	0	0	5	5	0	0	–10	–5
T5	m	19	Nantou	1:0	R	10	5	0	5	–5	5	5	0	0	5	0
T6	f	29	Kaohsiung	1:10	L	5	5	5	5	5	10	5	15	0	5	0
T7	f	20	Kaohsiung	1:6	L	15	5	10	10	10	10	15	10	10	10	10
T8	m	32	Taipei	3:0	R	5	0	0	5	10	0	0	5	0	–5	0
T9	f	27	Kaohsiung	1:0	L	15	10	10	10	10	10	10	5	5	15	5
T10	f	18	Taipei	0:11	R	5	5	0	0	5	5	–5	10	10	10	10
T11	f	35	Taipei	10:0	L	10	10	15	15	15	10	0	10	15	10	0
T12	m	31	Taipei	6:0	R	5	5	5	0	5	10	10	10	10	0	5
T13	f	19	Taichung	0:6	L	5	0	5	5	10	0	5	5	0	0	–5
T14	f	18	Yingko	0:6	L	5	5	5	5	5	5	5	–5	0	–10	–5
T15	m	27	Changhua	1:2	L	5	5	5	5	0	0	0	5	0	–5	0
T16	f	19	Taoyuan	0:11	L	10	5	0	0	0	5	0	5	5	0	–5

[a]Subjects that begin with *M* are Mainland subjects; subjects that begin with *T* are Taiwanese.
[b]Time lengths are in years and months notation.

evaluation, the resulting word lists were digitally adjusted to be relatively psychometrically equivalent (Nissen et al., 2005a, 2005b, 2007).

Procedure

All WR testing was performed in a double-walled sound booth using a Grason-Stadler model 1761 audiometer with a single TDH-50 Telephonics supra-aural earphone that met ANSI S3.1 standards for maximum permissible ambient noise levels for the ears not covered condition using one-third octave-band measurements (American National Standards Institute, 1999).

Each subject participated in a hearing screening and two subsequent test sessions which were similar in content and presentation format. Two sessions of WR testing of four lists in each dialect were performed for each participant to examine the test-retest reliability of the two sets of materials. All materials were presented in the ear with the lowest pure tone average for common speech frequencies (500 to 2000 Hz) for each participant. The retest session was similar to the initial test session in that all procedures were precisely replicated, including maintaining the same distance from the microphone, using the same headphones, and being tested in the same ear. Each participant was presented with four lists of 50 words each at four different intensity levels (0, 4, 8, and 12 dB HL) and in each of the two Mandarin dialects. The four intensity levels were chosen based on previous research (Nissen, 2005a), with the specified intensity levels corresponding to percentages of correct responses at 20%, 40%, 60%, and 80% of the psychometric function slopes for these materials as determined by logistic regression. These intensities were utilized to decrease the likelihood of floor and ceiling effects when

the materials were re-evaluated in the current study. Each participant was randomly assigned to one of four predetermined sequences of word list presentation which were counter-balanced for the different presentation intensity levels, subject dialect, and the dialect of the materials being presented.

Test sessions were one week apart, both of which were administered at the same time of day. Participants took at least one 5-minute break per session to avoid possible listener fatigue. The order of words within each list was randomly generated by customized software. Prior to each administration of the WR test, each participant was given instructions in English as follows:

> You will hear Mandarin words at a number of different loudness levels. Each word is two syllables in length. At the very soft loudness levels, it may be difficult for you to hear the words. Please listen carefully and repeat out loud the word that you hear. If you are unsure of the word, you are encouraged to guess. If you have no guess, say, *I don't know*. Do you have any questions?

A native interpreter was available to assist with any questions the participants had prior to the collection of data.

Participants were not familiarized with the bisyllabic words prior to testing. Each list was presented one time per participant during each WR evaluation session. Participants were asked to verbally respond to the stimuli by repeating the perceived word. Responses were scored as either correct or incorrect by a native judge of each dialect and subsequently recorded into an Excel spreadsheet.

Results

Statistical Analysis

The reliability and validity of the test materials described in this study were evaluated based on descriptive and inferential statistics. A Chi-Square analysis was used to determine if the listeners' ability to perceive the presented stimuli differed as function of the intensity (0, 4, 8, 12 dB HL), talker dialect (Mainland Mandarin and Taiwan Mandarin), listener dialect (Mainland Mandarin and Taiwan Mandarin), or the session (test and retest). Results were considered to be statistically significant at the alpha level of .01. A modified logistic regression equation was designed (Nissen et al., 2005a, 2007) and used to calculate the psychometric slope and intercept for each of the eight word lists (four Mainland Mandarin and four Taiwan Mandarin) across intensity level, talker dialect, and session. Any significant differences between these results were investigated with a Chi-Square analysis, with descriptive statistics being reported as the psychometric function slope (%/dB) at 50% and from 20 - 80%, as well as the intensity required for 50% recognition. Differences between the two test sessions are reported in terms of an odds ratio and comparisons between the performances of the two interpreters in terms of overall percentage of agreement across all stimulus presentations.

