2013-03-20

Test-Retest Reliability of Speech Recognition Threshold Material in Individuals with a Wide Range of Hearing Abilities

Karin Leola Caswell
Brigham Young University - Provo

Test-Retest Reliability of Speech Recognition Threshold Material in Individuals with a Wide Range of Hearing Abilities

Karin L. Caswell

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
David L. McPherson
Shawn L. Nissen

Department of Communication Disorders

Brigham Young University

March 2013

ABSTRACT

Test-Retest Reliability of Speech Recognition Threshold Material in Individuals with a wide range of hearing loss

Karin L. Caswell

Department of Communication Disorders, Brigham Young University
Master of Science

The purpose of this study was to evaluate an updated list of digitally recorded Speech Recognition Threshold (SRT) materials for test-retest reliability. Chipman (2003) identified 33 psychometrically equated spondaic words that are frequently occurring in English today. These digitally recorded words were used to determine the SRT of 40 participants using the American Speech-Language Hearing Association guidelines. The participants were between the ages of 19 and 83 years and presented with hearing impairment ranging from normal to severe. The individual's pure-tone averages classified 16 participants with normal hearing to slight loss, 12 participants with mild loss, and 12 participants with moderate to severe hearing loss. The speech materials were presented to participants in one randomly selected ear. The SRT was measured for the same ear in both the test and retest conditions. The average SRT for the test condition was 22.7 dB HL and 22.8 dB HL in the retest condition with an improvement of 0.1 dB for retest but no significant difference was identified. Using a modified variance equation to determine test-retest reliability resulted in a 0.98, indicating almost perfect reliability. Therefore the test-retest reliability was determined to be exceptional for the new SRT words.

Keywords: speech recognition threshold, word recognition, speech audiometry, pure-tone averages, test-retest reliability, digitally recorded materials.

ACKNOWLEDGMENTS

This research has taken many dedicated hours to complete and I am appreciative of my thesis committee and Dr. Harris for his willingness to teach and to advise throughout this process. Additionally, thank you Meghan Grange for all your help with the technical aspects of completing my thesis and for being my thesis partner. I am especially grateful for my husband Clinton and my four children who have supported, encouraged, and sacrificed so I could go back to school and follow my dreams. Thank you.

TABLE OF CONTENTS

LIST OF TABLES v

LIST OF APPENDIXES vi

DESCRIPTION OF STRUCTURE AND CONTENT vii

Introduction 1

Method 6

Materials 6

Participants 6

Calibration 7

Procedure 8

Results 10

Discussion 12

Conclusions 15

References 18

LIST OF TABLES

Table Page

1. Hearing Classification of Participants Based on Pure-Tone Averages..... 7

2. Descriptive Statistics for Participant's Age, Speech Recognition Threshold (SRT), Pure-Tone Average (PTA), Two-Frequency Pure-Tone Average (2PTA) and Differences Between Means..... 10

LIST OF APPENDIXES

Appendix Page

A. Annotated Bibliography 21

B. List of Spondaic Words 41

C. Informed Consent 42

DESCRIPTION OF STRUCTURE AND CONTENT

The body of this thesis is written as a manuscript suitable for submission to a peer-reviewed journal in speech-language pathology. An annotated bibliography is presented in Appendix A.

Introduction

An important part of evaluating hearing is to determine the ability to hear and understand speech. A hearing evaluation begins by determining hearing at specific frequencies using pure-tones. These pure-tones can identify hearing loss, but they cannot determine how well a person can understand speech. Speech materials (words, sentences) are needed to determine the threshold at which speech can be heard and understood. Speech audiometry serves many purposes in a hearing evaluation; it validates pure-tone hearing thresholds and identifies how well an individual perceives and understands speech. Speech audiometry is also useful for evaluating individuals that are difficult to test, aid in audiological rehabilitation, and in evaluating hearing aids (American Speech-Language Hearing Association [ASHA], 1988).

Speech recognition thresholds (SRT) are used to find the minimal level at which speech can be understood 50% of the time. The speech thresholds are found by asking the person to repeat the perceived words from a closed list of words. The words used in SRT are phonetically dissimilar to ensure that the individual is not just discriminating between similar phonemic words but perceives the correct word. In English spondaic words are typically used to measure the SRT. Spondaic words are bisyllabic words with equal stress on each syllable (e.g., cowboy). These words are used because they have the highest homogeneity in English or are equally intelligible (Egan, 1948).

The pure-tone audiogram can be verified by the SRT measurement (ASHA, 1988; Fletcher, 1950; Wilson, Morgan, & Dirks, 1973). An individual's pure-tone average (PTA), the mean hearing threshold level for the frequencies of 500, 1000, and 2000 Hz correlates well with the individual's SRT. The PTA can also be calculated using the two best pure-tone thresholds at 500, 1000, and 2000 Hz (usually 500 and 1000 Hz) if the individual presents with a sloping

hearing loss. This PTA is called two-frequency PTA (2PTA). It is normal for the SRT and PTA to fall within about a 5 dB range (ASHA, 1988). Disagreement of more than 10 dB between an individual's PTA and SRT can indicate pseudohypacusis. It also can indicate problems within the audiological exam due to equipment defects or failure to explain or understand test instructions.

Speech audiometry materials are well developed in English and are a comprehensive part of an audiological evaluation (Nissen, Harris, & Slade, 2007). Throughout the years research has attempted to standardize SRT testing to achieve valid and reliable results. Ostergard (1983) outlined the various factors that affect speech audiometry measures and increase variability in testing. Test words, instructions, procedures, talker differences, presetation mode are some factors that can influnce variability and affect or invalidate speech audiometry results. Standards have been recommended by ASHA (1988) for conducting and measuring SRT to reduce variability. Research has reviewed (a) methods for presenting SRT words, (b) proposed ways to control for statistical variance, and (c) identified the appropriate words for testing.

There are two methods of presenting SRT words: (a) monitored-live-voice (MLV) and (b) recorded materials. Audiologists present speech materials through MLV presentation more often than using recorded materials in clinical settings (Martin, Champlin, & Chambers, 1998). However, the preferred and recommended method for SRT is the use of recorded materials (ASHA, 1988). The intensity of each word can be controlled for in recorded materials whereas in live presentation intensity may become variable. Also the talkers themselves can influence variability and intelligibility of the words (Hood & Poole, 1980; Kreul, Bell, & Nixon, 1969). Tape recorded materials can deteriorate and become distorted over time and therefore digital

recordings are the preferred method for speech audiometry testing (Kamm, Carterette, Morgan, & Dirks, 1980).

Recorded materials are more consistent across all settings and evaluations providing a standardized method of delivery (Mendel & Owen, 2011). Hood and Poole (1980) insisted that recorded speech materials are mandatory to achieve standardization for SRT presentation. Standardizing the way in which SRT materials are presented reduces variability and ensures repeatability and consistency of the test results (Di Berardino et al., 2010; Hood & Poole, 1980).

Another factor important in speech audiometry is to identify the variability within and between subjects that can affect the SRT measurement. One way of identifying statistical variation is by determining intrasubject variability. This variability can be accounted for through the use of mathematical models (Ostergard, 1983; Raffin & Thornton, 1980, Thornton & Raffin, 1978). Thornton and Raffin identified two sources of errors that can affect statistical variation: the relationship between test performance and communicative function (test validity) and the consistency across test forms (test reliability). Reliability refers to a precise measurement that will remain stable across repeated measurements. A test is considered reliable if an individual achieves approximately the same results over multiple testing times. The test-retest reliability is what will be evaluated in the present investigation.

Hudgins, Hawkins, Karlin, and Stevens (1947) identified four crucial elements in creating appropriate speech audiometry materials. First, the words used in measuring the SRT should be familiar to the listeners. Familiarity is important because SRT should only measure the threshold of word intelligibility and not the individual's vocabulary knowledge. Second, phonetically dissimilar words should be used to avoid difficulty in discriminating among similar sounding words. Third, words should contain a representative sample of all speech sounds in English.

Hudgins et al. noted that all speech sounds were not always necessary for developing a reliable measurement. Finally, words with maximum homogeneity with regards to intelligibility should be utilized. Words that are homogeneous in regards to audibility have similar levels of intensity when the threshold of hearing is reached. In order to maximize homogeneity, spondaic words are used in determining SRT, and the individual word's thresholds can be equated for equal intelligibility (Hudgins et al. 1947, Wilson & Strouse, 1999). Words for SRT can also be homogeneous in terms of psychometric function slope. It is desirable to utilize words which are homogenous with respect to audibility and psychometric function slope when developing stimuli to measure the SRT (Harris, Nissen, Pola, McPherson, Tavartkiladze, & Eggett, 2007). Words that have similar and steep psychometric function slopes of about 10%/dB or greater at the 50% threshold point are preferred for SRT testing (Hudgins et al., 1947). Words with a steeper slope are used so fewer words and less time are needed to determine an individual's SRT.

Hirsh, Davis, Silverman, Reynolds, Eldert, and Benson (1952) developed the 36 Central Institute for the Deaf (CID) W-1 spondaic words for SRT testing. More recently, Chipman (2003) examined this 36 word list to determine familiarity of the words and word frequency in everyday usage. Chipman (2003) used the Standard Corpus of Present-Day American English (Francis & Kučera, 1982) and the Frown Corpus (Hundt, Sand, & Skandera, 1999) for comparison of word familiarity and usage. Chipman determined that many of the spondaic words of the CID W-1 list were less familiar than other more frequently occurring spondaic words. The CID W-1 spondaic words were examined, and 10 out of the 36 words were not included in the top 20,000 words, and 14 of the words were not in the top 10,000 words commonly used in the English language today. Following the guidelines of Hudgins et al. (1947) criteria for producing speech audiometry stimuli, Chipman produced a list of 33 spondaic

words which included 19 original CID W-1 words and 14 new spondaic words, all of which are commonly used and therefore more familiar to listeners. The words (Appendix A) replaced the less frequently used and therefore less familiar words. The 33 words were digitally recorded by both male and female talkers, and the slopes of the words were psychometrically equated at the 50% correct point. The recorded words were determined to have a steep psychometric function slope with average slopes of 16.2%/dB for male talker recordings and 15.2% /dB for female talker recordings.

