2012-03-28

Test-Retest Reliability in the Determination of the Speech Recognition Threshold

Alyssa Montierth Jacobs
Brigham Young University - Provo

Test-Retest Reliability in the Determination of the Speech Recognition Threshold

Alyssa M. Jacobs

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
Ron W. Channell

Department of Communication Disorders

Brigham Young University

June 2012

ABSTRACT

Test-Retest Reliability in the Determination

of the Speech Recognition Threshold

Alyssa M. Jacobs

Department of Communication Disorders

Master of Science

For many years, speech recognition threshold (SRT) testing has been used as an indicator of audiologic health. However, with changing methods and technology, test-retest reliability has not been reviewed extensively with newer digitally recorded spondaic words which meet a published criterion of listener familiarity. This study examined the test-retest reliability of 33 high frequency usage and psychometrically equated spondaic words. The American Speech-Language-Hearing Association recommended method (2-dB decrements) was used to measure the left and right SRT of 40 participants using both male and female talker recordings. For each participant, four SRTs were found during the test condition and four SRTs were found during the retest condition. All of the SRT scores were analyzed and the averaged SRT values found using a male talker recording resulted in an average retest SRT to be 1.4 dB better than the average test SRT. The averaged SRT values found using a female talker recording resulted in an averaged retest SRT to be 1.2 dB better than the averaged test SRT. The SRT scores also showed high validity when compared to each participant's pure tone average (PTA). This study additionally found no significant interaction in using a male versus a female talker when using digitally recorded and psychometrically equated spondaic words.

Keywords: speech recognition threshold, test-retest reliability, digitally recorded materials

ACKNOWLEDGMENTS

Completion of this project was made possible through the tremendous support from my husband Corbin and my parents who have encouraged me throughout my life to pursue my dreams. Additionally, Dr. Harris has been my mentor throughout my entire graduate school experience and has helped me to achieve academic excellence though his continued guidance and patience. Thank you.

Table of Contents

List of Tables

Table Page

List of Appendixes

Description of Structure and Content

This thesis is presented in a hybrid format where current journal publication formatting is blended with traditional thesis requirements. The introductory pages are therefore a reflection of the most up to date university requirements while the thesis report reflects current length and style standards for research published in peer reviewed journals for communication disorders. Appendix A includes an annotated bibliography. Appendix B is the informed consent form used for the participants in the research study. Appendix C contains a list of the 33 spondaic words tested throughout this research project.

Introduction

The fundamental purpose of auditory testing is to evaluate an individual's hearing ability and to identify any subsequent hearing handicap. Ideally, the audiological evaluation includes measures which test the reliability of the auditory system as a whole and successive measures which are used for a differential diagnosis if abnormalities or pathologies are found. To quantify the amount of hearing impairment at specific frequencies, testing begins with pure-tone audiometry (ASHA, 1978; Gelfand, 2009). Results from pure-tone audiometry are used to make a comprehensive diagnosis of a normal or abnormal hearing level and also to quantify and qualify the type and degree of hearing impairments present (Roeser & Clark, 2007). However, solely measuring the frequency specific deficits found in pure-tone testing provides only a partial understanding as it does not give any analysis of a patient's ability to hear and understand speech.

Speech audiometry uses speech stimuli to further determine a patient's auditory status (Gelfand, 2009), and testing includes finding the speech recognition threshold (SRT). The SRT is used clinically to validate pure-tone hearing thresholds (Fletcher, 1950; Wilson, Morgan, & Dirks, 1973). The SRT is also useful for establishing a baseline for other tests such as word recognition score (WRS) testing, in evaluating hearing aid performance, and for confirming the results of pure-tone audiometry in difficult testing situations, such as testing young children or patients suspected of malingering (Carhart, 1971; Hirsh, et al., 1952; Hudgins, Hawkins, Karlin, & Stevens, 1947).

SRT testing requires the listener to specify the word spoken when presented aurally with a record, tape, digital recording, or by the clinician using monitored live voice (MLV). The speech stimuli for SRT testing are comprised of phonetically dissimilar words to make the task

one of word identification, not phonemic discrimination between similar words. When finding the SRT in English, spondaic words (two-syllable words with equal emphasis on each syllable) are presented and repeated while the audiologist finds the intensity level at which the patient is 50% accurate in their response (ASHA, 1988). However, the specific parameters of SRT testing have evolved over time through the method of presentation, understanding the statistical variation among subjects, and maintaining the standardization of words.

Method of Presentation

When measuring the SRT, researchers have debated the validity of administration with MLV versus the current technology of digitally recorded words presented via CD. Many audiologists use MLV, a more common method of presenting the stimulus for speech audiometry testing (Martin, Champlin, & Chambers, 1998). However, there is a large body of literature available which has found that using recorded stimuli is more reliable than MLV (Hood & Poole, 1980; Martin & Clark, 2009) and more reliable than tape or record due to print-through or natural degeneration. There is also considerable variability in scores when material is presented by different talkers, as each talker may have a different dialect or minor articulation errors. This variability in speech stimuli can amount to over an 8% difference in discrimination scores (Penrod, 1979). Kreul, Bell, and Nixon (1969) argued that SRT word lists should not be thought of as a series of printed words, but as acoustic signals which need to be standardized.

One of the components for effectively finding the SRT should be standardized measures of delivery, according to Roeser and Clark (2007). Audiologists must have standardized measures for SRT testing because variability jeopardizes the accuracy and consistency of test results (Di Berardino et al., 2010; Mendel & Owen, 2011). Moreover, these standardized measures are only applicable in the specific circumstances for which they are adapted and the

population for whom the tests are normed (Kreul, et al., 1969). For this reason, the spondaic words used and tested in this study were previously recorded and psychometrically equated for consistency and standardization among all normally hearing participants (Chipman, 2003).

Statistical Variation

Another factor affecting the administration and interpretation of the SRT is the statistical variation found within and between subjects. Mathematical models have been developed to show the statistical variation of intrasubject variability in speech audiometry (Ostergard, 1983; Raffin & Thornton, 1980; Thornton & Raffin, 1978). Thornton and Raffin cited two statistical interests: the relationship between test performance and communicative function (validity of the test) and the consistency across test forms (reliability of the test). Reliability refers to the precision of the measurement possible with the particular test and assumes that results would be consistent if the participant were given the test multiple times. It is the consistency across test forms, or test-retest reliability, which is of concern in this study.