Test Material Reliability and Validity

Differences in talker dialect and listener dialect were also found to be statistically significant, $\chi^2(1, N = 32) = 8.8$, $p = 0.003$, and $\chi^2(1, N = 32) = 46.9$, $p < 0.0001$. In addition, a significant interaction between talker dialect and listener dialect was found to be significant, $\chi^2(1, N = 32) = 46.9$, $p < 0.0001$. Collapsing across test session and listener dialect, the materials produced by a talker from Mainland China exhibited steeper psychometric slope values at 50% and 20 to 80% (9.12 %/dB and 7.89 %/dB), as

compared to the Taiwan Mandarin materials (8.39 %/dB and 7.26 %/dB). The intensity required for 50% recognition of the Mainland and Taiwan Mandarin materials was 3.06 and 3.69 dB HL, respectively. Overall, the listeners from Mainland China recognized both sets of materials (50% correct recognition) at a lower mean intensity (2.89 dB HL) than the listeners from Taiwan (3.86 dB HL). Rank ordered, the best performance was with Mainland participants listening to Mainland materials (2.26 dB HL), Mainland listeners with Taiwan Mandarin materials (3.52 dB HL), and Taiwan listeners with either Taiwan or Mandarin materials (both at 3.86 dB HL). These mean slope and intensity results are illustrated in Figures 1 - 3.

As expected, the statistical analysis indicated a significant difference in the listeners' performance across the differing intensity levels, $\chi^2(1, N = 32) = 4817.5$, $p < 0.0001$, with an increase in word recognition percentage with increasing intensity. In addition, a main effect of test session was also found to be statistically significant, $\chi^2(1, N = 32) = 144.1$, $p < 0.0001$. Comparison of the test-retest statistics showed that listeners' performance was more likely to improve on the retest session, with an odds ratio of 0.7: 1.0 and a confidence interval of .661 to .743. No interactions between the test session and other variables were found to be statistically significant.

List Differences

The psychometric function slopes for the Mainland and Taiwan listeners' performance on the lists are illustrated in Figures 4 and 5, respectively. For both the Mainland and Taiwan Mandarin listeners the lists (both Mainland and Taiwan materials) differed significantly, $\chi^2(1, N = 32) = 561.2$, $p < 0.0001$, and $\chi^2(1, N = 32) = 926.9$, $p < 0.001$. The average differences among the Mainland lists for Mainland listeners was

.61 %/dB in slope at 50% of the psychometric function, .53 %/dB from 20 to 80% of the

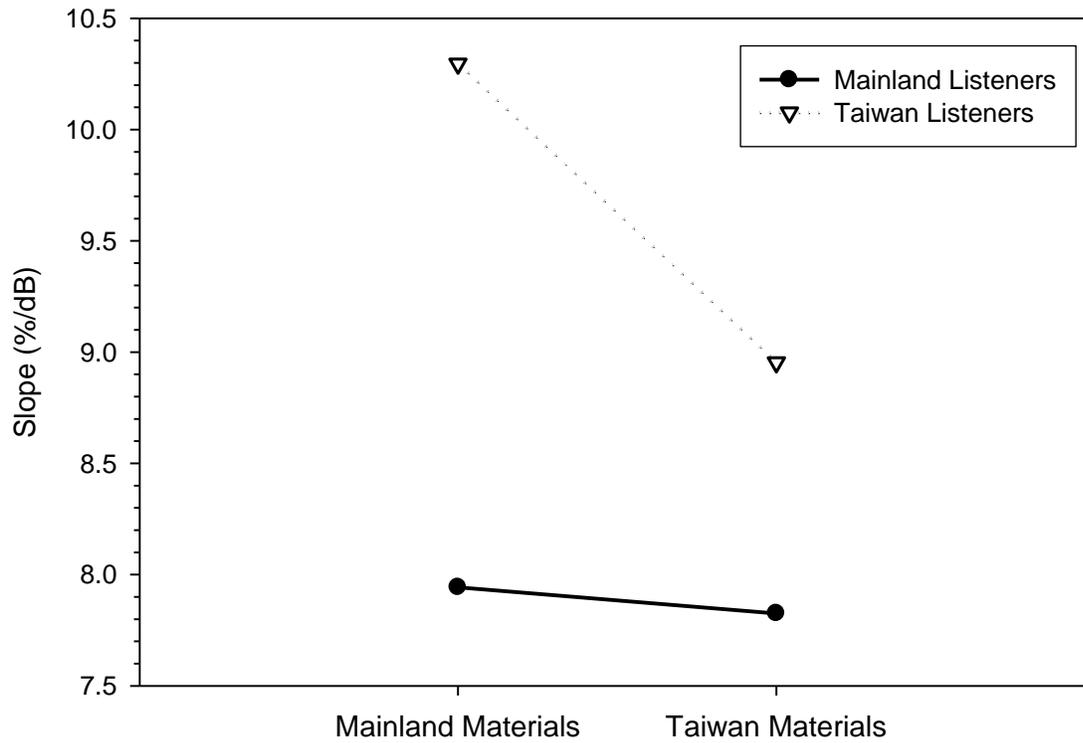function, and .45 dB HL for 50% recognition. The difference in intensity for 50%