Jacobs (2012) conducted a study in order to establish test-retest reliability of the 33 spondaic SRT words developed by Chipman (2003). Jacobs compared the SRT test-retest measurements for normal hearing participants. Eight speech thresholds (four test measurements, four retest measurements) were determined for each of the 40 participants using the male and female talker recordings. Jacobs also evaluated if gender of talker influenced test-retest reliability for measuring the SRT. It was determined that talker gender had little effect on test-retest reliability with male and female talkers differing by only 0.2 – 0.3 dB with no significant difference for determining SRT. Jacobs determined that the SRT related to the individuals PTA, and the difference between SRT test and retest and the PTA were clinically acceptable. The average SRT for the male talker resulted in 1.4 dB improvement from test to retest condition, and for the female talker retest improved by 1.2 dB. The difference of SRT test-retest was identified as significant; however, Jacobs determined that the results were clinically acceptable because they were within the margin of error +/- 5 dB as related to pure-tone testing (ANSI, 2004). Jacobs suggested that the differences between test-retest could be due to learning effects.

However, using a modified variability equation to measure test-retest reliability identified the reliability to be poor at 0.47. Jacobs noted that participants with only normal hearing result

in a small variance among subjects. The smaller variance produced a smaller value which reduced test-retest reliability. An increased variance of hearing ability among participants might produce better reliability for test-retest measurement of the SRT. The purpose of the present investigation was to determine a clinically adequate test-retest reliability range for the SRT list developed by Chipman in subjects with differing degrees of hearing abilities and impairments. This investigation also examined the correlation between the PTA and SRT.

Method

Materials

The speech materials used in this study are the 33 spondaic words selected and digitally recorded at Brigham Young University by Chipman (2003). Chipman produced both male and female talker recordings; however since no significant difference between talkers was identified for SRT the male talker recordings of the words was selected for use in this study (Jacobs, 2012). See (Appendix B) for the full list of words.

Participants

The participants were 40 (18 female and 22 male) native English speakers with no identified cognitive impairments. The participants were adults between 19 to 83 years of age that signed an informed consent for participation in this study approved by the Brigham Young University Institutional Review Board (Appendix B). Throughout this study basic ethical considerations were made for the protection of the research participants.

Each individual participated in a hearing evaluation prior to beginning the study to determine type and severity of hearing loss. The hearing evaluation included pure-tone, bone conduction, and tympanometry testing, and all findings were documented. To increase intrasubject variability, the goal of this study was to find 40 participants with hearing abilities

that were representative of normal to moderately severe hearing impairment ranges. The Scale of Hearing Impairment developed by Goodman (1965) and then later adapted by Clark (1981) was used to identify the subjects hearing levels. Normal hearing was identified as –10 to 15 dB HL, slight hearing loss 16 to 25 dB HL, mild hearing loss 26 to 40 dB HL, moderate hearing loss 41 to 55 dB HL, moderately severe hearing loss 56 to 70 dB HL, and severe hearing loss 71 to 90 dB HL. The participants hearing ability ranged between normal to moderately severe sensorineural hearing loss with one participant having severe hearing loss. The participants hearing classification was determined based on the individual's PTA. The participants PTA ranged between -5.0 dB HL to 71.7 dB HL. For a list of participants hearing classification see Table 1.

Table 1

Hearing Classification of Participants Based on Pure-Tone Averages

	Range dB HL	Number of Participants
Normal	-10 to 15	10
Slight	16 to 25	6
Mild	26 to 40	12
Moderate	41 to 55	6
Moderately severe	56 to 70	5
Severe	71 to 90	1

Calibration

A double-walled sound suite was used for all SRT Measurements. This sound suite met American National Standards Institute ANSI S3.1 (1991) standards for maximum permissible

ambient noise levels for test enclosure. A Grason- Stadler GSI 61 (model 1761) Clinical Audiometer was used to conduct all testing. The audiometer was calibrated to ANSI S3.6 (2010) standards at the beginning of data collection and then weekly throughout the study.

Procedure

Each participant was asked to complete one session lasting no longer than 75 to 90 minutes. In the first part of the session the individual participated in a pure-tone and tympanometry testing to document type and severity of hearing loss to ensure participants with a wide variety of hearing abilities. Once the participants were determined eligible for the study, SRT testing was conducted with a 10 minute break between the test and retest measurement of the SRT. Instructions were read to each participant prior to familiarization and testing to ensure understanding of the task, appropriate response for the task, and understanding that a response was needed even at faint listening levels (ASHA, 1988).

> You will now hear two-syllable words at a number of different loudness levels. At the very soft loudness levels it may be difficult for you to hear the words. For each word, listen carefully and then repeat what you think the word was. If you are not sure you may guess. If you have no guess wait silently for the next word. Do you have any questions?

All participants were provide with a printed list of the 33 spondaic words and the digitally recorded words were presented at 50 dB HL or a comfortable listening level as indicated by the individual's PTA and the individual themself. This familiarized the participants with the spondaic words to be used in measuring the SRT. Familiarization of speech materials aids participant's response accuracy and controls for effects that prior knowledge of test vocabulary can have on SRT measurements (ASHA, 1988; Tillman & Jerger, 1959).

The SRTs were measured using ASHA (1988) guidelines for SRT testing using the 2 dB descending approach. Custom software was used to randomize and present the spondaic words to the participants. All words were presented through a single TDH-50P head phone to control for variability. The SRT starting level was established by presenting one spondaic word at 30 to 40 dB above the estimated SRT as determined by the individual's PTA (ASHA, 1988). The intensity was decreased by 10 dB for each of the correct responses to the subsequent spondaic words (ASHA, 1988; Martin & Stauffer, 1975). When the participant provided an incorrect response a second word was presented at that same level. The starting level was obtained when the participant could no longer provide a correct response for two consecutive words at the same intensity level. The SRT starting level was established at 10 dB above the level in which two consecutive words could no longer be repeated correctly.

Once the SRT starting level was established, two spondaic words were presented at each intensity level as it decreased by 2 dB for each successive presentation. Five out of the first six spondaic words needed to be correct to continue the process of decreasing the intensity levels by 2 dB. However, if the correct responses were not obtained then the starting level was increased by 6 dB until five out of six words were repeated correctly. The test was completed when five out of the last six words presented were incorrect. To determine the participants SRT with the least amount of variability the total number of correct responses was subtracted from the starting level and a 1 dB correction factor was added (ASHA, 1988; Finney, 1952).

The speech stimulus was presented in the same ear (right or left) for both the test and retest conditions. The ear in which speech stimuli was presented were randomly selected prior to each participant beginning data collection. The SRT was found for the right ear in 21 participants and for the left ear in 19 participants. The participant's SRT was determined in the

test condition and then a 10 minute break was provided. The SRT was again established for the retest condition in order to determine test-retest reliability.

Results

The mean age, mean PTA, and mean 2PTA were determined for the study's participants. The data for each SRT test and SRT retest measurements were calculated and averaged. The averaged SRT measurement for the test condition resulted in 22.7 dB HL and 22.8 dB HL for the retest condition. The mean SRT measured during retest averaged 0.1 dB better than the test SRT. The descriptive statistics are presented in Table 2.

Table 2

Descriptive Statistics for the Participant's Age, Speech Recognition Threshold (SRT), Pure-Tone Average (PTA), Two-Frequency Pure-Tone Average (2PTA) and Differences Between Means.

	M	*SD*	Minimum	Maximum
Age (years)	53.4	18.4	19.0	83.0
PTA	31.2	20.3	-5.0	71.7
2PTA	24.4	19.6	-7.5	60.0
SRT test	22.7	17.6	0.0	62.0
SRT retest	22.8	18.2	0.0	59.0
Test-retest difference	-0.1	3.7	-9.0	9.0
Difference				
PTA - SRT test	8.5	8.2	-5.0	25.7
PTA - SRT retest	8.4	7.1	-6.0	25.7
2PTA - SRT test	1.7	4.6	-8.0	13.0
2PTA - SRT retest	1.6	4.2	-8.50	11.0

Note. The SRT, PTA, 2PTA, and test differences are all dB HL

A paired *t*-test revealed that the difference between SRT test and retest was not significant; $t\,(39) = 0.215$; $p = .831$. To measure validity of the new SRT words, a variability equation was adapted (Shavelson & Webb, 1991) in order to analyze test-retest reliability. The original variability equation which was modified for this study is as follows:

$$P'_{xx} = \frac{\sigma_p^2}{\sigma_p^2 + \frac{[\sigma_{pi,e}^2]}{n_i'}} \quad (1)$$

Equation 1 was adapted to compare the variance within subjects with the variance between subjects. The SAS 9.3 software was used to calculate the mathematical estimation model (proc varcomp) for estimating the variance. The test-retest reliability can be determined by calculating the variance within subjects and the variance between subjects then inserting into the equation as follows:

$$\text{Reliability} = 1 - \frac{\text{variance within subjects}}{\text{variance between subjects}} \quad (2)$$

Using Equation 2 the test-retest reliability was determined to be 0.98. Due to the large variability among the participants the test-retest reliability of the SRT measure with the 33 words was expected.

One of the purposes for determining an individual's SRT in a hearing evaluation is to validate their PTA. Therefore it was important that the participant's SRT was compared to their PTA. Also due to the participants differing hearing abilities the pure-tone audiograms demonstrated a variety of shapes of hearing loss, including sloping hearing loss. A 2PTA was determined for all participants as suggested by ASHA (1988) to control for the effects of a

sloping hearing loss (Fletcher, 1950). Both the 2PTA and PTA were compared with the SRT test and retest measure. The PTA averaged 8.5 dB better than the SRT test measurement and the PTA averaged 8.4 dB better than the SRT retest measurement. When the 2PTA was compared with the SRT, the 2PTA resulted in an average 1.7 dB better than the SRT test measurement and averaged 1.6 dB HL better for the retest measurement.

The Pearson correlations of the 40 participants PTA was compared with the SRT test measurement resulting in a strong correlation $r(38) = .92$, $p < .001$ and a comparison of the individuals PTA with SRT retest measurement resulted in a $r(38) = .94$, $p < .001$ correlation. Also correlations were determined for the participants 2PTA with the SRT test and retest conditions resulting in $r(38) = .98$, $p < .001$ correlations for test and retest conditions. The strong correlations of the 2PTA and SRT indicate that the SRT words are a reliable measurement for speech audiometry.