Standardization of Words

Hudgins et al. (1947) gave four parameters for developing spondaic words. First, the words should be familiar to the listener. If the spondaic words were obscure to the listener, the test could become a measure of vocabulary knowledge, rather than a measure of speech intelligibility. Second, the words should be phonetically dissimilar. For example, if words containing similar sounds such as *eyeball* and *highball* were included in the testing materials, the examination no longer becomes a test of simple word recognition, but of fine discrimination. Third, the words should contain a normal sampling of English speech sounds. However, this parameter was only a general guideline, since Hudgins et al. reported no evidence that every English sound needed to be represented for a SRT measurement. Lastly, the words should be

homogeneous with respect to basic audibility. This homogeneity among words includes having a steep psychometric function with a slope of about 10%/dB at the 50% correct point (Wilson & Strouse, 1999), meaning a 10% increase in the likelihood of a correct response for each decibel of increased intensity. This is to estimate the SRT with as few words as possible.

A study done by Chipman (2003) examined the frequency usage of words from the Central Institute for the Deaf (CID) W-1 word lists (Hirsh et al., 1952). Chipman analyzed the frequency of occurrence of these words against the Standard Corpus of Present-Day American English, (Francis & Kučera, 1982) and the Frown Corpus (Hunt, Sand, & Skandera, 1999) to determine if these words had a high probability of being familiar to the patient. Chipman found that only 13 of the 36 CID W-1 words occurred in the top 10,000 frequently used words. In addition, four words did not occur at all in the over two million words sampled in the two corpora. Chipman concluded that more than one third of the 36 CID W-1 words were not frequently used and may be less familiar than other spondaic words which occur more frequently. Chipman selected new spondaic words (appendix C) and replaced the less frequently used words. These words were digitally recorded and psychometrically equated to have slopes of 16.2%/dB (male talker recording) and 15.2%/dB (female talker recording) for the 33 selected words at the 50% correct point. These slopes are higher than other reported slopes of 10%/dB (Hudgins et al., 1947) because the stimuli were digitally recorded and the data analysis used logistic regression to create psychometric functions to fit the data, rather than third order polynomials (Wilson & Strouse, 1999).

These newer spondaic words were not evaluated for test-retest reliability when developed. The purpose of this thesis was to analyze the test-retest reliability of the SRT found using the spondaic word list developed by Chipman (2003). The test scores were compared to

the retest scores to find a satisfactory clinical range and validated against the participants' PTA. In addition, Chipman's male and female talker recordings were used to determine whether the gender of the talker influenced the test-retest reliability of the SRT.

Method

Development of Materials

The Standard Corpus of Present-Day American English (Francis & Kučera, 1982) and the Frown corpus (Hundt, Sand, & Skandera, 1999) were used to obtain the 10,000 most frequently used English words. The 36 CID W-1 spondaic words (Hirsh, et al., 1952) were also included. A total of 98 spondaic words were selected for recording and evaluation. Test recordings were made using a male and female talker previously selected for the production of speech audiometry materials (Harris & Hilton, 1991). Both talkers were native to the United States and used a standard American English dialect. These talkers were judged by a panel of American English speakers who indicated that the vocal quality and accent of each talker was acceptable. Thirty-three spondaic words were selected for the final list tested in this research study and for eventual clinical application (Appendix C).

All recordings were made in the large anechoic chamber located on the Brigham Young University campus in Provo, Utah, USA. The ambient background noise levels in the anechoic chamber were approximately 60-65 dB down from the speech levels measured during recordings. A Larson-Davis model 2541 microphone was positioned at a 0 azimuth and was covered by a 7.62 cm windscreen at a distance of 15 cm from the male talker and 6 cm from the female talker. The microphone was connected to a Larson-Davis model 900B microphone preamp, and the preamp was coupled to a Larson-Davis model 2200C preamp power supply. The signal from the preamp power supply was then routed through an Apogee AD-8000 24-bit

analog-to-digital converter and the digitized signal was stored on a hard drive for later editing. A 44.1 kHz sampling rate with 24-bit quantization was used for all recordings and every effort was made to utilize the full range of the 24-bit analog-to-digital converter. Once recorded, the words were edited using Sadie Disk Editor software.

During the recording sessions, the talker was asked to pronounce each word four times. Two judges rated each word for perceived quality of production and the best production of each word was then selected for evaluation. After the rating process, the intensity of each word to be included on the CD was edited to yield the same intensity as that of the 1000 Hz calibration tone contained on the CD (American National Standards Institute, 1996). The CD containing the 33 final edited words was produced on a recordable CD-ROM drive using a 44.1 kHz sampling rate and 16-bit resolution. These words were psychometrically equated and had slopes at the 50% point of 15.2-16.2%/dB.

Participants

Forty participants between the ages of 18 and 35 years agreed to participate and signed an informed consent approved by the Brigham Young University Institutional Review Board for participation in this study (appendix B). Basic ethical considerations were made for the protection of the research participants throughout the duration of this study. All of the participants reported English as their primary language and had a history with no auditory pathologies.

Each individual was qualified to participate in the study by passing a hearing screening which included (a) pure-tone testing with thresholds at 15 dB HL or better at all test frequencies (125-8000 Hz., including mid-octave frequencies), (b) tympanometry with a type A

tympanogram and static acoustic admittance between 0.3 and 1.4 mmhos with peak pressure between –100 and +40 daPa, and (c) ipsilateral acoustic reflexes at 90 dB HL (ASHA, 1990).

Calibration

All pure-tone and SRT testing was conducted in a double walled sound booth meeting ANSI standards for maximum permissible ambient noise levels (American National Standards Institute, 1999). Testing was conducted with a Grason-Stadler GSI 61 (model 1761) Clinical Audiometer calibrated to ANSI S3.6-2004 standards (American National Standards Institute, 2004a). The audiometer was calibrated at the beginning of the study and weekly during data collection.

Procedure

All participants were given the following written instructions to orient them to the nature of the task, to specify their mode of response, to indicate that the test material was speech, and to stress the need for the client to respond at faint listening levels (ASHA, 1988).

> You will hear words at a number of different loudness levels. Each word is two syllables in length. At the very soft loudness levels, it may be difficult for you to hear the words. 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?

The participants were given a written list of the 33 spondaic words and were aurally presented these words at 50 dB HL by the use of the same digital recording used to find the SRT. This was done so that each participant was familiar with the test stimulus and could auditorily recognize each test word. Also, it aided in the accuracy of the participants' responses to be interpreted by

the clinician and to control the effects of prior knowledge of test vocabulary on measurement of the SRT (ASHA, 1988; Tillman & Jerger, 1959).