Figure 1. The psychometric function slope at 50% of the function curve for Mainland and

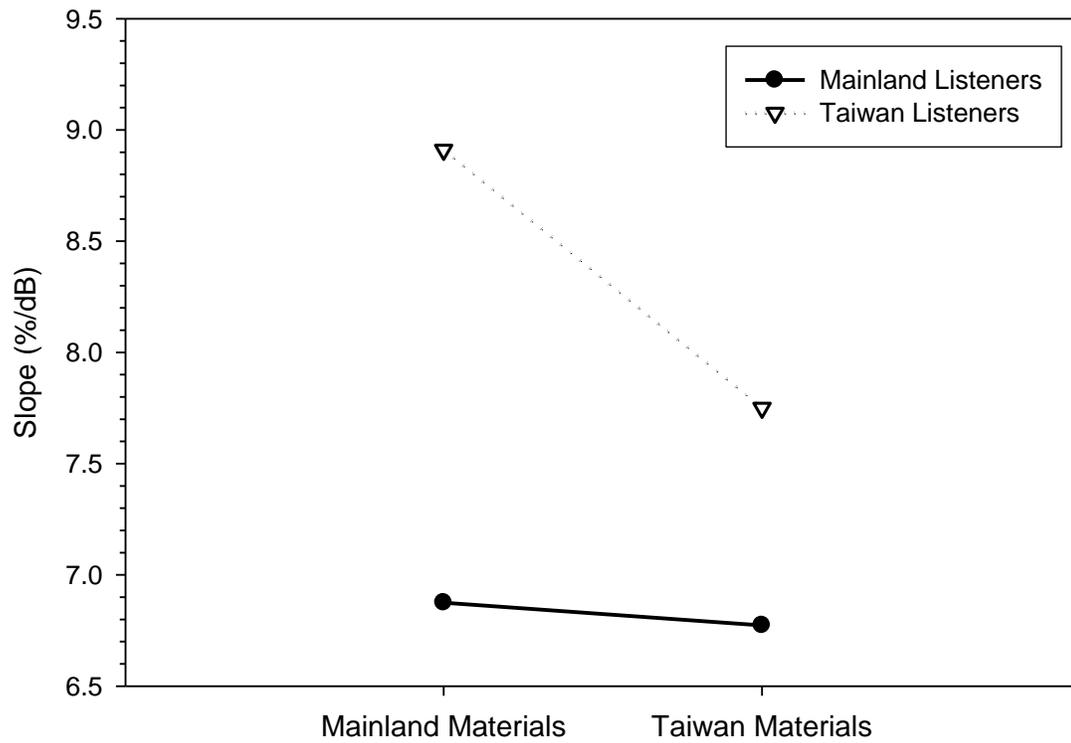Taiwan Mandarin materials across Mainland and Taiwan listeners.