Discussion

The purpose of the study was to determine test-retest reliability of the 33 spondaic words developed by Chipman (2003) used for measuring individual's SRT. As Ostergard (1983) had suggested there is a high amount of variability in speech audiometry, Chipman attempted to control variability by selecting familiar words that were homogeneous with regards to audibility with each word having equal thresholds and with respect to having steep psychometric function slopes. The SRT words were then digitally recorded to provide a standardized method of presentation to control for variability. In the present study the goal was to control variability by insuring that the 33 words were a reliable SRT measure.

Jacobs (2012) attempted to demonstrate test-retest reliability of the 33 spondaic words; however, the investigation resulted in poor reliability of 0.47. The participant's hearing ability

was the significant difference between Jacobs and the current research. In the present investigation hearing abilities ranged between normal to severe sensorineural hearing loss, and in Jacobs study all participants presented with normal hearing . The participants in the current study were selected based on their wide range of hearing impairments to increase variability. The individuals in this investigation average PTA was 31.2 dB HL with hearing range of -5.0 to 71.6 dB HL, whereas Jacobs average PTA was 1.8 dB HL with a hearing range of -6.7 to 10.0 dB HL. The greater diversity of the participant's PTA improved the test-retest reliability score from Jacobs data of 0.47 to the present data of 0.98 or in other words from poor reliability to almost perfect reliability.

The selection of participants with wide range of hearing impairments also improved the mean test-retest SRT measurement as compared to Jacobs (2012). The mean SRT measurement improved 0.1 dB from test-retest with a standard deviation of 3.7 indicating low variability between the two measurements with no significant difference. The mean SRT measurement for Jacob's normal hearing participants improved 1.4 dB for male talker and 1.2 dB for female talker from test to retest. These differences between test and retest were determined to be significant although within clinically acceptable range. Therefore in the current investigation the 33 words demonstrated better reliability across measurements and in general, proved to be very reliable for measuring SRT.

The current investigation, similar to Jacobs (2012) examined the agreement of the SRT and the individual's PTA. Jacobs averaged the test and retest SRT, for female talker the mean SRT averaged 2.0 dB better than the PTA and for male talker the mean SRT averaged 2.3 dB better than the PTA for normal hearing individuals. In the current study, the individual's PTA and mean SRT did not demonstrate a very good agreement with averaging 8.4 dB and 8.5 dB for

test and retest, respectively. However, since the participants had a wide range of hearing impairments, many presented with a sloping hearing loss and a 2PTA was a more accurate measurement of participant's hearing and demonstrated a better agreement with the individual's SRT. The mean 2PTA averaged 1.7 dB better than test SRT and averaged 1.6 dB better than retest SRT for individuals with a wide range of hearing impairments. The differences between SRT and 2PTA were well within the acceptable range of 0.3 – 3.1 dB (ASHA, 1988). The 33 words proved to be a valid and precise measurement for SRT.

This study identified a strong correlation of the individuals PTA with the SRT test measurements and the SRT retest measurements of 0.92 and 0.94, respective, and a stronger correlation was found when using the 2PTA with SRT test and retest of 0.98. The correlation coefficients of 0.92 for PTA and SRT test and 0.94 for PTA and SRT retest are not as strong as reported by ASHA (0.95 - 0.98) however the 2PTA correlations with SRT test and retest (0.98) are extremely strong and well within the ASHA guidelines (1988). However, it should be noted that the participants presented with greater hearing variability and the increased variability between subjects would contribute to the strong correlation between PTA and SRT.

As compared to Jacobs (2012), the present investigation was able to demonstrate that the 33 spondaic words developed for SRT are very valid and reliable measures. Jacobs presented good test-retest reliability based on the mean SRT test-retest difference and the good agreement between SRT and individual's PTA; however achieved poor test-retest reliability results with the mathematical model for normal hearing participants. Using participants with increased variability of hearing impairments the current study found exceptional test-retest reliability based on both the mathematical model and SRT test-retest difference. The SRT words were able to

validate the individual's PTA and proved to have extremely good correlation between PTA and SRT.

These words can be used clinically with a high degree of confidence that the same results will be obtained across measurements and testing situations. It is important that SRT measurements are reliable in a clinical setting. The SRT measure is used clinically most often to validate the individual's PTA but it is also used to measure threshold, evaluate hearing aid performance, assess suprathreshold intelligibility, a tool for otologic surgery, aid in peripheral and central auditory diagnosis, and in rehabilitation research (ASHA, 1988; Jerger, Speaks, & Trammell, 1968). Clinicians need to be able to rely on SRT measures to be consistent and reliable to evaluate and assess hearing impaired patients. The 33 words developed to measure SRT demonstrated exceptional test-retest reliability. The SRT words can be used with a high degree of confidence that the same measurement will be obtained from one testing to the next. Also the SRT words were excellent measures that verified the individual's PTA.

Conclusions

In summary, the purpose of the study was to examine the test-retest reliability of the 33 spondaic words developed by Chipman (2003). The SRT for both test and retest condition was determined for 40 participants with a wide range of hearing impairments. The investigation found very good test-retest reliability (.98) between the two SRT measurements. Also the individual's 2PTA was in good agreement with the SRT measure.

Many important factors were identified in this investigation. First, the 33 spondaic words are reliable measures for determining an individual's SRT. In the present study the participants were chosen based on their wide range of hearing impairments. In order to achieve good test-retest reliability, a wide range of variability was needed among the participants. The increased

variability of hearing (normal to severe hearing loss) of the participants produced exceptional test-retest reliability, thus proving that the spondaic words will produce reliable and dependable SRT measurements.

Second, the 33 spondaic words are valid speech stimuli for determining the SRT. The SRT measurements demonstrate high correlation with the individual's 2PTA, as well as other developed SRT materials and are well within the acceptable range of 0.3 – 3.1 dB (ASHA, 1988). Also Chipman (2003) followed the appropriate criteria for word selection and homogeneity with regards to audibility and psychometric function slope. The 33 words have demonstrated to be a very good and valid measurement for determining an individual's SRT.

Third, the PTA is not necessarily the best average to use with participants with hearing loss. Although PTA may be the most widely used method for averaging hearing loss, it is not the only method for establishing a pure-tone average (ASHA, 1988). The participant's shape of hearing loss needs to be considered for determining the best method for identifying the pure-tone average of the individuals. Individuals with hearing loss often have sloping loss at high frequencies and the 2PTA would provide a better representation of the individuals overall hearing and may provide a better agreement with the SRT measurement (Fletcher, 1950). This was the case in the current study, the participant's PTA and SRT were not in close agreement; however when the 2PTA was compared with the SRT a very good agreement could be established.

Lastly, when determining test-retest reliability of a test or a measure it is essential to have participants with a large amount of variability. In this study, participants with a wide range of hearing impairments were needed to increase variability but other factors may be needed to increase variability for other research. When variability was increased among participants, all

other correlations and measurements also improved in the current investigation in part due to the increased variability.

Future research may include SRT measurements in noise to identify how well an individual can communicate in normal everyday situations. Also the use of the female recording (Chipman, 2013) could be used with a wide range of hearing impaired individuals to identify if a different talker would have an effect on the SRT measurement.

References

American National Standards Institute. (1991). *Maximum permissible ambient noise levels for audiometric test rooms*. ANSI S3.1-1991. New York: ANSI.

American National Standards Institute. (2004). *Methods for manual pure-tone threshold audiometry.* ANSI S3.21-2004. New York: ANSI

American National Standards Institute. (2010). *Specification for audiometers.* ANSI S3.6-2010. New York: ANSI.

American Speech-Language Hearing Association. (1988). Guidelines for determining threshold level for speech. *American Speech-Language Hearing Association, 30*(3), 85-89. doi: 10.1044/policy GL 1988-00008

Chipman, S. (2003). *Psychometrically equivalent English spondaic words* (Unpublished master's thesis). Brigham Young University, Provo, Utah.

Clark, J. G. (1981). Uses and abuses of hearing loss classification. *American Speech-Language and Hearing Association, 23*(7), 493-500.

Di Berardino, F., Tognola, G., Paglialonga, A., Alpini, D., Grandori, F., & Cesarani, A. (2010). Influence of compact disk recording protocols on reliability and comparability of speech audiometry outcomes: Acoustic analysis. *Journal of Laryngology and Otology 124*(8), 859-863. doi: 10.1017/S0022215110000782

Egan, J. P. (1948). Articulation testing methods. *The Laryngoscope, 58*(9), 955-91.

Finney, D. J. (1952). *Statistical method in biological assay*. London: C. Griffen.

Fletcher, H. (1950). A method of calculating hearing loss from an audiogram. *Acta Oto-laryngologica Supplementum, 90*, 26-37.

Francis, W. N., & Kučera, H. (1982). *Frequency analysis of English usage: lexicon and grammar*. Boston: Houghton Mifflin Company.

Goodman, A. C. (1965). Reference zero levels for pure-tone audiometer. *American Speech-Language and Hearing Association, 75*(7), 262-263.

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,* 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.

Hudgins, C. V., Hawkins, J. E., Karlin, J. E., & Stevens, S. S. (1947). The development of recorded auditory tests for measuring hearing loss for speech. *The Laryngoscope, 57*, 57-89.

Hundt, M., Sand, A., & Skandera, P. (1999). Manual of information to accompany the Freiburg-Brown Corpus of American English ('Frown'), from http://khnt.hit.uib.no/icame/manuals/frown/INDEX.HTM

Jacobs, A. M. (2012). *Test-retest reliability in the determination of the speech recognition threshold* (Unpublished master's thesis). Brigham Young University, Provo, Utah.

Jerger, J., Speaks, C., & Trammell, J. (1968). A new approach to speech audiometry. *The Journal of Speech and Hearing Disorders, 33*(4), 318-328.

Kamm, C., Carterette, E. C., Morgan, D. E., & Dirks, D. D. (1980). Use of digitized speech materials in audiological research. *Journal of Speech and Hearing Research,* 23, 709-21.