The SRT testing followed the ASHA method using 2 dB decrements (ASHA, 1988). For each participant, the list of spondaic words was randomized and delivered using wav audio player software (version 1.0.3). The preliminary phase to determine the SRT starting level was done by presenting one spondaic word to the participant at 40 dB HL. The sound level was decreased in 10 dB decrements for every correct response (Martin & Stauffer, 1975). When a word was missed, a second word was presented at the same level. This process was continued until two consecutive words were missed at the same hearing level. Subsequent SRT testing began 10 dB above the level where these two consecutive words were missed, referred to as the SRT starting level.

Two spondaic words were presented at the starting level and at each successive 2 dB decrement. This process was continued if five out of the first six words were repeated correctly. When this was not the case, the starting level was increased by 10 dB and words were presented at each successive 2 dB decrement until five out of the six words were repeated correctly. The testing was completed when the participant responded incorrectly to five of the last six words presented. The threshold was calculated by subtracting the total number of correct responses from the starting level and adding a correction factor of 1 dB (ASHA, 1988; Finney, 1952). The SRT found with this method has been determined to be standard and have the lowest amount of variability (Penrod, 1979).

A randomized block design with eight combinations within was used to vary the presentation order for each of the participants, to determine the gender of the talker, and the ear in which the stimulus was presented. This process was repeated four different times. After a

short break, the procedure was repeated to examine the test-retest reliability for a total of eight SRT found. This block design was repeated five times in the duration of the study for a total of 40 participants. See table 1 for a visual representation of the design.

Table 1

Randomized Block Design in the Method of Presentation

	Test				Retest			
Participant	Trial 1		Trial 2		Trial 3		Trial 4	
1	MR	ML	FR	FL	MR	ML	FR	FL
2	MR	ML	FR	FL	FR	FL	MR	ML
3	FR	FL	MR	ML	FR	FL	MR	ML
4	FR	FL	MR	ML	MR	ML	FR	FL
5	ML	MR	FL	FR	ML	MR	FL	FR
6	ML	MR	FL	FR	FL	FR	ML	MR
7	FL	FR	ML	MR	FL	FR	ML	MR
8	FL	FR	ML	MR	ML	MR	FL	FR

Note. M = male talker for SRT and F = female talker for SRT; R = right ear tested and L = left ear tested. This block design was repeated five times throughout the study for a total of 40 participants tested.

In addition, the SRT data were analyzed according to testing order (sequence of SRT presented). Each of the 40 participants was assigned one of four sequences during the testing period and again during the retest period (see table 2). For example, order one implies the first SRT administered to the participant regardless of whether the SRT was found during the test or retest condition. Therefore, for the first participant order one would be the SRT found with the

male talker in the right ear (test condition) averaged with the SRT found with the male talker in the right ear (retest condition). Order two would be the SRT found with male talker in the left ear (test condition) averaged with the SRT found with the male talker in the left ear (retest condition), and so forth. See table 2 for a visual representation of this design.

Table 2

Testing Order Among Participants in Block Design

	Test				Retest			
	Order 1	Order 2	Order 3	Order 4	Order 1	Order 2	Order 3	Order 4
1	MR	ML	FR	FL	MR	ML	FR	FL
2	MR	ML	FR	FL	FR	FL	MR	ML
3	FR	FL	MR	ML	FR	FL	MR	ML
4	FR	FL	MR	ML	MR	ML	FR	FL
5	ML	MR	FL	FR	ML	MR	FL	FR
6	ML	MR	FL	FR	FL	FR	ML	MR
7	FL	FR	ML	MR	FL	FR	ML	MR
8	FL	FR	ML	MR	ML	MR	FL	FR

Note. M = male talker for SRT and F = female talker for SRT; R = right ear tested and L = left ear tested. This block design was repeated five times throughout the study for a total of 40 participants tested.

Results

After the raw data were collected, the SRT scores found with both the male and female talker recordings were averaged across test and retest conditions. The averaged SRT values found using male talker yielded test and retest values of 0.2 dB HL and -1.2 dB HL respectively,

resulting in an average retest SRT 1.4 dB better than the average test SRT. The averaged SRT values found using a female talker yielded test and retest SRTs of 0.4 dB HL and -0.8 dB HL respectively, resulting in an averaged retest SRT 1.2 dB better than the averaged test SRT. For a full list of descriptive statistics, see table 3.

Table 3

SRT Descriptive Statistics for Talker Gender for Test and Retest (dB HL)

	Mean	S.D.	Minimum	Maximum
Female SRT test	0.4	2.8	-5.0	7.0
Female SRT retest	-0.8	3.0	-6.0	9.0
Male SRT test	0.2	3.1	-6.0	7.0
Male SRT retest	-1.2	2.9	-6.0	7.0

Test-retest reliability was analyzed using a modified variance equation (Shavelson & Webb, 1991) designed to calculate the validity of the SRT using the more recently developed spondaic words. The original variability equation is as follows:

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

For the present research design, equation 1 can be simplified to a comparison of variance within subjects and variance between subjects. By inserting a calculated variance within subjects and a calculated variance between subjects, it is possible to find a mathematical estimation of test-retest reliability.

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

In this study, the variance within subjects was found to be 3.18 dB and the variance between subjects was found to be 5.99 dB, resulting in a calculated test-retest reliability of 0.47. Because this total number is far from 1.0, this mathematical model indicates poor test-retest reliability. In order to predict higher test-retest reliability, variance between subjects would need to increase. If the variance between subjects is small then the test-retest reliability value will also be small. If the variance between subjects is large, given the same variance within subjects, the test-retest value will also be large. It is acknowledged that satisfactory calculated test-retest reliability using this equation would come from a study which included participants with a wide variety of hearing impairment (increasing the variance of scores), a limitation of this study and an area for future research. However, it is important to note that the actual difference between test and retest SRT scores for both the male and female talkers was clinically quite good (1.2-1.4 dB). Relating to pure-tone testing, acceptable margin of error is +/- 5 dB (American National Standards Institute, 2004b).