Figure 2. The psychometric function slope at 20% to 80% of the function curve for

Mainland and Taiwan Mandarin materials across Mainland and Taiwan listeners.
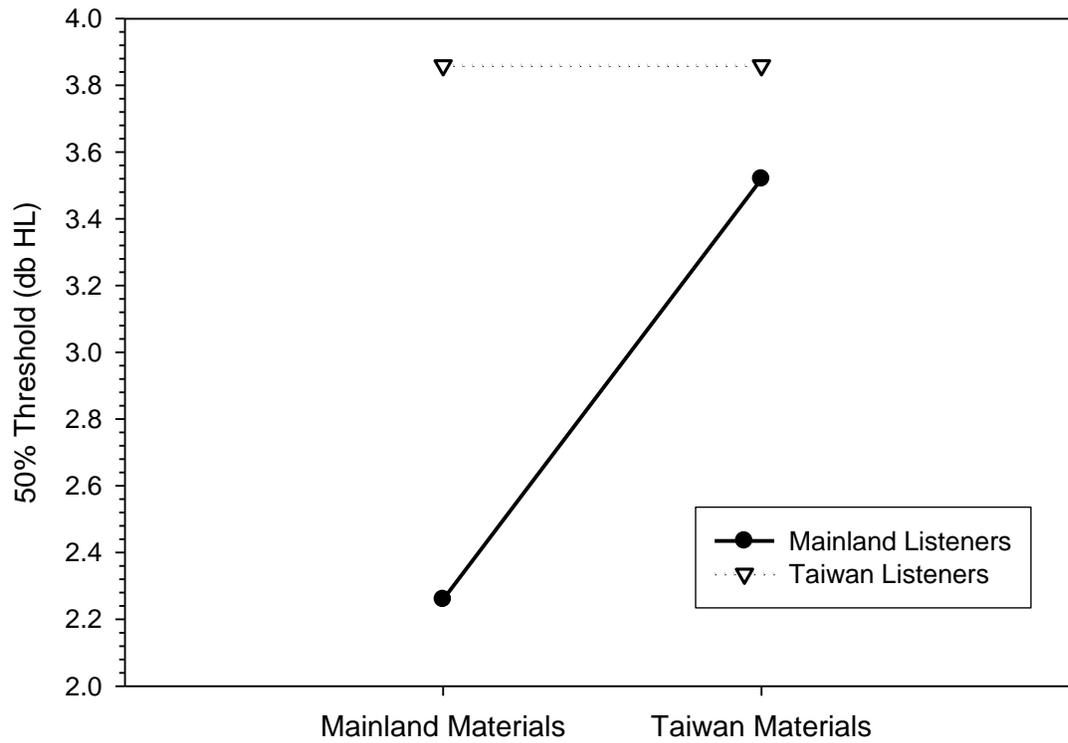
Figure 3. Intensity required for 50% intelligibility for Mainland and Taiwan Mandarin

materials across Mainland and Taiwan listeners.

Figure 4. Psychometric functions for Mainland listeners on Mainland and Taiwan

Mandarin word recognition lists [lines A through D represent Mainland Mandarin word

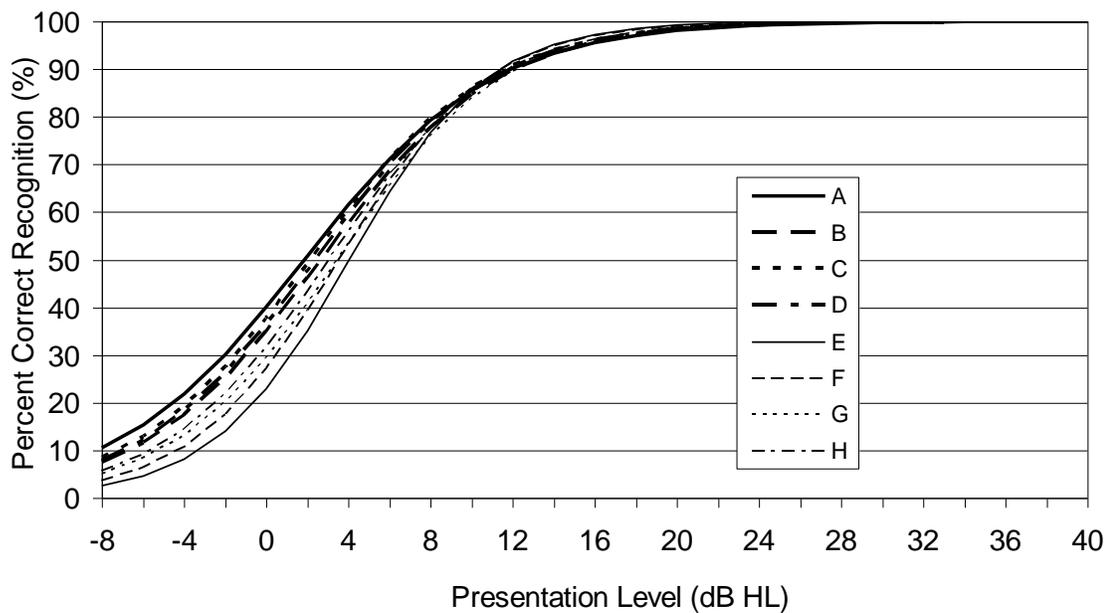recognition lists; lines E through H represent Taiwan Mandarin word recognition lists].

