Test-retest Reliability in Word Recognition Testing in Subjects with Varying Levels of Hearing Loss

Meghan Elizabeth Grange
Brigham Young University - Provo
Test-Retest Reliability in Word Recognition Testing in Subjects with Varying Levels of Hearing Loss

Meghan E. Grange

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

Richard W. Harris, Chair
Shawn L. Nissen
David L. McPherson

Department of Communication Disorders Brigham Young University March 2013

Copyright © March 2013 Meghan E. Grange All Rights Reserved
ABSTRACT

Test-Retest Reliability in Word Recognition Testing in Subjects with Varying Levels of Hearing Loss

Meghan E. Grange
Department of Communication Disorders
Master of Science

The purpose of this study was to determine the test-retest reliability of digitally recorded word recognition materials. Word recognition testing is included in a complete audiological evaluation to measure an individual’s ability to discriminate what they hear. A phonetically balanced list of 50 monosyllabic words was presented to each participant at four different sensation levels (SL) using the American Speech Language Hearing Association recommended protocol for word recognition score testing. Each participant took a 10 minute break before the test was readministered. Participants included 40 subjects with varying levels of hearing loss, from normal hearing to severe hearing loss. The test and retest scores of all participants were analyzed to estimate the test-retest reliability to be .65 at 10 dB SL, .87 at 20 dB SL, .88 at 30 dB SL, and .95 at 40 dB SL. It was concluded that the word lists have strong test-retest reliability at 20, 30, and 40 dB SL and that the reliability increases as the presentation level increases.

Keywords: word recognition score, speech audiometry WRS, speech discrimination, test-retest reliability, digitally recorded materials
ACKNOWLEDGMENTS

This project has been a rigorous learning experience, and I am so grateful to have had the expertise of Dr. Harris and my committee members Dr. McPherson and Dr. Nissen throughout the thesis process. I am indebted to my research partner, Karin Caswell, who inspires me with her dedication and determination. Additionally, I am grateful to my family members, classmates, and roommates who have encouraged me throughout my academic journey. Thank you.
TABLE OF CONTENTS

LIST OF TABLES .......................................................................................................................... v

LIST OF APPENDIXES .............................................................................................................. vi

DESCRIPTION OF STRUCTURE AND CONTENT ..................................................................... vii

Introduction ............................................................................................................................... 1

Method of Presentation ........................................................................................................... 4

Test-retest Reliability ............................................................................................................. 5

Method ....................................................................................................................................... 6

Participants .............................................................................................................................. 6

Calibration ............................................................................................................................... 7

Procedure .................................................................................................................................. 8

Results ....................................................................................................................................... 9

Discussion ................................................................................................................................ 11

Conclusion .............................................................................................................................. 14

References ............................................................................................................................... 15
# LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Hearing Classification of Participants based on Pure-Tone Averages (N = 40)</td>
<td>7</td>
</tr>
<tr>
<td>2. Word Recognition Score (WRS) Descriptive Statistics in Percentage Correct (N = 40)</td>
<td>10</td>
</tr>
<tr>
<td>3. Word Recognition Score (WRS) Descriptive Statistics for Simple Main Effects (N = 40)</td>
<td>11</td>
</tr>
</tbody>
</table>
## LIST OF APPENDIXES

<table>
<thead>
<tr>
<th>Appendix</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>A. Annotated Bibliography</td>
<td>17</td>
</tr>
<tr>
<td>B. Informed Consent</td>
<td>29</td>
</tr>
<tr>
<td>C. CID W-22 Word Lists</td>
<td>30</td>
</tr>
</tbody>
</table>
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

Communication is an essential part of daily living. In today’s society, modern technology provides a myriad of modalities for transmitting messages. However, oral/aural communication continues to dominate interpersonal interaction. Deficits in communication due to hearing loss can markedly impact an individual’s daily activities and ability to function in society. When a hearing loss is suspected, several assessments are used in a complete audiological evaluation to determine hearing ability. While pure tone audiometry, tympanometry, otoscopy, and acoustic reflex testing are necessary for establishing the general health of the hearing system, speech audiometry tests can provide critical information about the function of the hearing system in daily situations, as they assess a listener’s ability to perceive and understand speech stimuli (ASHA, 1977, 1988). Perceiving and processing speech stimuli are some of the most important functions of the auditory system because it is through producing and understanding speech that we communicate. Two of the most commonly used speech audiometry measures are Speech Recognition Threshold (SRT) testing and Word Recognition Score (WRS) testing. Both of these assessment tools involve presenting words to the listener and asking them to repeat each word following its presentation.