To further examine the validity of Chipman's spondaic words, the SRT data were compared to the PTA of each participant. The averaged SRT found with the female talker was 2.0 dB better than the PTA and the averaged SRT found with the male talker was 2.3 dB better than the PTA. For a full list of PTA and SRT comparison data, see table 4. Additionally, a paired *t*-test was performed to compare the averaged right and left PTA against the SRT found with the male talker and female talker. Statistically significant differences were found comparing the male talker SRT and PTA $t(79) = 1.42$, $p < 0.0001$ and female talker SRT and PTA $t(79) = 1.63$, $p < 0.0001$. However, the *t*-test data is clinically irrelevant as a difference of 2.0-2.3 dB between the SRT and PTA is well within the margin of error accepted by practicing audiologists. According to ASHA, a range of 0.3-3.1 dB is acceptable when comparing the SRT

Table 4

Descriptive Statistics for the PTA and SRT (dB HL)

	Mean	S.D.	Minimum	Maximum
Right ear PTA	2.7	3.1	-3.3	10.0
Female right SRT, test	0.6	2.6	-5.0	7.0
Female right SRT, retest	-0.8	2.9	-6.0	6.0
Male right SRT, test	0.6	2.9	-4.0	7.0
Male right SRT, retest	-0.8	2.9	-6.0	7.0
Left ear PTA	1.0	3.9	-6.7	10.0
Female left SRT, test	0.2	3.1	-5.0	7.0
Female left SRT, retest	-0.8	3.2	-6.0	9.0
Male left SRT, test	-0.2	3.3	-6.0	7.0
Male left SRT, retest	-1.6	2.8	-6.0	4.0
Right and left PTA, averaged	1.8	3.6	-6.7	10.0
Female SRT, averaged	-0.2	3.0	-6.0	9.0
Male SRT, averaged	-0.5	3.0	-6.0	7.0

Note. The averaged SRT scores found using a male and female talker were averaged across test and retest values.

to the PTA (ASHA, 1988). In addition, these data show that using a male versus female talker to find the SRT yields no clinical significance, as the averaged difference is 0.3 dB.

The SRT scores were then averaged across testing order (the sequence of SRT presented). During the examination, each of the 40 participants was assigned one of four orders (see table 2). For example, order one implies the first SRT administered to the participant regardless of whether the SRT was found during the test or retest condition. Therefore, for the first participant order one would be the SRT observed with the male talker in the right ear (test condition) averaged with the SRT found with the male talker in the right ear (retest condition). Order two would be the SRT found with male talker in the left ear (test condition) averaged with the SRT found with the male talker in the left ear (retest condition), and so forth. See table 5 for the descriptive statistics for the testing orders averaged across test and retest values.

Table 5

Descriptive Statistics for the Testing Order SRT (dB HL) Averaged Across Test and Retest

	Mean	S.D.	Minimum	Maximum
Order 1	0.14	2.8	-6.0	7.0
Order 2	-0.26	3.1	-6.0	9.0
Order 3	-0.65	2.9	-6.0	7.0
Order 4	-0.56	3.3	-6.0	7.0

Note. Orders 1-4 represent the averaged first through fourth SRT found in a given sequence, regardless of test or retest condition.

A mixed model ANOVA (blocking over subjects) was done to determine significance among testing order, gender of the talker, and the ear tested. In addition, significant interactions

were probed between: the gender of the talker and the ear which was tested, the gender of the talker and test-retest reliability, the ear tested and test-retest reliability, and test-retest reliability and testing order. For a full list of terms included, see table 6.

A *p* value less than 0.15 was used to eliminate non-significant items for the final statistical model. A significant interaction was found in the effects of testing between the test SRT and the re-test SRT, $F(1,79) = 52.29$, $p < 0.0001$. This implies a statistically significant difference in the values from the SRT scores found during the test period and the retest period. Although the data from the ANOVA suggest the retest SRT scores are significantly different, the actual difference is small (1.2-1.4 dB) when viewed for clinical application.

Table 6

Mixed Model ANOVA with All Terms Included

Source	df	*F*	*p*
Order	3, 77	4.23	0.008
Gender	1, 78	2.74	0.102
Ear	1, 77	0.71	0.402
Test-Retest	1, 78	51.72	<0.001
Gender x Ear	1, 78	2.74	0.102
Gender x Test-Retest	1, 77	0.45	0.507
Ear x Test-Retest	1, 78	0.15	0.701
Order x Test-Retest	3, 75	0.82	0.486

A significant interaction was also found in the testing order (sequence of SRT presented), $F(3,78) = 4.24$, $p = 0.008$. To determine which of the four testing order effects showed significance, a *post-hoc* analysis was done using a Tukey-Kramer adjustment. The SRT measured in the order 1 condition was statistically different from the SRT measured in the order 3 condition, $t(78) = 3.21$, $p = 0.002$. No other order effects were found to be significant. For a summary of these items, the reader is directed to table 7.

Table 7

Mixed Model ANOVA Final Results

	df	*F*	*p*
Order	3, 78	4.24	0.008
Gender	1, 79	2.68	0.106
Test-Retest	1, 79	52.29	0.0001

Discussion

The main purpose of this study was to examine the test-retest reliability of the 33 spondaic words developed by Chipman (2003). These 33 spondaic words were developed to maintain specifications set by Hudgins et al. (1947) which state that spondaic words must meet four parameters in order to maintain validity. These four parameters are: (a) the words should be familiar to the listener, (b) the words should be phonetically dissimilar, (c) the words should contain a normal sampling of English speech sounds, and (d) the words should be homogeneous with respect to basic audibility, including a steep psychometric function with a slope of about 10%/dB. Chipman's list of 33 spondaic words was created to meet all four of Hudgins et al.'s

criteria for current, valid spondaic words. There was emphasis placed on the parameter of familiarity in the language being tested, as this was a specific criticism of the CID W-1 list of spondaic words. When examined, 14 of the original 36 CID W-1 words did not occur in the top 10,000 words combined from the Standard Corpus of Present-Day American English and the Frown Corpus. Four of the 36 CID W-1 words did not occur in the top two million words sampled. Chipman's list contained spondaic words all within the top 10,000 most frequently used words in the English language and included some of the original CID W-1 words. However, because these words were digitally recorded and then psychometrically equated, the slopes of the psychometric functions were higher than the CID W-1 words at 15.2-16.2%/dB as opposed to 10%/dB reported by Hudgins et al.

In examining the test-retest reliability of the 33 spondaic words developed by Chipman, the retest values were 1.4 dB better on average (male talker) and 1.2 dB better on average (female talker) than the test values. These values are good clinical indicators of the reliability of the spondaic words as they are within an acceptable margin of error of +/- 5 dB. The improvement in the retest scores can be justified by the learning effects of each participant as they proceeded through each SRT test.