Figure 5. Psychometric functions for Taiwan listeners on Mainland and Taiwan

Mandarin word recognition lists [lines A through D represent Mainland Mandarin word

recognition lists; lines E through H represent Taiwan Mandarin word recognition lists].
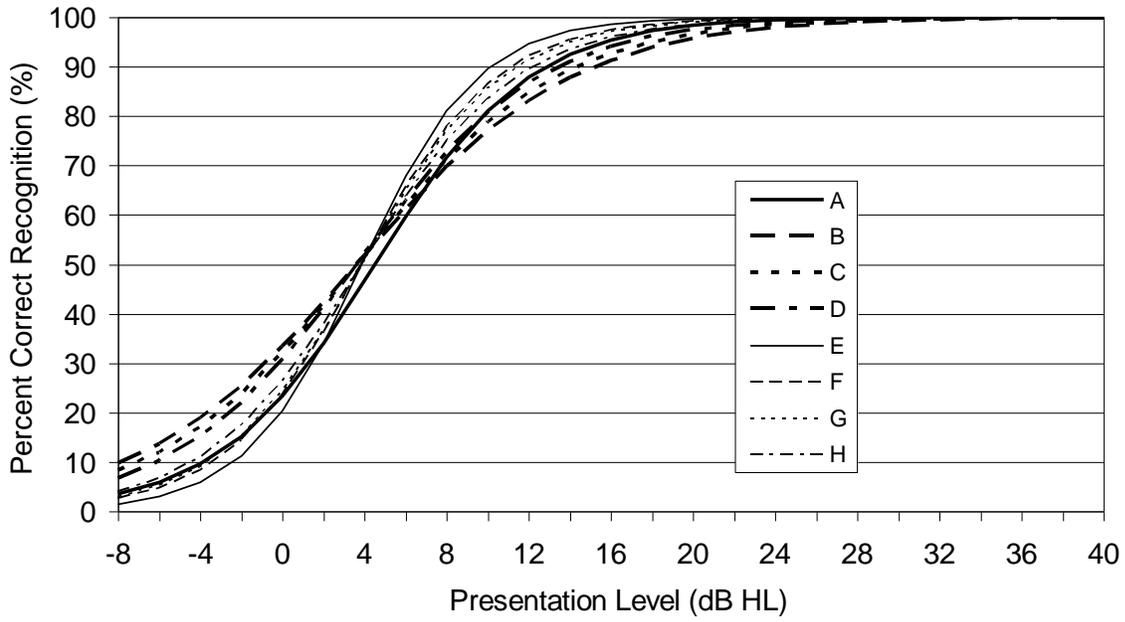
 recognition between the lists ranged from .24 to .81 dB HL. The average differences

between the Taiwan lists for Taiwan listeners were .64 %/dB HL in slope at 50% of the

psychometric function, .56 %/dB HL from 20 to 80% of the function, and .01%/ dB HL

for 50% recognition. The difference in intensity for 50% recognition between the lists

ranged from .004 to .02 dB HL. A detailed listing of the values for the psychometric

function slope (%/dB HL) at 50% and from 20 - 80%, as well as the intensity required for

50% recognition for both sets of lists is reported in Tables 2 and 3, respectively.

Inter-judge Reliability Across Interpreters

        Each listener responses were scored as either correct or incorrect by one regional

and one non-regional native judge. The two judges recorded the same judgment of correct

or incorrect for 25,383 out of 25,600 participant responses, for an overall percentage of

agreement of 99.1% across all stimulus presentations.

Table 2

Mean Performance of Mainland Chinese Listeners on Mainland and Taiwan Lists

List	a^a	b^b	Slope at 50% [c]	Slope from 20-80% [d]	Threshold[e]
Mainland A	0.4029	-0.2180	7.9	6.9	1.8
Mainland B	0.6221	-0.2337	7.4	6.4	2.7
Mainland C	0.5006	-0.2334	8.6	7.4	2.1
Mainland D	0.5607	-0.2352	7.9	6.8	2.4
M	0.5216	-0.2301	7.9	6.8	2.2
Minimum	0.4029	-0.2337	7.4	6.4	1.8
Maximum	0.6221	-0.2180	8.6	6.9	2.7
Range	0.2192	0.0157	1.2	0.5	0.9
SD	0.0934	0.0081	0.5	0.1	0.4
Taiwan A	1.2164	-0.3005	8.2	7.1	4.0
Taiwan B	0.9897	-0.2806	7.3	6.3	3.5
Taiwan C	0.8723	-0.2526	8.6	7.4	3.4
Taiwan D	0.7710	-0.2525	7.2	6.3	3.0
M	0.9623	-0.2715	7.8	6.8	3.5
Minimum	0.7710	-0.3005	7.2	6.3	3.0
Maximum	1.2164	-0.2525	8.6	7.4	4.0
Range	0.4454	0.0480	1.4	1.1	1.0
SD	0.1915	0.0234	0.7	0.1	0.4