Kreul, E. J., Bell, D. W., & Nixon, J. C. (1969). Factors affecting speech discrimination test difficulty. *Journal of Speech and Hearing Research, 12*(2), 281-287.

Martin, F. N., Champlin, C. A., & Chambers, J. A. (1998). Seventh survey of audiometric practices in the United States. *Journal of the American Academy of Audiology, 9*(2), 95-104.

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.

Mendel, L. L., & Owen, S. R. (2011). A study of recorded versus live voice word recognition. *International Journal of Audiology, 50*(10), 688-693.

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,* 449-58. doi:10,1080/1499202070136129

Ostergard, C. A. (1983). Factors influencing the validity and reliability of speech audiometry. *Seminars in Hearing, 4*(3), 221-239.

Raffin, M. J., & Thornton, A. R. (1980). Confidence levels for differences between speech-discrimination scores. A research note. *Journal of Speech and Hearing Research, 23*(1), 5-18.

Shavelson, R., & Webb, M. (1991). *Generalizability theory: A primer*. California: Sage Publications, Inc.

Thornton, A., & Raffin, M. J. (1978). Speech-discrimination scores modeling as a binomial variable. *Journal of Speech and Hearing Research, 21*, 507-518.

Tillman, T. W., & Jerger, J. F. (1959). Some factors affecting the spondee threshold in normal hearing subjects. *Journal of Speech and Hearing Research, 2*, 141-146.

Wilson, R. H., Morgan, D. E., & Dirks, D. D. (1973). A proposed SRT procedure and its statistical precedent. *Journal of Speech and Hearing Disorders, 38*(2), 184-191.

Wilson, R. H., & Strouse, A. (1999). Psychometrically equivalent spondaic words spoken by a female speaker. *Journal of Speech, Language, and Hearing Research, 42*(6), 1336-1346.

Appendix A

Annotated Bibliography

American National Standards Institute. (1991). *Maximum permissible ambient noise levels for audiometric test rooms.* ANSI S3.1-1991. New York: ANSI.

Purpose of work: The purpose of this work was to set the standards for maximum permissible ambient noise levels while performing any kind of audiometric test. These standards will insure that ambient noise does not distort the hearing test results.
Summary: ANSI outlines the acceptable measurements of ambient noise levels under different conditions. It defines all necessary terms associated with hearing test. It describes the conditions: insert earphones, ears covered, ears not covered and the acceptable ambient noise level in each test frequency range. It describes the proper measurement instrumentation that should be used in testing. It also describes how the instruments should be used and in what conditions.
Conclusions: All persons conducting or performing audiometric test should comply with these standards to insure reliable results.
Relevance to the current work: This work was referenced to insure compliance with proper specifications of ambient noise levels used in the current research.

American National Standards Institute. (2004). *Methods for manual pure-tone threshold audiometry.* ANSI S3.21-2004. New York: ANSI

Purpose of work: The purpose of the current paper is to outline the proper procedures for conducting pure-tone threshold audiometry. It provides standards for pure-tone audiometry.
Summary: The manual defines terms associated with pure-tone testing. It explains general requirements of ear canal, earphone placement, instructions, response, and interpretation of responses when testing. It outlines the procedures for determining thresholds, the standards for air conduction, and bone conduction measures. It presents research about variability of threshold measures and provides reliability of pure-tone measures.
Conclusions: The paper presents standards for the proper procedures in determining pure-tone thresholds that should be followed in evaluating hearing.
Relevance to the current work: ANSI provides information about the reliability of pure-tone measures. It presents research and standards of +/-5 dB consistency in pure-tone testing.

American National Standards Institute. (2010). *Specification for audiometers.* ANSI S3.6-2010. New York: ANSI.

Purpose of work: The purpose is to provide a standard of specification and tolerances for audiometers to ensure hearing test will be reliable across different settings.
Summary: ANSI defines relevant terms and identifies the types of equipment used in hearing test. It outlines requirements for specific types of audiometers. It explains proper

set up and use of the equipment. It explains how to deal with and test for unwanted sounds in different conditions (earphone, bone vibrator). ANSI reports required frequencies and hearing levels for various audiometers. It explains specifications, measurements, and maximal permissible harmonic distortions. ANSI describes details for using audiometers with speech material and specification for masking sounds. It describes signal level controls, tone switching, reference signal facilities, transducers specifications, requirements for device marking, and instruction manuals.
Conclusions: Compliance to all requirements for audiometers should be followed for reliable hearing evaluations.
Relevance to the current work: ANSI identifies the appropriate specifications and calibrations for the audiometer used in the current study

American Speech-Language Hearing Association. (1988). Guidelines for determining threshold level for speech. *ASHA*, 30, 85-89. doi:10.1044/policy.GL 1988-00008

Purpose of work: The purpose of this work is to set guidelines for speech testing. These guidelines supersede the guidelines of 1979. There were many concerns about the guidelines of 1979 due to lack of evidence based practice, time-consuming procedures, and lack of defined procedures. These new guidelines define the terminology and procedures that are supported by relevant research.
Summary: This study identifies the purpose of the speech recognition threshold as an accepted procedure to validate the pure-tone average to rule out pseudohypacusis. It defines speech threshold terms and how the terms differ from each other. It outlines proper procedures in obtaining the SRT. It describes the acceptable and preferred environment, instrumentation, materials, and responses. It defines recorded vs. live voice presentation of the test material and explains some of the possible problems with each method. It describes the procedures of how to give instructions and familiarization of the test list. It discusses the way to present each word in different testing conditions. It also outlines the way to calculate the speech threshold using ether 2 dB or 5 dB increments.
Conclusions: These guidelines are updated with current research for acceptable practice in determining speech thresholds. This current work establishes further reliability and validity to support these recommendations.
Relevance to the current work: ASHA defines SRT testing and the reasons for the testing. It also outlines why recorded presentation of materials are the preferred method. It also outlines the procedure used in the current study for determining speech thresholds.

American Speech-Language-Hearing Association. (1997). Guidelines for audiologic screening. *ASHA*, doi:10.1044/policy.GL 1997-00199

Purpose of work: The purpose of this work is to consolidate the current research, technology, and audiologic screening guidelines for all ages in a single document.
Summary: The study outlines screening, performance, definitions, and framework for hearing screenings. It identifies the development, management, and proper procedures for a screening test. It also presents the needed criteria to pass a screening test and in what case to refer for a full hearing evaluation. The ASHA describes the guidelines for hearing screenings that cover from pediatric to all ages over the life span. It presents rationale and appropriate personnel for audiologic screenings. It discusses the proper

procedures for conducting screenings with children for outer and middle ear disorders and the criteria for passing tympanogram or referring for a full hearing evaluation. It describes the recommended procedures to measure infants hearing. It outlines the way to use auditory brainstem response or the evoked otoacoustic emission to screen infants hearing. The paper outlines all procedures for conducting hearing screenings for all age levels from infant through adulthood.
Conclusions: The paper presents guidelines and procedures for conducting audiologic screenings based on current research.
Relevance to current work: This paper provides background knowledge for hearing screenings and for understanding appropriate recommended procedures.

Chipman, S. (2003). *Psychometrically equivalent English spondaic words* (Unpublished master's thesis). Brigham Young University, Provo, Utah.

Purpose of the study: The pupose of this work is to create a list of digitly recorded SRT words that meet the criteria as recommended by Hudgins,et al. (1947) in that the spondaic words are more homogeneous and more familiar than the 36 CID W-1 lists.
Method:Twenty normal hearing subjects between the ages of 18 and 23 participated in this study. All subjects were English speaking individuals. Male and female talkers jugded to have good vocal quality and accents were selected to record a list of 98 spondaic words more frequently used in todays English language. Twenty subjects listened to the 98 words at 13 different intensity levels.
Results: The data were analized by using the logistic regression equation used to find the regression slope and intercept. The pecent correct then could be predicted at any specified intensity level. Thirty three words with the steepest psychometirc function were identified. To improve homogeneity of the 33 words the intensity was digitaly adjusted to correspond with the individual's PTA. An ANOVA was used to compare the word slopes at the 50% point of the 33 recorded words with the recorded 36 CID W-1 words. A significance difference in slopes were found between the words with the 33 recorded words having a steeper slope average than the CID W-1 words.
Conclusions: The list of 33 words were more frequently used and therefore more familiar. Also these recordings of words demonstrated steeper slopes and greater homogeneity.
Relevance to the current work: The purpose of the current study is to determine test-retest reliablity of these 33 spondaic words.

Clark, J. G. (1981). Uses and abuses of hearing loss classification. *American Speech-Language and Hearing Association,*23(7), 493-500.

Purpose of work: The purpose of this paper is to describe the history of hearing loss classifications and the appropriate use for the different types of classifications.
Summary: Over the years many different types of hearing classifications have been developed to describe hearing loss. Different methods for classifing hearing loss have been devised for medicolegal, clinical, educational, and research. Percentage classifications were developed for medicolegal use however from 1922 until the current time this method has been revised multiple times. There has been several limitations noted with percentage classification. Many versions tend to exclued high frequency, low

frequency hearing loss, or do not include normal hearing. This system continues to have flaws and is not always used appropriately by audiologists and physicians. Another method of describing hearing loss is the use of adjective descriptors. Most descriptors are based on the average pure-tone thresholds which tend to ignore high frequency losses. Also many descriptor methods identify hearing loss starting at 25 dB HL when research has indicated that a slight loss begins at 15 dB HL which can affect childrens development of speech. It is recommended that *slight loss* be included in the classification system. The problems are a lack of standardization in hearing loss labels. This can cause confusion when different professionals use different labels for the same hearing loss. Additional classifications systems have been developed by Risberg and Martony (audiogram classification), Carhart (deshon classification), and Pearson, Kell, and Taylor (index of hearing impariment). Labels are limited and more factors need to be examine to understand a individuals hearing loss limitations.
Conclusions: Whatever classifications used for individuals, it is important that an understanding of the individuals hearing loss and abilities are identified and described in several different ways.
Relevance to current work: The classification of hearing loss (Scale of Hearing Impairment) presented in this paper was used for the current study to describe participants hearing abilities.