When analyzing the improvement of scores across order, a statistically significant interaction was found between the SRT scores of orders 1 and 3. These differences may be justified by the participants' increasing familiarity with the task and spondaic words; however, the SRT measured in order 4 did not show as much improvement as the SRT measured in order 3, showing no statistical significance. Therefore, this interaction between the SRT of orders 1 and 3 are perhaps due to chance alone.

The SRTs were then compared to the PTAs of each participant and a difference of

2.0-2.3 dB was observed. According to the ASHA, a range of 0.3-3.1 dB is acceptable when comparing the SRT to the PTA (ASHA, 1988). Therefore, the present results were within an acceptable margin of error.

Also of interest in this study was the relationship between using a male versus a female talker in finding the SRT with digitally recorded and psychometrically equated spondaic words. The mean SRT data of the male talker versus a female talker yielded a difference of 0.3 dB, suggesting that there is no clinical significance in using a male talker versus a female talker.

Conclusion

This study found there to be good test-retest reliability when the SRT are found using the new word list developed by Chipman. The differences in the test and retest scores when compared to each other and against the PTA fall within the margins of clinical error accepted by practicing audiologists. Finding the SRT by these more frequently used (familiar) words maintains published standards for validity in audiologic testing (Hudgins et al., 1947). It is imperative that the specification of frequency usage be met as the CID W-1 words may not be as valid for those in the coming generation as they were in the past. In addition, with increasing availability of technology it is encouraged that speech audiometry is completed by the use of digitally recorded materials to increase efficiency and standardization in the clinical setting, as opposed to presentation with MLV or non-digital recording.

Mathematical calculations of variance showed that the test-retest scores could be improved by increasing the variance between subjects (see equation 2). When increasing the variance between subjects, the denominator in the equation increases. The resulting value will then be closer to 1.0. A larger value (close to 1.0) indicates better test-retest scores, according to this mathematical model. To increase the variance between subjects, the words should be tested

on participants with a wide range of hearing impairments, including those with mild, moderate, severe, and perhaps profound hearing impairments. The SRT scores should then be compared to the PTA and compared across test and retest conditions, an area for future research.

References

American National Standards Institute. (1996). *Specification for Audiometers*. ANSI S3.6-1996. New York: ANSI.

American National Standards Institute. (1999). *Maximum Permissible Ambient Noise Levels for Audiometric Test Rooms*. ANSI S3.1-1999. New York: ANSI.

American National Standards Institute. (2004a). *Methods for Manual Pure-Tone Threshold Audiometry.* ANSI S3.21-2004. New York: ANSI.

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

ASHA. (1978). Guidelines for manual pure-tone threshold audiometry. *American Speech-Language Hearing Association, 20*(4), 297-301.

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

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

Carhart, R. (1971). Observations on relations between thresholds for pure tones and for speech. *Journal of Speech and Hearing Disorders, 36*, 476-483.

Chipman, S. (2003). *Psychometrically equivalent English spondaic words.* M.S., Brigham Young University.

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

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.

Gelfand, S. (2009). *Essentials of audiology* (3rd ed.). New York Thieme Medical Publishers, Inc.

Harris, R.W., & Hilton, L.M. (1991). English digitally recorded speech audiometry materials. Disk #1. Provo, UT, Brigham Young University.

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

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., & Clark, J. G. (2009). Speech Audiometry *Introduction to Audiology* (10th ed., 126-164). Boston: Allyn and Bacon.

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.

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

Penrod, J. P. (1979). Talker effects on word-discrimination scores of adults with sensorineural hearing impairment. *Journal of Speech and Hearing Disorders, 44*(3), 340-349.

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.

Roeser, R. J., & Clark, J. L. (2007). Pure tone tests. In R. J. Roeser, M. Valente & H. Hosford-Dunn (Eds.), *Audiology: Diagnosis* (2nd ed., pp. 238-260). New York: Thieme Medical Publishers, Inc.

Shavelson, R., & Webb, M. (1991). *Generalizability theory: A primer*. California: Sage Publications, Inc., 94.

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.

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. (1996). *Specification for Audiometers*. ANSI S3.6-1996. New York: ANSI.

Relevance to the current work: The audiometers covered in this specification are devices designed for use in determining the hearing threshold level of an individual in comparison with a chosen standard reference threshold level. This standard provides specifications and tolerances for pure tone, speech, and masking signals and describes the minimum test capabilities of different types of audiometers. The calibration tone on the CD used was set to these standards.

American National Standards Institute. (1999). *Maximum Permissible Ambient Noise Levels for Audiometric Test Rooms*. ANSI S3.1-1999. New York: ANSI.

Relevance to the current work: This article specifies maximum permissible ambient noise levels which produce negligible masking of test signals presented at reference equivalent threshold levels specified in ANSI S3.6-1996 (see above). The maximum permissible ambient noise levels are specified from 125 to 8000 Hz in octave and one-third octave band intervals for two audiometric testing conditions (ears covered and ears not covered) and for three test frequency ranges (125 to 8000 Hz, 250 to 8000 Hz, and 500 to 8000 Hz). The sound booth used in this study was calibrated according to these standards.

American National Standards Institute. (2004 a). *Methods for Manual Pure-Tone Threshold Audiometry*. ANSI S3.21-2004. New York: ANSI.

Relevance to the current work: This article specifies the margin of error when finding the pure-tone thresholds in audiologic testing. These standards can be compared to SRT testing when finding test-retest reliability. Because ASHA accepts +/- 5 dB, one can assume that the data found in this study is within the acceptable margins of error.

American National Standards Institute. (2004 b). *Specification for audiometers*. ANSI S3.6-2004. New York: Acoustical Society of America.

Relevance to the current work: The audiometers covered in this specification are devices designed for use in determining the hearing threshold level of an individual in comparison with a chosen standard reference threshold level. This standard provides specifications and tolerances for pure tone, speech, and masking signals and describes the minimum test capabilities of different types of audiometers. The audiometer used in this study was calibrated according to these standards.

ASHA. (1978). Guidelines for manual pure-tone threshold audiometry. *American Speech-Language Hearing Association, 20*(4), 297-301.

Relevance to the current work: These guidelines provide a standard set of procedures representing a concensus of recommendations found in the literature. The testing done in the current study follows the protocol stated in these guidelines for the determination of pure-tone thresholds and standard procedures for monitoring and diagnostic air conduction measures.

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

Relevance to the current work: This American Speech-Language Hearing Association protocol defines standards for general conditions during audiometric testing, including instrumentation and calibration, the testing environment, and test material. It also designates appropriate methods of delivery and response for finding the SRT, including the preferred method of using previously recorded spondaic words. Also included are the purposes and instructions for administering speech recognition testing.