[a]a = regression intercept. [b]b = regression slope. [c]Psychometric function slope (%/dB) at 50% was calculated from 49.999 to 50.001%. [d]Psychometric function slope (%/dB) from 20-80%. [e]Intensity required for 50% intelligibility.

Table 3

Mean Performance of Taiwan Listeners on Mainland and Taiwan Lists

List	a[a]	b[b]	Slope at at 50% [c]	Slope from 20-80% [d]	Threshold[e]
Mainland A	1.1891	-0.2641	9.3	8.1	4.5
Mainland B	0.6938	-0.1905	11.3	9.8	3.6
Mainland C	0.7527	-0.2063	9.6	8.3	3.6
Mainland D	0.8179	-0.2251	11.0	9.5	3.6
M	0.8634	-0.2215	10.3	8.9	3.8
Minimum	0.6938	-0.2641	9.3	8.1	3.6
Maximum	1.1891	-0.1905	11.3	9.8	4.5
Range	1.1972	0.0736	2.0	1.7	0.9
SD	0.2229	0.0317	0.9	0.8	0.4
Taiwan A	1.3642	-0.3526	8.4	7.3	3.9
Taiwan B	1.1720	-0.3039	9.1	7.9	3.8
Taiwan C	1.1309	-0.2920	8.7	7.6	3.9
Taiwan D	1.0237	-0.2653	9.6	8.2	3.8
M	1.1727	-0.3034	8.9	7.7	3.8
Minimum	1.0237	-0.3526	8.4	7.3	3.8
Maximum	1.3642	-0.2653	9.6	8.2	3.9
Range	0.3405	0.0873	1.2	0.9	0.1
SD	0.1421	0.0365	0.5	0.4	0.1

[a]*a* = regression intercept. [b]*b* = regression slope. [c]Psychometric function slope (%/dB) at 50% was calculated from 49.999 to 50.001%. [d]Psychometric function slope (%/dB) from 20-80%. [e]Intensity required for 50% intelligibility.

Discussion

The aims of this study were to examine the reliability and validity of using two dialects of Mandarin word recognition materials to evaluate the speech perception abilities of regional and non-regional speakers of the presented dialects and to determine the reliability of these materials across multiple sessions and between individual lists. In addition, this study evaluated whether a native speaker of one Mandarin dialect is able to accurately administer and score word recognition testing for speakers of a different regional dialect.

The main effect of session was found to be significant. The participants were able to identify 50% of the words in a list at a significantly lower intensity (odds ratio of 0.7: 1.0) during the retest session. Thus, a learning effect was likely responsible for this increase in performance. The participants may have been more familiar with the words presented or the testing procedure in general. In other words, the participants better understood what was expected of them the second time that they participated, and thus their scores improved. It is possible that addition time between the test and retest session may have reduced such differences. Although session differences were statistically significant, it is unclear if such differences are clinically significant. In addition, scores between the lists remained relatively stable across test session, indicating that the materials maintained good inter-list equivalence across multiple presentations.

Statistically, differences in listener performance varied across the regional dialect of the materials and the participants. Taiwan listeners presented with a steeper slope at 50% for both sets of dialect materials as shown in Figure 1. Similar results were noted in Figure 2 for the slope at 20 to 80%. In each of these slope measurements, the slope for Taiwan listeners was more than 2%/dB greater than Mainland listeners for Mainland

materials and more than 1%/dB greater than Mainland listeners for Taiwan materials. Figure 3 illustrates that threshold 50% intelligibility for Taiwan listeners was 1.6 dB HL greater than Mainland listeners for Mainland materials and 0.4 dB greater than Mainland listeners for Taiwan materials.

A possible explanation for these differences may talker differences in the recorded materials. The talker selected to record the Mainland materials was a professional speaker in his native dialect. Informal analysis of these records by native speakers indicated that his style of speech was characterized by increased articulation, precision and prosody, and a slower speech rate than the Taiwan talker. In addition, the Mainland talker was relatively older than many of the participants, whereas the Taiwan speaker was approximately the same age as the listeners. These factors may have made the favored the intelligibility of the Mainland talker over the Taiwan talker.