Di Berardino, F., Tognola, G., Paglialonga, A., Alpini, D., Grandori, F., & Cesarani, A. (2010). Influence of compact disk recording protocols on reliability and comparability of speech audiometry outcomes: Acoustic analysis. *Journal of Laryngology and Otology 124*(8), 859-863. doi:10.1017/S0022215110000782

Pupose of the study: The pupose of the study was to examine the different protocols used in CD recording for speech materials to assess reliability of outcomes. The acoustic analysis was measured and compared.
Method: Four different CDs were selected that are used clinically in Italy. One list of words from each CD and the CDs calibration signal were acousticly analysed by measuring the RCA analogue output and by measuring the sound pressure level over time. Psychoacoustic evaluation was also used by finding the psychometric curves for the spondee list of the four CDs for 12 normal hearing participants.
Results: The calibration signals and the recording level of the speech stimuli were measured for the four CDs and no CD demonstrated equal levels. For CD 3 and CD 4 the calibration signal was recorded at a lower level than the speech material and for CD 1 and CD 2 the speech material was recorded at a lower level than the calibration signal resulting in lower recongnition scores for CD 1 and CD 2 than CD 3 and CD 4. The Friedman test was used to determine differences among SRT and maximum intelligibility thresholds of the CDs. A significant difference was found and the Wilcoxon signed-rank post-hoc analysis indicated that CD 1 and CD 2 lead to higher SRT and maximum intelligibility thresholds than CD 3 and CD 4.
Conclusions: The study found that different CD recordings of speech material leads to significantly different speech thresholds due to the lack of standards for equalization of speech test material. There is a need to specify the protocols taken in producing the CD to ensure reliability and consistency of the results.

Relevance to the current work: The article documents the importance of using recorded materials in SRT testing and the importance of standardizing the presentation to insure reliability and consistency across testing.

Egan, J. P. (1948). Articulation testing methods. *The Laryngoscope, 58*(9), 955-91.

Purpose of work: The purpose of this paper was to identify and explain the factors that affect articulation testing. It is proposed that both the test items and the procedures used in testing affect the scores obtained in speech testing.
Summary: Articulation materials can consist of nonsense syllables, monosyllabic words, polysyllabic words, or sentences. It is important for the speech materials to be representative of all speech sounds used in everyday language. The speech materials should also be sensitive measuring instruments with an appropriate distribution of difficulty. Egan also outlines the purposes, advantages, and disadvantages for each type of speech materials. Factors that can affect articulation testing are the procedures, environment, and equipment used in testing. The articulation scores can be affected by announcers, microphones, amplifiers, earphones, noises, listeners, and test lists all. It is important to be aware of these factors and take steps to control for them in order to achieve the most reliable results. Also the author describes and identifies the purpose for determining the thresholds for speech.
Conclusions: The articulation testing methods are described and the purposes of each test are explained. The author sets forth factors that can affect articulation testing and identifies ways to solve and control for them in order to achieve reliable results.
Relevance to the current work: The paper describes that spondaic words are the highest homogeneous words in English and are the most appropriate word for SRT testing.

Finney, D.J. (1952). *Statistical method in biological assay.* London: C. Griffen.

The purpose of work: The purpose of this book was to use biological assay with experimental techniques and to use measurements for comparing the potencies of treatments.
Summary: The book provides formulas and measurement for analyzing data for a variety of different quantitative biological assays.
Relevance to the current work: The book provided information and reasoning for the 1 dB correction needed in calculating the individuals SRT.

Fletcher, H. (1950). A method of calculating hearing loss from an audiogram. *Acta Oto-laryngologica Supplementum, 90*, 26-37.

The purpose of work: The purpose of the paper is to present a new formula that is more accurate then calculating hearing from averaging hearing ability from 500, 1000, and 2000 Hz pure-tone thresholds.
Summary: The formula is presented and described for calculating hearing loss. The equations are also identified and explained. Comparisons of calculating hearing loss are compared and it is suggested that taking the 2 best values at 500, 1000, 2000 Hz would be more accurate for determining hearing averages.

Conclusions: Averaging hearing ability from 500, 1000, and 2000 Hz is not the most accurate way for identifying hearing loss.
Relevance to the current work: The paper presents the procedure of identifying hearing loss through finding the 2 best values from 500, 1000, and 2000 Hz. This procedure was used due to the variety of shapes of hearing loss in our participants for comparing with individual's SRT.

Francis, W. N., & Kučera, H. (1982). *Frequency analysis of English usage: Lexicon and grammar*. Boston: Houghton Mifflin Company.

Pupose of work: The work compiles a standard corpus of present-day American English.
Summary: A list of over a million words were compled in the computer data base between 1963 and 1964 at Brown University. It is commonly know as the Brown Corpus and is published in over 15 genres.
Relevance to the current work: Chipman (2003) used this resource to identify familiar words used in the updated SRT list. It was also used to compare the CID W-1 list to identify words not commonly used in English. The purpose of the current study is to determine test-retest reliability of the words selected by Chipman.

Goodman, A. C. (1965). Reference zero levels for pure-tone audiometer. *American Speech-Language and Hearing Association, 75*(7), 262-263.

Pupose of work: The purpose of this paper was to clarify the new 1964 ISO reference zero.
Summary: The article explained the new standards for calibration of pure-tone audiometers and the differing results that could affect the persons classification of hearing. The new standards were more sensitive to a hearing difference than the 1952 ASA reference and a new classification of hearing impairment scale was needed
Conclusions: A new scale of hearing impairment was established and general guidelines were provided for interpreting results.
Relevance to the current work: The scale of hearing impairment was used to identify participants with different hearing abilities and ensure a good distribution of hearing impairments were representative in our study.

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,* 47-66. doi:10.1080/14992020601058117

Purpose of the study: The purpose of this study was to develop speech materials in Russian. The goal was to make digital recordings of both word recognition list and SRT words for use in audiological evaluations.
Method: Twenty native Russian subjects from two regions of Russia participated in the study. They were between the ages of 16 and 50 and all had hearing within normal limits. The word lists were 280 monosyllabic frequently used words, 30 of the words were eliminated. Seventy bisyllabic initially stressed words were identified for use in

SRT testing. The initial recording was made by 13 (six male and seven female) native Russians. Six subjects (two male and four female) evaluated each talker for best pronunciation, vocal quality, and standard dialect. The highest ranked female and male were identified for future recordings. The speakers were asked to say each word four times for recording. The best production of the word was included in the list of words. The subjects participated in three sessions where SRT stimuli were presented in the first two sessions and in the third session they listened to word recognition stimuli. For the word recognition stimuli, 250 monosyllabic words were randomly put in to lists of 25 words each and presented to half the listeners. They were listened to at 10 levels between -5 and 40 dB HL. They were then randomly regrouped and presented to the remaining half of the listeners at the same 10 levels. For the SRT stimuli, 70 bisyllabic words were presented to all listeners at 13 intensity levels. The listener was asked to repeat the perceived word.

Results: Monosyllabic word recognition: the words were ranked according to the successful identification by the listeners. The top 200 perceptible words were selected to be included in four lists (50 words each) and eight half-list (25 words each). The 200 ranked-ordered words were assigned to four lists. They were counterbalanced to insure each lists contained equal difficulty. The eight half list were created the same way as the four full lists of words. To find the psychometric functions of each list the logistic regression was calculated. Then the values of the regression slope and intercept were used in a modified logistic regression equation to calculate the percent correct at each intensity level (-8 to 40 dB HL in 2 dB increments). The threshold slope at 50%, and slope from 20% to 80% intelligibility were determined for each lists and half lists. A chi-square was performed to ascertain if any significant differences exists between the lists and half lists. No significant differences were identified. Small intensity adjustments were made to the lists to equate performances. Bisyllabic SRT words: the psychometric functions were calculated using logistic regression for each word. Then the values of the regression slope and intercept were used in a modified logistic regression equation to calculate the percentage of correct recognition at each intensity level (-10 to 18 dB HL in 1 dB increments). The threshold slope at 50% and the slope from 20% to 80% intelligibility were determined. The top 25 words with homogeneity and steepest psychometric function slopes were selected to be included in the final recording. To decrease variability the intensity of each word was adjusted so that 50% threshold of the words would match the mean PTA of the subjects.

Conclusions: The researchers were able to develop homogenous word list for measuring word recognition and were able develop SRT materials. These are familiar and are homogenous material for audiological evaluation in Russian.

Relevance to work: The article outlines the need for and appropriate procedures for developing familiar speech materials for all individuals. This study provided an understanding for calculating the psychometric function slopes to achieve homogeneity.

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.

Purpose of work: The purpose of this study was to develop CID Auditory Test W-2, CID Auditory Test W-1, and CID Auditory Test W-22 to imporve words and presentation

of speech materials for hearing evaluations in clinical use. The sudy presented words that were more familiar and phonetically balanced with the use of magnetic tape recordings.
Method: For the development of CID Auditory Test W-1 84 spondee words were selected from PAL Test No. 9 that were judged to be more familiar. Six listener with normal hearing listened to the words and 36 spondee words with equally intelligiblity were identified and recorded on tape. Six word orders were recorded and then both inexperienced and experienced listeners listen to words and the words were equated by + 2 dB for difficult words and - 2 dB for easy words. CID Auditory Test W-2 was developed by using the 36 spondee word list of W-1 with the rate of attenuation of 3 dB every three words for a faster pace of estimating the threshold of intelligibility. Six experienced listeners were selected to listen to the six lists of words to determine if any differences in difficulty among the lists were identified. CID Auditory Test W-22 consisted of four lists of 50 monosyllabic words. Words were selceted based on sylables, familiarity, and phonetic composition. Three groups of five participants listened to all 24 list that were recorded on tape.
Results: The words that were too easy and too dificult were eliminated from the list of words in list W-2. The degree of difficulty of the word was correlated with the intensity of the word. An analysis of variance was measured to determine thresholds obtianed by using diferent word orders and did not find significant differences. In test W-22 the articulation scores were found to be simular for three of the four lists, therefore list 1 intensity was increased to match the other three lists.
Conclusions: Three new speech test were constructed to improve familiarity, phonetic balance of words and list.
Relevance to the current work: Lists were constructed in 1952 and using the same requirements of familiary for constructing a new list of spondiac words was developed for SRT testing. This updated list of words were developed to ensure familiary for hearing evaluations today. It is this list of spondiac words that are being examined for test-retest reliability.