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

Relevance to the current work: These American Speech-Language Hearing Association guidelines define ranges for normal hearing. It also outlines parameters when screening for hearing impairments. These ranges and guidelines were used when qualifying participants for the research.

Carhart, R. (1971). Observations on relations between thresholds for pure tones and for speech. *Journal of Speech and Hearing Disorders, 36*, 476-483.

Purpose of the work: This work reviews standards in pure-tone and SRT testing.

Summary: The purposes of using a pure-tone averaging system are discussed and the implications of using specific frequencies are reviewed. Carhart explains that it is fundamental to have correct calibration and standards for audiometric equipment so that a correction coeffient would not have to be calculated into a PTA. In addition, Carhart raises the question of whether or not patients with differing audiogram shapes and types of hearing loss should have the same predictive SRT formula by using their PTA.

Relevance to the current work: Carhart concludes that it is within the clinician's scope of judgement to decide whether or not to utilize the PTA comprised of two frequencies (500 Hz and 1000 Hz) or to add a third frequency (2000 Hz) and suggests an appropriate correction factor to finding the SRT is 2 dB when comparing to the PTA.

Chipman, S. (2003). *Psychometrically equivalent English spondaic words.* M.S., Brigham Young University.

Purpose of the study: Chipman analyzed the 36 CID W-1 words commonly used for finding the SRT during an audiologic exam, finding that not all of the words met a previously published standard for frequency usage and familiarity to the listener.

Method: As a response to the previously used CID W-1 words, Chipman analyzed two different English language corpora to find spondaic words which met the published criterion for familiarity to the listener, phonetic dissimilarity, normal sampling of English speech sounds, and homogeneity with respect to basic audibility, including a steep psychometric function with a slope of about 10%/dB. Chipman recorded 98 spondaic words and had a male and female talker record each word four times. Each of the words was judged by a panel of native English speakers resulting in the selection of 33 new spondaic words for use in SRT testing.

Results: Chipman developed 33 new spondaic words for use in SRT testing which theoretically hold higher validity than the CID W-1 spondaic words. The words which Chipman recorded are more commonly used in the English language, appropriate for a hearing impaired population changing to a younger age and demographic.

Relevance to the current work: This study by Chipman is the basis for the current research project. The 33 spondaic words developed by Chipman needed to be examined for test-retest reliability, validity in comparison to the PTA, and determine whether using a male or female recorded talker made a clinically relevant influence on finding the SRT.

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

Purpose of the study: The objective was to assess whether using different CD recording protocols affected the reliability and comparability of SRT testing.

Method: The researchers initiated an acoustic analysis of CD recordings currently used in clinical practice to determine whether the speech material was recorded with similar procedures. Normal hearing participants were tested using CDs which had been prepared differently in order to assess the impact of different recording methods and procedures. Once completed the psychometric curves of each participant were compared.

Results: After analysis, the researchers found that the speech material on the CDs had not been recorded in a standardized manner. They found that SRT and maximum intelligibility thresholds differed significantly between CDs and were influenced by factors such as the recording level of the speech material.

Conclusion: The researchers state as a result of their study, clinicians must check for possible differences in the recording gains of the digitally recorded materials used in their practices.

Relevance to the current work: This article claims that the reliability and comparability of speech test outcomes obtained using different CDs will only be maintained if the clincian checks for any possible differences used in preparing the speech materials. If differences are found, compensations must be made. The audio recordings prepared for this study were standardized and the same recordings were used for all participants.

Finney, D. J. (1952). *Statistical Method in Biological Assay*. London: C. Griffen.

Relevance to the current work: This reference outlines the statistics for the 1 dB correction factor used when finding the SRT via the 2 dB method. In the present study, the ASHA 2 dB method was used and therefore a 1 dB correction factor was added.

Fletcher, H. (1929). *Speech and Hearing*. New York: Van Nosteand.

Relevance to the current work: This article is used as a standard in audiologic research and practice. Fletcher states that averaging the pure-tone frequencies at 500 Hz, 1000 Hz, and 2000 Hz will give an accurate pure-tone average which is a good predictor for the SRT. The pure-tone averages calculated in the present study were found by averaging pure-tones found at these three frequencies.

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

Relevance to the current work: Using the three frequencies (500 Hz, 1000 Hz, and 2000 Hz) for finding the PTA is commonly accepted in audiologic research. This article discusses the validity of the PTA in predicting the SRT for patients who have a hearing loss which may impact their ability to understand speech. The Fletcher average is introduced, which is used to compare (instead of the PTA) the SRT when there is a significant hearing loss at either 500 Hz, 1000 Hz, or 2000 Hz. The two best frequencies are solely used for comparison in cases with a sloping hearing loss.

Francis, W. N., & Kučera, H. (1982). *Frequency Analysis of English Usage: Lexicon and Grammar*. Boston: Houghton Mifflin Company.

Relevance to the current work: The Standard Corpus of Present-Day American English, commonly known as the Brown Corpus, is a compilation of current American English. This corpus is widely used and frequently cited in linguistics research. The CID W-1 words were analyzed against this corpus for frequency usage to develop new spondaic words in Chipman's research. The test-retest reliability of these new spondaic words was the focus of the research in the present article.

Gelfand, S. (2009). *Essentials of audiology* (3rd ed.). New York Thieme Medical Publishers, Inc.

Purpose of the work: Standardization in calibration and finding the SRT is imperative for students studying audiology. Guidelines for each are set.

Summary: This work discusses that calibration is a necessary technical step in speech audiometry and that it is advised to calibrate the test material before each use. In addition, the SRT has several clinical functions. These are (a) to serve as a measure for corroborating pure-tone thresholds, (b) to serve as a reference point for deciding an appropriate level to administer suprathreshold speech recognition tests, (c) to determine hearing aid needs and performance, (d) to ascertain the need for aural rehabilitation, and (e) to determine hearing sensitivity for young children and others who are hard to test.

Relevance to the current work: This work presented the purposes and standards for finding the SRT. The present study's objectives were aimed to compare the PTA and SRT as stated in the text.

Gelfand, S., & Silman, S. (1985). Functional hearing loss and its relationship to resolved hearing levels. *Ear and Hearing, 6*(3), 151-158.

Purpose of the study: This article studied the nature of functional hearing loss with respect to hearing sensitivity.