A complete audiological evaluation generally begins with pure-tone testing of both air conduction and bone conduction. Finding air conduction thresholds involves presenting the listener with pure tones at 250, 500, 1000, 2000, 3000, 4000, 6000, and 8000 Hz via headphones or insert phones and determining how loud the sound must be for the listener to perceive the tone 50% of the time (ASHA, 1977). The process is similar for bone conduction. The stimulus is presented at the frequencies of 250, 500, 1000, 2000, 3000, and 4000 Hz (ASHA, 1977). In this test, the stimulus is introduced via a vibrating bone conductor which is mounted either on the
mastoid process of the skull, or the forehead (ASHA, 1977). After pure-tone thresholds have been determined through both air conduction and bone conduction, speech audiometry is often performed next, followed by acoustic reflex testing and tympanometry (Brandy, 2002).

Pure-tone testing is an essential component of a complete audiological evaluation, however augmenting pure-tone testing with additional measures often provides a more comprehensive understanding of an individual’s hearing abilities. Speech audiometry addresses the listener’s speech reception, speech recognition, and speech understanding abilities (Hamid & Brookler, 2006). These assessment tools can help determine the precise degree and type of hearing loss as well as aid in the selection of an appropriate assistive device (Hamid & Brookler, 2006). Speech audiometry tests are routinely administered to both children and adults during a hearing evaluation (Brandy, 2002). According to a survey of audiological practices conducted by Martin, Armstrong, and Champlin (1994), 99% of audiologists use speech audiometry regularly as part of a general hearing evaluation.

There are two basic types of speech audiometry, SRT and WRS testing. There are three general purposes for performing SRT: (a) to check the results of the pure-tone air conduction testing by comparing SRT results to the air conduction pure-tone average (PTA), (b) to evaluate the hearing sensitivity level for speech, and (c) to provide a baseline used to determine the intensity level for subsequent suprathreshold testing, such as WRS (Brandy, 2002).

Extensive research has examined the correlation between the PTA and the SRT. Carhart (1971) determined that the average of the pure-tone air conduction thresholds at 500, 1000, and 2000 Hz is a relatively accurate predictor of an individual’s SRT, because the bulk of conversational speech is spoken within these frequencies. When assessing individuals whose pure-tone threshold curve has sharp peaks or valleys, a two frequency PTA may be appropriate,
in which the best two frequencies are selected. Also known as Fletcher’s average, this measurement has been shown to correlate better with the SRT in individuals whose air conduction thresholds are not flat (Brandy, 2002). When the PTA and the SRT fall within 6 dB of each other, the agreement is said to be good. If the scores fall within 7 to 12 dB of each other, they are said to have fair agreement. However, if the difference between the two scores is greater than 13 dB, the agreement is poor (Brandy, 2002). The SRT or the PTA, or both should be retested if the agreement is poor and cannot be logically explained by the type of patient (Brandy, 2002).

While the PTA is sometimes used as a single number indication of a person’s hearing sensitivity, the SRT is more indicative of the ability to understand speech and better reflects a functional level (Brandy, 2002). An SRT of 20 dB HL or less represents normal hearing sensitivity in mature children and adults (Brandy, 2002). Audiologists use the SRT to determine the beginning intensity for performing suprathreshold testing. A common rule of thumb in suprathreshold testing is to begin 30 dB above the SRT (ASHA, 1988; Brandy, 2002).