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

Purpose of the study: The purpose of the study was to validate 20 MRC word lists recorded by a professional BBC announcer and to determine if the speaker's voice had larger effect on the articulation curve verses the phonetic balance of the lists of words.
Method: Forty five participants were selected to listen to lists of 25 words produced from the Harvard phonetically balanced lists. The lists were presented six times at 10, 15, 20, 25, 30, and 35 dB HL. All 500 words were presented 36 times to all or some of the participants. The words were shown to be in order of difficulty and by using a x^2 test, it was determined that the recoded words were graded as to difficulty and not just random chance. From the 500 words, 25 easy words were selected and 25 difficult words were selected and the articulation scores were examined. They were treated as lists of words and placed in an articulation curve with the easy words forming a steep slope and the difficult words forming a shallow slope. The words were re-recorded as single lists, and five normal hearing participants listened to the words at 5 dB up to 25 dB HL in 5 dB increments for the easy list and up to 45 dB HL for the difficult list. The words demonstrated the same order of difficulty with the five subjects as they did with the 45

subjects. The two lists of words were re-recorded by two additional speakers and the same procedure repeated with five listeners.
Results: The curves of the lists were clearly different from the easy list to the difficult list with speaker l as the list moved closer together. Speaker 2 and with speaker 3 were almost on top of each other thus changing the difficulties of the lists of words. The phonetic structure, word familiarity, word environment, and inter- and intrasubject variability were examine to determine their effects on word difficulty order.
Conclusion: The study demonstrated that the speaker has more effect on word difficulty order and the articulation curve then the words themselves. It is important to be aware of the effect the speaker has when developing speech audiometry materials.
Relevance to the current work: The work provides reasoning for the use of recorded speech materials due to the effect that different speakers have on the articulation curve and the difficulty of the words for speech audiometry.

Hudgins, C., Hawkins, J., Karlin, J., & Stevens, S. (1947). The development of recorded auditory tests for measuring hearing loss for speech. *Laryngoscope, 57*, 57-89.

Purpose of work: The purpose of this study was to identify problems with the current audiometric tests, to develop test that are more precise measurements for all types of hearing loss, and to differentiate between high frequency deafness and uniform deafness.
Summary: The paper outlines the Western Electric 4C audiometer speech test. This test uses digits and decreases in steps of 3dB per pair of digit with a 33 dB range. It is a coarse screening test that can only be used with limited hearing loss. The 4 important elements for speech test are described. These include familiarity of words selected, words should be phonetically dissimilar, words should be representative of English speech sounds, and words should be homogeneous for audibility. This paper discussed the development of two new speech audiometry tests: Auditory Test No. 9 threshold of hearing for words and Auditory Test No. 12 threshold of hearing for sentences. Test No. 9 consists of two lists of 42 spondee words with words found to have an average slope of 10% decibel between 20 and 80%. The words were recorded in seven groups of six words and each group decreased by 4 dB with a range of 24 dB. The No. 12 test consisted of eight lists of short questions made up of 28 items except for list 1 which contains 21 items. List 1 is divided into seven groups with three items each. The groups were recorded at a decrease of 6 dB with a range 24 dB. List 1 included recorded instructions and is a screener to determine intensity level that all other list should be presented. List 2-8 were divided in to seven groups with four items and recorded at a decreasing intensity of 4 dB with an overall range of 24 dB. To improve homogeneity the sentences levels were adjusted by determining the level that the average sentence was intelligible to a normal listener. The standard error of measurement was determined to be 2 dB with 30 normal hearing participants under good conditions. Test No. 9 and 2.were 4 dB in less than good conditions for 37 normal hearing participants For 70 hearing impaired at four different places the standard error of measurement was found to be from 2.1 to 2.8 dB. Therefore the authors concluded that an individual will achieve a score that falls within 2.8 dB two-thirds of the time. For Test No. 12 two groups of normal hearing participants 16 and 52 individuals obtained standard errors of measurement of 2.8 and 2.2 dB, respectively. Standard errors of measurement for 2 groups of hearing impaired participants 21 and 28 individuals resulted in 2.3 and 1.4 dB. Therefore an

individual will achieve a score that falls within 2.8 dB of the true score two-thirds of the time. The two-thirds of the time restriction are due to the slightly different levels of the different forms of the test.
Results: Two test with the four criteria for audiometry speech test: familiarity, phonetic dissimilar, representation of English, and homogeneity for audibility were developed for clinical use. Instructions were provided for properly administering and scoring the test.
Relevance to the current work: The established criteria for the words selected in speech audiometry were followed in producing a new SRT word list by Chipman. It is this new list of SRT words that will be used to establish test-retest reliability.

Hundt, M., Sand, A., & Skandera, P. (1999). Manual of information to accompany the Freiburg-Brown Corpus of American English ('Frown'), from http://khnt.hit.uib.no/icame/manuals/frown/INDEX.HTM

Purpose of work: The purpose of the manual was to compile a set of corpora that match Brown and LOB corpora with language that is representative of the 1990s.
Relevance to the current work: The manual was used to identify commonly used and familiar words that were included in selecting words for the new SRT list of 33 spondaic words.

Jacobs, A. M. (2012). *Test-retest reliability in the determination of the speech recognition threshold* (Unpublished master's thesis). Brigham Young University, Provo, Utah.

Purpose of the study: The purpose of this study was to establish test-retest reliability for newly developed SRT spondiac words. The list of spondiac words were updated to include more familiar words that were more commonly used in todays English.
Method: The speech stimuli used in this study were developed by Chipman (2003). Fourty participants with normal hearing listened to the speech stimuli by male and female talkers. In the testing condition the participants SRT was determined for each ear for both the male and female talker (four SRTs) and then in the retest condition the SRTs (four SRT) were again found. The study used a randomized block design for ears and talkers for both the test and retest condition.
Results: A modified variance equation was used to determine test and retest reliability by calculating variance of within subjects with variance between subjects. The reliability was determined to be poor at 0.47. However when examining the scores between test and retest there was only an average 1.4 dB (male talker) and 1.2 dB (female talker) improvement in the retest condition which is clinically acceptable. The SRT data was compared to the individual's PTA and a *t*-test determined a difference; however, it was within the margin of error that is clinically accepted. No significant difference was found between the male and female talkers.
Conclusions: Overall test-retest reliability was found to be good for the 33 spondiac words. For the mathematical calcultions the reliability was poor due to similarity of the participants hearing. More variability in hearing ability is needed to achieve a greater reliability between test and retest conditions.
Relevance to current work: This study is being repeated for the current study with participants that have a wide variety of hearing abilities to increase between subject variance and achieve a greater test-retest reliability.

Jerger, J., Speaks, C., & Trammell, J. (1968). A new approach to speech audiometry. *The Journal of Speech and Hearing Disorders, 33*(4), 318-328.

Purpose of the study: The purpose of the study was to develop new speech materials and new test procedures to eliminate problems of responses and prior knowledge.
Methods: The materials used were synthetic sentences. They contained components of real sentences but for the sequence of words. The sentences were homogeneous as they all contained seven words and were based on the word-triplet rule (word dependence on surrounding words). The participants were asked to identify the message from a closed set of response alternatives instead of repeating what was heard. The study compared the sentence identification task with the conventional word list. Over 150 patients with hearing loss listen to both conventional word list and the synthetic sentences.
Results: It was determined that the synthetic sentences achieved the equivalent results as the traditionally used word list.
Conclusions: The synthetic sentences have many benefits over the traditional word list. The procedures for the sentences used a closed set which eliminates prior background and familiarity of language. Other benefits included unambiguous scoring is eliminated as a source of error, sentences have greater face validity, and generating equivalent forms in easily completed.
Relevance to the current work: The article outlines and describes the purposes of speech audiometry.

Kamm, C., Carterette, E. C., Morgan, D. E., & Dirks, D. D. (1980). Use of digitized speech materials in audiological research. *Journal of Speech and Hearing Research,* 23, 709-21.

Purpose of the work: The purpose of this paper was to describe the advantages of using digital recordings for audiological research.
Summary: The paper attempts to explain the technical issues of the digital signal processing. It describes the process of converting analog to digital signal. Digital recordings have superior signal-to-noise ratio when compared to amplifiers or tape recordings. Also when information is stored in digital form it can be changed without any loss of information. When speech stimuli are recorded using digital signal the stimuli can be easily reproduced, manipulated, and stored. The speech materials can be speeded up, slowed down, mixed, edited, filtered, generated, and timed with digitization. This improves the quality of speech materials used in research.
Conclusions: Digital recordings of audiologic material can reduce distortions of speech material, addapted and changed easily, and increase efficiency.
Relevance to the current work: The study provided evidence for the preferred method of presenting speech stimuli in audiological research.

Kreul, E. J., Bell, D. W., & Nixon, J. C. (1969). Factors affecting speech discrimination test difficulty. *Journal of Speech and Hearing Research, 12*(2), 281-287.

Purpose of the study: The purpose of this study was to examine possible factors that affect speech discrimination. These factors include carrier phrase, different noise levels, different talkers, and repetition of speech stimuli.

Methods: Two talkers were selected to produce recordings of the Modified Rhyme Test (300 monosyllabic words in six 50-word list). The recording involved different factors that may affect speech discrimination results. The recordings took place over two sessions, two different carrier phrases were recorded by the talkers, and noise was introduced when presenting list of words to listeners. Twenty-three individuals in three groups, all with normal hearing listened to the recorded list of speech stimuli. All listeners were presented with two different carrier phrase (two carrier phrase 1, two carrier phrase 2) each day. Two groups listened to only talker 1. One group listened to the eight orders of the MRT list by talker 1 at 0 and -10 dB S/N. The other group listened to talker one at -10 and -15 dB S/N and 15 participants heard all eight list by talker 2 at -10 dB S/N. The speech level was constant at 75 dB.
Results: Descriptive statistics were presented and Friedman analysis of variance was used to analyze repeat testing which were not significantly different. It indicated difficulty remained stable. A significant difference was determined with the two carrier phrases. The list difficulty remained with all S/N levels.
Conclusions: The talker and carrier phrase can change test difficulty however in the study the two different recording by talkers were not different. Also noise did not affect the difficulty of the list.
Relevance to current work: The study presented research to identify that in speech discrimination test different talkers can affect speech audiometry results.