Anecdotally, the island of Taiwan is relatively small in land area when compared to Mainland China and likely contains less dialectal variation in the standard language. Most of the Taiwan participants in this study originated from geographic locations that were relatively close to the origin of the talker and the other participants, whereas the Mainland participants originated from a much larger geographic area. There may also be a differing degree of exposure to the non-regional dialect for the participants. Participants from Taiwan may be more familiar with the dialect structure of Mainland Mandarin than Mainland subjects may be with Taiwan Mandarin. Participants in this study were from various locations of Mainland China and Taiwan; there are other dialect influences in their regional speech that cannot be accounted for within the scope of this study. Because of these factors, we expected to find that Taiwan participants would have performed

better on threshold measurements than the Mainland participants. However, the results favored the Mainland participants rather than Taiwan participants for threshold measurements. Another possible explanation for this finding may be that the study was limited in its sample size, with 32 listeners participating in the study. A study that included a larger sample size of participants may be more representative of the native Mandarin-speaking population.

Probably more importantly, although the results of this study were statistically different across talker and listener dialect, it is unclear if such differences are large enough to make a clinically significant difference. Most audiologists test in increments of 5 dB, so differences less than 5 dB may not alter clinical test interpretation. Also it is unclear if the methodology employed in this study is sensitive enough to make definitive conclusions based on such differences.

The psychometric function slopes and intensity thresholds for the Mainland and Taiwan listeners' performance on the lists were also found to be statistically significant. However, once again it is not known if such differences are clinically significant. For example, the difference in intensity for 50% recognition between the lists ranged from .24 to .81 dB HL for the Mainland lists and .004 to .02 dB HL for the Taiwan lists. In terms of clinical relevance, such differences are minimal and indicate a relatively high degree of inter-list reliability.

The high percentage of agreement between the two interpreters from the two different regional dialects of Mandarin provides evidence that a speaker from either dialect could accurately administer and score the WR results from a speaker from the other dialect. However, this result may be language specific to Mainland and Taiwan

Mandarin, and may not apply to regional dialects of other languages. This may not be the case for dialects that are not as mutually intelligible as the dialects in the present study.

The findings of this study are similar to the findings of Schneider's study (1992), in that no clinically significant differences were found between talkers of different dialects for speech audiometry materials. Once again, one possible reason for similar findings in both studies may be the high mutual intelligibility of the dialects used for each study. Another possible reason that findings of this study may be more similar to those of Schneider rather than Weisleder and Hodgson (1989) may be that Schneider used two dialects of Spanish materials which were presented to participants, while Weisleder and Hodgson presented only one dialect of Spanish to speakers from six different countries. The design of the current experiment was more similar in nature to the design of Schneider's experiment, in that materials from two dialects were presented.

An area of further research might be to evaluate the impact of regional dialect on two dialects that are less mutually intelligible. In addition, it would be interesting to examine if these findings also generalize across children and individuals with a hearing impairment. Despite these limitations, it is hoped that the results of this study will inform the continued development of addition digital speech audiometry materials for a wide variety of languages and dialects.

References

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

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

Brandy, W. T. (2002). Speech audiometry. In J. Katz (Ed.), *Handbook of Clinical Audiology* (pp. 96-110). Philadelphia: Lippincott Williams & Wilkins.

Cambell, G. L. (1991). *Compendium of the world's languages.* London & New York: Routledge.

Cheng, R. L. (1985). A comparison of Taiwanese, Taiwan Mandarin and Peking Mandarin. *Language and Cognitive Processes, 61*, 352–377.

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*, 164–168.

Deutsch, L. J., & Kruger, B. (1971). The systematic selection of 25 monosyllables which predict the CID-W22 speech discrimination score. *Journal of Audiology Research, 11,* 286-290.

Egan, J. P. (1948). Articulation testing methods. *Laryngoscope, 58,* 955-991.

Elpern, B. S. (1961). The relative stability of half-list and full-list discrimination tests. *Laryngoscope, 71,* 30-36.

Epstein, A. (1978). Speech Audiometry. *Otolaryngolic clinics of North America, 11*, 667–676.

Giolas, T. G., & Epstein, A. (1963). Comparative intelligibility of word lists and continuous discourse. *Journal of Speech and Hearing Research, 6*, 349-358.

Harris, R. W., Nissen, S. L., Slade, K., & Dukes, A. (2007). *Taiwan Mandarin Chinese speech audiometry materials*, [Compact Disc]. Provo, UT: Brigham Young University.