Method: The nature of functional hearing loss was studied retroactively in 126 ears. The difference between the functional and resolved thresholds was related to the resolved hearing levels.

Results: Test results showed that the size of the functional overlay was essentially the same for normal to mild hearing losses. With sloping high-frequency losses, the magnitude of funcional overlay became smaller for the impaired frequencies.

Conclusion: The findings of this study suggest that pure-tone audiometric configuration and amount of functional hearing loss bilaterally is accounted for on the basis of explainable auditory factors.

Relevance to the current work: This is an additional resource to discuss hearing loss and the comparison of the PTA and the SRT, although this method was not used in the present study.

Harris, R.W., & Hilton, L.M. (1991). English digitally recorded speech audiometry materials. Disk #1. Provo, UT, Brigham Young University.

Relevance to the current work: The digitally recorded spondaic words tested in the present research were made using the male talker and female talker previously selected for the production of these speech audiometry materials. Both talkers were native to the

United States, spoke a standard American English dialect, and were judged by a panel of native American English speakers to have an acceptable accent and vocal quality.

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 the study: The objective of this study was to modify earlier speech audiometry materials and tests to correct deficiencies found clinically. Two improvements stated were to restrict the word lists in order to promote familiarity to the listener and to record these words on magnetic tape.

Method: In order to create the CID W-1 list of spondaic words, the researchers begain with the 84 words found in the PAL Auditory Test No. 9 and then were rated for familiarity by independent judges. Two acetate discs were cut and the talker monitored the carrier phrase, 'say the word' on a VU meter in order to maintain equal emphasis. Six experienced listeners and six inexperienced listeners then judged the more familiar CID W-1 words at +4 dB, +2 dB, 0 dB, -2 dB, -4 dB, and -6 dB relative to the threshold found with the PAL Test 9.

Results: There was no significant difference found in the data among the experienced listeners and inexperienced listeners. The articulation vs. gain function showed an articulation score which rose from 0-100% within 20 dB. There was an increase from 20% to 80% within a range of 8 dB and within this range the rise in score is approximately 8%/dB.

Conclusion: The researchers conclude that these CID W-1 spondaic words are a better clinical alternative than the PAL Auditory Test No. 9.

Relevance to the current work: It is these words which were revised in the research done by Chipman (2003) and tested for test-retest reliability in the present study.

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 work: The objective of this research was to determine the characteristics of recorded word articulation material.

Method: 45 listeners judged 20 phonetically balanced word lists comprising five words which were recorded on tape by a professional announcer. The researchers assigned a grade of difficulty and compiled two lists of 25 reported difficult words and 25 reported easy words. These words were re-recorded by two additional speakers.

Conclusion: The characteristics of recorded materials are determined by the speaker and recording technique.

Relevance to the current work: There is considerable variability in scores when material is presented by different talkers; therefore, standardization in recordings should be present in SRT testing. It is a standardized recording that was used in the present research.

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.

Purpose of the work: Specifications for speech audiometry materials are given.

Summary: This work states that there are four parameters when developing speech audiometry materials. These four parameters are (a) the words should be familiar to the listener, (b) the words should be phonetically dissimilar, (c) the words should contain a normal sampling of English speech sounds, and (d) the words should be homogeneous with respect to basic audibility.

Relevance to the current work: Chipman (2003) discovered that some of the CID W-1 words no longer met the criteria of familiarity. When Chipman developed a new list of spondaic words which met all four criteria developed by Hudgins et al., the list of spondaic words needed to be tested for 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

Relevance to the current work: The text of the Freiburg-Brown (Frown) corpus was selected to closely match the work of the Standard Corpus of Present-Day American English (Brown corpus) in the sampling of words selected in Chipman's (2003) work of developing new materials for SRT testing. These words were tested for test-retest reliability in the present study.

Jerger, J., & Hayes, D. (1977). Diagnostic speech audiometry. *Archives of Otolaryngology, 103*(4), 216-222.

Purpose of the work: The scope of speech audiometry can include useful diagnostic information in clinical testing.

Summary: When there is comparison of performance vs. intensity functions for phonetically balanced words in addition to synthetic sentence identification, a clinically useful pattern is shown which differentiates peripheral and central sites of auditory disorders.

Relevance to the current work: Although the current work focuses on the SRT testing portion of speech audiometry, it is important to understand the relevance of all parts of the speech audiometry examination and the clinical implications of the data found.

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 work: Specifications for speech audiometry are given.

Summary: Kreul, Bell, and Nixon argued that SRT tests should be thought of in terms of acoustic signals which need to be standardized and not a series of printed words. In addition, test standards are only applicable in specific circumstances in which they are adapted and for the population for which the tests are normed.

Relevance to the current work: The test stimuli used in the present research study was digitally recorded for standardization among participants.

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 work: The objective of this article was to determine if clinical practices were being retained, modified, or replaced among practicing audiologists.

Method: A 5-page questionnaire was sent to 500 audiologists randomly selected from members of the American Academy of Audiology. The results were then compared to the data from similar studies done in 1971, 1972, 1978, 1985, 1989, and 1994.

Results: Most audiologists continue to use monitored live voice (MLV) as the method of presentation for speech audiometry.

Conclusion: Using MLV as a method of presentation for speech audiometry is not as reliable as using digitally recorded materials as a method of presentation.

Relevance to the current work: The present study used digitally recorded materials as the method of presentation for SRT testing.

Martin, F. N., & Clark, J. G. (2009). Speech Audiometry *Introduction to Audiology* (10th ed., pp. 126-164). Boston: Allyn and Bacon.

Purpose of the work: This chapter introduces speech audiometry and how to interpret the data obtained.

Summary: There are advantages to using digitally recorded materials rather than MLV. Recorded materials provide significantly more reliable measures than that of MLV because of their standardized nature.

Relevance to the current work: The spondaic words used in the SRT testing were digitally recorded in order to obtain standardization in the testing.

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.

Relevance to the current work: The ASHA 2 dB method for finding the SRT is based upon the modification of the Tillman-Olsen method for finding the SRT. When finding the starting level for the SRT, the clinician is to set the hearing level 30-40 dB above the estimated SRT and present one word to the patient. If the response is correct, the clinican presents a word decending in 10 dB decrements until the patient misses two words consecutively at the same level. The starting level is then 10 dB above the level where the last two spondaic words were missed. The ASHA 2 dB method was used in the present study.

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 work: The objective of the researchers was to determine the amount of time needed for word recognition with MLV versus digitally recorded materials.

Method: 50 word NU-6 lists were presented via MLV, short ISI CD recordings, and long ISI CD recordings.