Martin, F. N., Champlin, C. A., & Chambers, J. A. (1998). Seventh survey of audiometric practices in the United States. *Journal of the American Academy of Audiology, 9*(2), 95-104.

Purpose of the study: The purpose of the study was to astertain most common audiologic procedures use by audiologist in hearing evaluations. Also to determine if any changes occurred in procedures and practices from 1994.
Method: Five hundred questionnaires with 98 multiple-choice questions were sent to members of the American Academy of Audiology. The members lived in 42 states and were required to be clinically active at least 15 hours a week. Returned surveys included 239 with 218 that met the requirements.
Results: The study identified percentage of audiologist that reported performing tasks of pure-tone testing, speech audiometry, masking in pure-tone and speech, immittance measures, electrophysologic test, central auditory processing disorders, hearing aids, and other hearing procedures. The study found that 99.5 % of repondents obtain SRT using spondaic words and 94% use monitored live voice. About 58% do not familiarize patients with the words prior to testing.
Conclusions: The information provided is a good baseline for other audiologist to compare with their current practices. Provides educators with most commonly used clinical procedures used and aids in teaching correct hearing evaluation procedures.
Relevance to the current work: The study identifies the most commonly used procedures in SRT testing with presentation of words in speech audiometry and endorses the preferred and most reliabile method of using recorded materials.

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.

Purpose of the study: The purpose of this study was to compare a new modified SRT approach with Tillman and Olsen's procedure and to determine if the same results could be obtain in a similar amount of time.
Method: The speech stimuli used in the study were the CID W-1 list of spondaic words. The participants were 20 normal hearing individuals (18 female and 2 male) between the ages of 20 -23 years. The participants were familiarized with the list of words prior to testing. The even number participants were presented with Tillman and Olsen's procedures for obtaining SRT and the odd number were presented with the new modified procedure.
Results: A t-test was used to compare means of Tillman and Olsen's procedure and the modified method demonstrated no significant differences. Also there were no differences in the amount of time required to determine the individual's SRT.
Conclusions: The modified procedure can be used without a prior estimate of the SRT for testing children, hearing evaluations, and organic hearing loss.
Relevance to the current work: The procedures developed and presented in this study for determining the SRT were used in the current study to find the participants SRT.

Mendel, L. L., & Owen, S. R. (2011). A study of recorded versus live voice word recognition. *International Journal of Audiology, 50*(10), 688-693.

Purpose of the study: The purpse of the study was to measure the amount of time needed for presentation of speech materials by way of Monitored live voice (MLV) and recordings.
Method: The participants between the ages of 20 to 80 years were put in to three groups. The two groups of normal hearing listeners consisted of 20 younger listeners in one group and 19 older listeners in group 2. The third group consisted of 20 normal hearing audiologist or Doctor of Audiology students that were the talkers for this study. The NU-6 word recognition list were used for testing. The talkers presented the NU-6 word lists to a participant from group 1 and 2 in three conditions: MLV, short ISI CD recording, and long ISI CD recording. All three administration times were recorded.
Results: The study found that MLV presentation took less time to administer but less than one minute per 50-word list which were not clinically significant.
Conclusions: The authors continued to support recorded materials use to insure less variabilitiy between testing and re-testing.
Relevance to the current work: The article supports using recorded speech materials to improve reliablility all though MLV presentation time maybe a few minutes faster.

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,* 449-58. doi:10,1080/14992020701361296

Purpose of the study: The purpose of this study is to develop SRT testing materials with steep psychometric functions in Taiwan Mandarin for both male and female.

Method: Twenty subjects between the ages of 18-39 years of age (3 male & 17 female) with normal hearing participated in the study. A list of 130 trisyllabic words were identified and three native judges rated them on a scale of 1 to 5 in familiarity (1=extremely familiar and 5=rarely used). Forty-one words were eliminated. Initial recordings were made by three male and three female native Taiwan Mandarin speakers. Eight native speakers evaluated the recordings for pronunciation, vocal quality, and standard dialect. The highest ranked female and male were identified for future recordings. In the recording sessions, the speaker said each word four times and the best production was selected. In two test sessions 89 trisyllabic words were presented at 14 different intensity levels to each listener. The listener was asked to repeat the perceived word.
Results: The psychometric functions were calculated using logistic regression for each word. Then the values of the regression slope and intercept were used in a modified logistic regression equation to calculate the percent correct at each intensity level (-10 to 18 dB HL in 1 dB increments). The threshold slope at 50%, and slope from 20% to 80% intelligibility were determined. The top 28 words with homogeneity and steepest psychometric function slopes were selected to be included in the SRT recording. To decrease variability the intensity of each word was adjusted so that 50% threshold of the words would match the mean PTA of the subjects.
Conclusions: Two lists, male and female, of psychometric equivalent trisyllabic words were developed for SRT testing in Taiwan Mandarin. After adjustments the 28 words proved to be more homogeneous and the mean psychometric function slopes were similar to those in other languages. Future studies should look at how regional dialects and accents affect SRT testing in Mandarin.
Relevance to the current work: The study relates to the current study establishing that SRT is well developed in English with established standards and norms.

Ostergard, C. A. (1983). Factors influencing the validity and reliability of speech audiometry. *Seminars in Hearing, 4*(3), 221-239.

Purpose of work: The purpose of the work is to describe and examine validity and reliability of speech audiometry.
Summary: All tests that measure speech should be constructed to be valid and reliable instruments. Also those that use the measurements should be concerned with validity and reliability of the instruments that are used. Validity is the test measure what it says it measures. There are different types of validity. Content, criterion related, and construct validity are important elements in ensuring the test is measuring what it says it is. Test makers and test users also need to be concerned with reliability of the test. Reliability is the precision of measurement. Test-retest correlation, alternate forms, and internal consistency are all types of reliability. Reliable test will achieve the same results in different testing environments and at different times. Other factors that are important in a test are sensitivity and specificity. A good test needs a good balance of both test sensitivity and specificity however it should be noted when adjusting one the other can be affected. Test sensitivity is the test can identify a difference when a difference exists and test specificity is the test will not identify a difference when no difference exists. Speech audiometry test are concerned with test performance (validity) and communicative function (reliability). Audiologist need to know that the tests are reliable measurements

that actually are measuring hearing and not just the conditions. The binomial model was proposed by Thornton and Raffin to evaluate the obtained scores on speech audiometry test and to judge if the scores are due to chance and estimate the variability. Speech test have a high variability and factors that contribute to the variability needed to be controlled in order to consider the test reliable. Factors that affect reliability in SRT testing are instructions, presentation of stimuli, familiarization, test words, procedures followed, calibration, talker, and transmissions speech are just a few factors.
Conclusions: There are many factors that affect the test validity and reliability and the binomial model can estimate some variability but not all. In order for speech test to be reliable audiologist need to ensure standardization with as many factors as possible.
Relevance to current work: The need for reliable speech audiometry test and to understand all the factors that affect reliability and validity of SRT testing.

Raffin, M. J., & Thornton, A. R. (1980). Confidence levels for differences between speech-discrimination scores: A research note. *Journal of Speech and Hearing Research, 23*(1), 5-18.

Purpose of work: The purpose of this paper was to provide computer generated tables of confidence levels for the probability of difference between speech discrimination scores.
Summary: The tables provide critical differences for 10 item, 25 item, 50 item, and 100 item test with confidence level of 0.05%. The tables also include the probabilities of chance differences for comparing scores obtained on test that have the same item numbers (10 and 10 item) and different item numbers test (10 and 25 item).
Conclusions: The tables provide clinical aplication of the binomial model to estimate variance of speech discrimination scores.
Relevance to the current work: The work provides an understanding of the binomial model and ability to mediate the variance of speech measurements.

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,* 23-28. doi:1059-0889/02/23-28

Purpose of the study: The purpose of the study was to provide familiar stimuli (digit pairs) for SRT testing for new learners of English in the United States. The goal was to determine if digit pairs were more accurate than the accepted stimuli used by the Central Institute for the Deaf CID W-1 for non-native English speakers.
Method: The participants were 12 non native English speakers and 12 native speakers. There were two male and 10 female subjects between the ages of 22 and 69 years of age with normal hearing. The non-native speaker's first languages included Spanish, French, Chinese, Farsi, and Russian. They spoke limited English and lived in the U. S. for less than a year. They had at least a high school education and receptive recognition ability as measured by the Adult-Language Assessment Scales. The stimuli were 56 pairs of numbers between one and nine (excluding seven) with no repetition within a pair (22). A recording of the 56 pairs and 36 CID W-1 items was made by a female American English audiologist. The PTA for all participants was determined and both the 56 pairs and 36

CID W-1 stimuli were presented through insert earphones. They were familiarized with the lists and the stimuli were presented beginning at 20 dB above their PTA.
Results: The researchers looked at two factors; conditions (PTA, D-SRT, and CID-SRT) and group (NE, NNE). A two-factor ANOVA with repeated measures on one factor (conditions) was performed. The researchers also calculated the strength of association and effect size. The most important findings were for the two-factor ANOVA for the group-by-condition interaction. Both groups mean PTA were similar, the D-SRT mean were similar to each other and similar to their mean PTA. In the CID-SRT condition for the native English speakers the means were similar to both their PTA and D-SRT means, however for the non-native English speakers the CID-SRT means were significantly different. A correlation and regression analysis was calculated for the three conditions. They found that there is a high correlation between the D-SRT and the CID-SRT for both groups and that D-SRT accurate for predicting hearing threshold for both groups.
Conclusions: The researchers found that D-SRT is an appropriate stimulus for non-native English speakers for hearing test. Familiar stimuli were found to be better stimuli for corresponding with the individual's PTA than unfamiliar stimuli.
Relevance to the current work: The research supports the need for familiar stimuli in SRT testing.

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.