Harris, R. W., Nissen, S. L., & Jennings, L. J. (2004). *Mandarin Chinese speech audiometry materials: Putonghua*, [Compact Disc]. Provo, UT: Brigham Young University.

Ingram, K., Bunta, F., & Ingram, D. (2004). Digital data collection and analysis: Application for clinical practice. *Language, Speech and Hearing Services in Schools, 35*, 112-121.

Jirsa, R. I., Hodgson, W. R., & Goetzinger, C. P. (1975). Unreliability of half-list discrimination tests. *Journal of the American Audiological Society, 1,* 47-49.

Kruel, I. J., Nixon, C., Kryter, K., Bell, D, Lang, J., & Schubert, I. (1968). A proposed clinical test of speech discrimination. *Journal of Speech and Hearing Research, 11,* 536-552.

Li, D. C. C. (1984). The sociolinguistic context of Mandarin in Taiwan: Trends and developments. *Paper presented at the Fourteenth International Conference on Sino-Tibetan Languages and Linguistics*, Gainesville, FL, USA.

Martin, F. N., Champlain, C. A., & Perez, D. D. (2000). The question of phonetic balance in word recognition testing. *Journal of the American Academy of Audiology, 11,* 489-493.

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*, 255–262.

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*, 379–390.

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

Nissen, S. L., Harris, R. W., & Slade, K. B. (2007). Development of speech reception threshold materials for speakers of Taiwan Mandarin. *International Journal of Audiology, 46*, 449–458.

Ostergard, C. A. (1983). Factors influencing the validity and reliability of speech audiometry. *Seminars in Hearing, 4*, 221–239.

Ramkissoon, I. (2001). Speech recognition thresholds for multilingual populations. *Communication Disorders Quarterly, 22*, 158–162.

Resnick, D. M. (1962). Reliability of the twenty-five word phonetically balanced lists. *Journal of Audiology Research, 2,* 5-12.

Roup, C. M., Wiley, T. L., Safady, S. H., & Stoppenbach, D. T. (1998). Tympanometric screening norms for adults. *American Journal of Audiology, 7*, 55–60.

Schneider, B. S. (1992). Effect of dialect on the determination of speech-reception thresholds in Spanish-speaking children. *Language, Speech, and Hearing Services in Schools, 23*, 159–162.

Schumaker, D.R. & Rintelmann, W. I. (1974). Half-list vs full-list speech discrimination testing in a clinical setting. *Journal of Audiology Research, 2,* 16-17.

Schwartz, D. M., Bess, I. H., & Larson, V. D. (1977). Split-half reliability of two word discrimination tests as a function of primary-to-secondary ratio. *Journal of Speech and Hearing Disorders, 42,* 440-445.

Skarakis-Doyle, E., & Mallet, C. A. (1991). Test-retest reliability: Another evaluation of the test of problem solving. *Language, Speech, and Hearing Services in Schools, 22*, 278–279.

Smith, A. (2004). *The fifteenth most serious health problem: The WHO perspective.* Paper presented at the IFHOH World Congress, Helsinki.

von Hapsburg, D., & Pena, E. D. (2002). Understanding bilingualism and its impact on speech audiometry. *Journal of Speech, Language, and Hearing Research, 45,* 202–213.

Weinstein, B. E. (1996). Treatment efficacy: Hearing aids in the management of hearing loss in adults. *Journal of Speech & Hearing Research, 39*, 37–45.

Weisleder, P., & Hodgson, W. R. (1989). Evaluation of four Spanish word-recognition-ability lists. *Ear and Hearing, 10*, 387–382.

Zhou, X., & Marslen-Wilsen, W. (1995). Morphological structure in the Chinese mental lexicon. *Language and Cognitive Processes, 10*, 545–600.

APPENDIX

Informed Consent

Research Participation Form

Participant:_____ Age:_____

      You are asked to participate in a research study sponsored by the Department of Audiology and Speech Language Pathology at Brigham Young University, Provo, Utah. The faculty director of this research is Richard W. Harris, Ph.D. Students in the Audiology and Speech-Language Pathology 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 participant 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 participants 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 have any questions regarding this research project you may contact Dr. Richard W. Harris, 131 TLRB, Brigham Young University, Provo, Utah 84602; phone (801) 422-6460. If you have any questions regarding your rights as a participant in a research project you may contact Dr. Shane Schulthies, Chair of the Institutional Review Board, 122A RB, Brigham Young University, Provo, UT 84602; phone (801) 422-5490.

      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