Results: The average time for administration was shortest using MLV rather than CD recordings. However, there was more variability in testing time with MLV than with a CD recording. There was no significant difference in administration times for the recorded lists.

Conclusion: Presentation with MLV was 49 seconds faster when testing patients with a hearing impairment. The researchers concluded that this is not a clinically significant amount of time.

Relevance to the current work: This article discussed some of the history of speech audiometry and also the techniques of monitored live voice (MLV) versus digitally recorded materials. Pros and cons of each method are discussed and the conclusion is that there is not a clinically significant difference in the amount of time needed for administration.

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

Purpose of the work: The objective of this work is to discuss the statistics and probability outcomes of speech audiometric testing.

Summary: Speech tests should be characterized by validity, reliability, sensitivity, and specificity. With test development, the clinician should be aware of how an individual

performs in a variety of circumstances and should be able to infer the degree to which the individual possesses the test construct. Moreover, test reliability is the "precision of measurement" (p. 224) possible with a particular test. One assumption of test reliability is that if the patient were given the test under the same circumstances, the results would be consistent.

Relevance to the current work: Test reliability is the primary focus of the present study. The research design was established to hypothesize test and retest results to be consistent among the group of participants.

Penrod, J. P. (1979). Talker effects on word-discrimination scores of adults with sensorineural hearing impairment. *Journal of Speech and Hearing Disorders, 44*(3), 340-349.

Purpose of the work: This article analyzed the interaction between the talker and listener in speech audiometric testing.

Method: 30 participants completed speech discrimination testing. Tape recordings of four talkers using the CID W-22 word lists were used and the listeners' responses were scored.

Results: 26 out of 30 participants (87%) showed variability greater than 8% between their lowest and highest word discrimination scores.

Conclusion: Statistical analysis indicated that there is an interaction between talker and listener.

Relevance to the current work: This article further supports the proposal that SRT testing must use standardized methods of delivery when presenting spondaic words.

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 the work: This article was written as an addition to Thornton & Raffin (1978) which proposed a model to describe variability in speech discrimination scores.

Summary: The authors published confidence intervals used in their statistical model. These tables are used to construct tables for critical differences at any confidence level for clinical reference.

Relevance to the current work: This work used in conjunction with Thornton & Raffin (1978) was used to hypothesize results from test-retest scores done in the present study.

Roeser, R. J., & Clark, J. L. (2007). Pure tone tests. In R. J. Roeser, M. Valente & H. Hosford-Dunn (Eds.), *Audiology: Diagnosis* (2nd ed., pp. 238-260). New York: Thieme Medical Publishers, Inc.

Purpose of the work: This work defines the nature and purpose of audiologic testing.

Summary: Details of pure-tone testing and the nature of an audiological exam are presented in this chapter. Results of the pure-tone test allow the clincian to make an intial diagnosis for the depth and breadth of audiologic diagnostic and rehabilitation precedures needed for each patient. In addition, pure-tone results allow the clincian to determine the type and extent of a patient's hearing loss.

Relevance to the current work: This chapter was utilized as a basic reference in learning the purpose of different audiologic tests, including pure-tone testing and SRT testing.

Siegenthaler, S., & Strand, R. (1971). Audiogram-average methods and SRT scores. In I. Ventry, J. Chaiklin & R. Dixon (Eds.), *Hearing Measurement: A Book of Readings* (pp. 251-255). Englewood Cliffs, New Jersey.

Purpose of the work: This chapter discusses the relationship between the PTA and the SRT scores.

Summary: This work discusses different audiogram averaging methods and the rationale behind using different frequencies. In addition, these methods are compared to SRT scores and examined for clinical relevance.

Relevance to the current work: This chapter served as a reference when deciding to use a three-frequency average for comparing the PTA to the SRT in the present study.

Shavelson, R., & Webb, M. (1991). *Generalizability theory: A primer*. California: Sage Publications, Inc., 94.

Relevance to the current work: This work served as a reference to the statistical analysis used in finding the test-retest reliability of the data collected in the present study.

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 the work: This study discusses variability in speech audiometry test data.

Method: Lists 1-4 of the CID W-22 were presented to 1030 ears with varying degree of hearing impairment. The scores were compared and a binomal distribution was created.

Results: Angular confidence intervals were used to create table of critical differences which could be used clinically.

Conclusion: The bionomial characteristics of speech audiometry tests make variability among test forms dependent upon two factors. These factors are the number of items in a test and the participant's true score.

Relevance to the current work: This work was referenced as a comparison of validity and reliability in speech audiometry testing.

Tillman, T. W., & Jerger, J. F. (1959). Some factors affecting the spondee threshold in normal hearing subjects.

Purpose of the work: ASHA recommends that the effects of prior knowledge of words must be controlled in order to obtain consistent results in SRT testing.

Summary: Differences in SRT scores are obtained if the patient is not familiar with the words presented. Prior familiarization with the spondaic words improved SRT scores by 4-5 dB.

Relevance to the current work: All participants were familiarized with the list of spondaic words before the testing began. The participants were given the list to read as the words were presented aurally via digital recording.

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 work: This article discusses differences in the ASHA 2 dB method and ASHA 5 dB method for finding the SRT.

Method: 100 ears were tested and 36 spondaic words were recorded on magnetic tape. The participants' SRT scores were found via the ASHA 2 dB and 5 dB methods and compared to the participants' PTA.

Results: The researchers found smaller inconsistencies with the ASHA 2 dB method.

Conclusion: Both methods are reliable and valid for clinical usage.

Relevance to the current work: The present study used the ASHA 2 dB method for finding the SRT and compared this data to the participants' PTA for validity measures.

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.

Purpose of the work: This study was designed to psychometrically equate spondaic words from the CID W-1 lists.

Method: Two studies were performed on two groups of listeners with normal hearing. Each group was comprised of 20 participants. In the first experiment, psychometric functions were established by a male and female talker. Based upon this threshold data, the words spoken by the female talker were adjusted digitally to produce equal intelligibility thresholds with the male talker. In experiment two, psychometric functions were established for the 36 spondaic words.

Results: The mean thresholds for the two experiments were the same but the standard deviations in experiment two were significantly smaller than in experiment one.

Conclusion: The psychometrically equated word recordings are now available and should be used in order to maintain test validity.

Relevance to the current work: The recorded words in the present study were homogeneous and had a steep psychometric function with a slope of over 10%/dB.

Appendix B

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 is normal 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

Appendix C

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