Purpose of the study: The purpose of the study is to reexamine the tympanometric data norms that were set by ASHA as proposed by Margolis and Heller (1987). It was the aim of the study to determine norms for young adults and control for age and gender.
Methods: The participants were 102 young adults between the age of 20 -30 years with normal hearing and otoscopic findings. For each participant the peak compensated static acoustic admittance, acoustic equivalent volume, and tympanometric width was measured for a randomly selected ear.
Results: The results were compared with the results of the Margolis and Heller (1987) study. A *t*-test was used to compare means and no differences were found for the peak compensated static acoustic admittance. However significant difference was found when comparing the acoustic equivalent volume and tympanometric width. The acoustic equivalent volume was higher in this study and the tympanometric width was significantly smaller. The Komogorov-Smirnov Two Sample Test was also completed to examine distribution of the three tympanometric measures and in the present study the acoustic equivalent volume extended to higher values and the tympanometric width extended to lower values. Gender was also compared with in the study and with the Margolis and Heller study. The peak compensated static acoustic admittance and the acoustic equivalent volume were lower and tympanometric width were higher for females when compared to males.
Conclusions: The study found differences between the two studies with both distribution of the tympanometric measures and gender differences. The results indicate that based on the distribution measures younger adults may be referred unnecessarily for testing. Significant gender differences were identified in the study. The authors suggested test sensitivity of middle ear screening may be loss if male and female data is combined as a norm.

Relevance to the current work: This study informs and provides an understanding of the tympanometric screening norms for the adults in the current study.

Shavelson, R., & Webb, M. (1991). *Generalizability theory: A primer*. California: Sage Publications, Inc.

Purpose of work: The purpose of the work is to present and describe the generalizability theory.
Summary: The work describes the generalizability theory as a statistical theory. It is used for evaluating the reliability of behavioral measurements by estimating variance. It provides formulas for calculating the variance.
Relevance to current work: The formula was used for measuring test-retest variability in the current work. It provided the needed understanding for using a modified variance equation for calculating the variance between subjects and the variance within subjects.

Thornton, A., & Raffin, M. J. (1978). Speech-discrimination scores modeling as a binomial variable. *Journal of Speech and Hearing Research, 21*, 507-518.

Purpose of work: This study's purpose was to describe variability across forms of speech test. It also developed a binomial model to estimate variance of test forms that could be used for test in the communications disorders field.
Method: The records of 4120 of hearing impaired listeners were examined and results from the CID W-22 were used in this study. The subjects were between the ages of 20 to 80 years old with the majority between 50 and 60 years of age. Each 50 word list for 1030 ears was divided in two lists of 25 and then into 5 list of 50. The scores were determined for the shorter lists and then compared and a binomal distribution was created.
Results: A table was created to determine significantly different scores (critical differences) using angular confidence intervals.
Conclusions: The participant's observed score and the number of items in a test are factors that affect variability of test forms. The table can help the clinician identify significant differences in scores and the varibility of the mesuring instrument.
Relevance to the current work: The study was used to substantiate the importance of reliability of speech audiometry testing.

Tillman, T. W., & Jerger, J. F. (1959). Some factors affecting the spondee threshold in normal hearing subjects. *Journal of Speech and Hearing Research, 2*, 141-146.

Purpose of the study: The purpose of this study was to examine the practice effects and prior knowledge of spondee words in SRT testing.
Method: The participants were divided in to three groups of 10 individuals. They were all female between the ages of 18 to 24 years with normal hearing. The participants had no prior experience listening to speech stimuli in a hearing evaluation. The CID Auditory Test W-1 lists were used as speech stimuli in this study. Two SRT's were found for the 10 participants in Group A. They listened to 1-18 spondee words of List E for the SRT then listen to 19-36 to determine their second SRT to examine practice effects only. Group B participants listen to 1-18 words (List E) to determine their first

threshold and then they listen to the same 18 words for their second threshold to examine both possible prior knowledge and practice effects. In group C the listeners were presented with the 18 words (List E) to be used for obtaining their thresholds prior to the first test. The same 18 words (List E) were used to determine both (2) thresholds to examine definite prior knowledge and practice effects.
Results: For group A the two conditions were compared and the thresholds were found to almost have equivalent thresholds. Also for group C no thresholds differences were identified indicating no practice affects. Group B the thresholds improved over the 2 conditions which indicate that the improvement was due to knowledge of the test vocabulary. When group C thresholds from both conditions were compared with group A, group C thresholds yielded 4 to 5 dB lower than group A, thus indicating a difference due to the participant prior knowledge of the words.
Conclusions: The authors concluded that no significant practice effects were demonstrated in SRT testing across 2 trials and prior knowledge of words lead to a lower SRT threshold.
Relevance to the current work: The article demonstrates the importance of familiarizing the participants with the SRT words prior to testing so changes in threshold will not be due to prior knowledge in the re-testing condition.

Wilson, R. H., Morgan, D. E., & Dirks, D. D. (1973). A proposed SRT procedure and its statistical precedent. *Journal of Speech and Hearing Disorders, 38*(2), 184-191.

Purpose of the study: The purpose of this study was to propose modification to the Tillman and Olsen procedure for determining SRT.
Method: The study outlines the procedures for obtaining the SRT using Tillman and Olsen's procedures and then introduces the modifications. The study suggests using a 5 dB decrement verse the 2 dB decrements. Every 5 dB decrement five words are presented at each intensity and when all five words are incorrect at one intensity the test is terminated. Then a correction factor of two is added to determine the Threshold. There were 76 participants that were having an audiologic examination for suspected hearing impairment. The individuals had wide variety hearing impairment. For each participant the PTA, SRT using the 2 dB decrement, and the SRT using the 5 dB decrement procedure were determined.
Results: A significant difference was determined between the 2 dB decrement procedure and the 5 dB decrement procedure. The author noted that the 5 dB decrement procedure is less precise but suggest that for clinical uses the difference is minimal. The new procedure produced similar mean thresholds for speech, pure-tone averages with high correlations.
Conclusions: This procedure can be used for determining an individual's SRT that may be more convenient for the audiologist that use audiometers that are difficult to perform the 2 dB decrement procedure.
Relevance to the current work: This study provides background information about valid procedures in determining SRT however the 2 dB decrement procedure was used in the present study.

Wilson, R. H., & McArdle, R. (2005). Speech signals used to evaluate functional status of the auditory system. *Journal of Rehabilitation Research and Development, 42*(4), 79-94. doi:10.1682/JRRD.2005.06.0096

Purpose of work: The purpose of this paper was to outline the history of speech audiometry and identify factors that contribute to an individual's ability to understand spoken language.
Summary: The paper presents the history behind pure-tone and speech audiometry. It discussed that speech recognition materials were basically developed over 75 years ago and identified key researchers (Hudgins, Hirsh) that have perfected the speech materials and procedures. The developments of the established speech materials (CID W-1, W-2, and NU-6) were presented. Audibility and distortion were presented and described as the two components of hearing loss that affect the hearing system. The authors present the issue of speech recognition in quiet and in noise. The number one complaint is that individuals with hearing loss have difficult hearing in noisy situations however most audiologists do not perform speech recognition in noise. The disadvantages and advantages of speech testing in noise are outlined. Research is presented in which supports speech testing in noise for helping the audiologist understand how well a person understands speech in real life environments. The use of words and sentences for speech recognition in audiological evaluations are outlined. Words are the preferred method by audiologist however studies supporting both methods are presented. The possible pros and cons of each method are described. The paper examines the roll-over phenomenon of speech recognition in noise as a function of presentation level. For some individuals as the recognition performance does not always increase as the intensity level increases. Lastly the paper examines the effects of age and hearing loss with speech recognition. It presents study's that demonstrates that with increasing age auditory ability declines. It also suggested that with age not only auditory ability decreases but also reaction-time decreases which can affect performance on speech recognition.
Conclusions: The paper outlines the factors that affect speech recognition and factors that should be understood and examined when performing an audiological evaluation.
Relevance to the current work: The article provided understanding of the history and the factors that affect speech recognition testing.

Wilson, R. H., & Strouse, A. (1999). Psychometrically equivalent spondaic words spoken by a female speaker. *Journal of Speech, Language, and Hearing Research, 42,* 1336-1346. doi:1092-4388/99/4206-1336

Purpose of the study: The purpose of this study was to psychometrically equate spondaic words to improve homogeneity to decrease variability.
Method: Two experiments were performed in this study. A digital copy of the Veteran's Administration (VA) Speech Recognition and Identification Materials by Hirsh and the female talker was made from the VA compact disc. In experiment 1: 14 randomizations of the 36 words were made and recorded. Twenty participants between 22 – 30 years with normal hearing listened to the words with 10 participants listening to the Hirsh words first and then to the female talker and the remaining 10 listening to speaker in the reverse order. In experiment 2: the individual words by the female talker only were adjusted for intensity and 14 randomizations were made and recorded. Twenty

participants between 19 -29 years with normal hearing listened to the adjusted words by the female talker. The participants in both experiments were familiarized with the words and then the 13 random lists were presented in 2-dB steps at random levels between -10 and 14 dB HL.

Results: In experiment one it was found the threshold for words are different if spoken by different talkers. Therefore listeners performance differed based on the talker of the words and not the words themselves. Based on experiment ones findings the words were adjusted for experiment two. In experiment two the inter word standard deviations were less with the adjusted words therefore producing more homogeneous word thresholds.

Conclusions: The study was able to reduce threshold variability of the words by equating words in experiment one to 0 vu and equating to intelligibility in experiment two. Although these words are less variable for normal hearing individuals they may not be for hearing impaired individuals.

Relevance to the current work: The study provides understanding of the process and importance of homogeneity of words used in SRT testing.

Appendix B

List of Spondaic Words

Aircraft
Airport
Bathtub
Birthday
Broadway
Cowboy
Daylight
Doorway
Downtown
Elsewhere
Hardware
Highway
Horseshoe
Iceberg
Ice cream
Mankind
Meanwhile
Nowhere
Outside
Playground
Railroad
Sailboat
Sidewalk
Somehow
Somewhere
Stairway
Suitcase
Sunlight
Weekend
Welfare
Whitewash
Woodwork
Workshop

Appendix C

Informed Consent

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 ability and that you are qualified for this study.

It will take approximately one hour to complete the test. 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 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