Following the administration of SRT testing, WRS is often the next assessment given to the patient during a comprehensive audiological evaluation. This assessment is designed to determine a listener’s capacity to understand speech that is presented at supra-threshold intensities. Typically, the WRS is measured for each ear individually using lists of phonemically balanced words, sometimes referred to as phonetically balanced (PB) word lists. Monosyllabic words are generally selected for WRS testing in the English language. The individual’s score on the assessment is reported as the percentage of words correct in each list (Brandy, 2002).

Historically, several different lists have been developed for WRS in the English language. In 1948, Egan (1948) compiled a list of 1000 words that was segmented into 20 different PB
word lists containing 50 words in each list. A prominent radio figure, Rush Hughes, made recordings of 8 of these lists, and these were commonly referred to as the PB-50, or the Rush Hughes test (Brandy, 2002). This test was one of the earliest standardized speech audiometry tools and was widely supported by audiologists, because it eliminated the speaker to speaker variability present in monitored live voice (MLV) presentation of the words (Brandy, 2002). However, researchers soon found that the Rush Hughes test was not without flaw. Clipped peaks in the recordings, lack of equivalency among the lists, and listener unfamiliarity with some of the words rendered the assessment tool somewhat unreliable (Brandy, 2002). Later, Hirsh et al. (1952) revised Egan’s original list of 1000 words and used a five member panel of judges to ensure that the words would be familiar to native English speakers (Brandy, 2002). The Hirsh et al. lists were recorded with the words being spoken with a descending pitch contour, and these speech materials are still available and used currently and are referred to as the CID Auditory Test W-22 (Hirsh et al., 1952). Other subsequently developed lists include the NU-6, developed by Tillman and Carhart (1966) which featured improved phonemic balancing, and the Maryland CNC developed by Causy et al., as cited by Brandy (2002), which resulted from further research of acoustic context surrounding the syllable nucleus in the monosyllabic words, and controlled coarticulatory effects. The materials used in the present investigation were updated digital recordings of the CID Auditory Test W-22, recorded on the campus of Brigham Young University in Provo, Utah.

**Method of Presentation**

As speech audiometry materials have developed, researchers have compared the benefits and drawbacks of using recorded word lists to the older, traditional method of MLV (Brandy, 1966). While MLV is more flexible in situations where the patient type necessitates some kind
of adaptation or adjustment to the normal testing environment, it is difficult to maintain consistency in the way the words are spoken across individuals and over time. While there are certainly some advantages to using MLV, research has shown recorded speech stimuli results in more accurate, consistent speech audiometry testing (Brandy, 1966).

**Test-retest Reliability**

When evaluating the effectiveness of a diagnostic tool, such as WRS, it is valuable to examine the variability of the measure. Researchers have developed ways to evaluate intrasubject reliability in word recognition tests (Ostergard, 1983; Raffin & Thornton, 1980). Two sources of error were noted by Raffin and Thornton (1980) as factors that potentially affect intrasubject variance in speech audiometry: firstly, the relationship between test performance and test validity, and secondly, test reliability. Test validity is a measurement of how well the test is measuring what it claims to measure. Test reliability indicates the degree to which a measure yields consistent results each time it is measured. The present investigation will focus on the matter of test reliability, specifically test-retest reliability.

A previous study by Jacobs (2012) evaluated the test-retest reliability of another speech audiometry measure, SRT scores, in a normal hearing population. In her study, Jacobs also addressed the factor of talker gender in the digitally recorded SRT materials. Jacobs found there was no significant difference between patient performance with male talker compared with performance using recordings with a female talker. The test-retest reliability results Jacobs reported showed weak test-retest reliability (.47). Jacobs noted that the homogeneity of hearing acuity among participants influenced the test-retest reliability. Increasing the variance of the population to include participants with a variety of hearing impairments would result in better reliability in a test-retest situation. When making inferences about the hearing system based
upon results from speech audiometry testing, reliability is a paramount concern. High reliability is achieved when variability of results are low. Variability refers to the change in an individual’s performance when it is measured multiple times under the same assessment conditions (Ostergard, 1983). When reliability for a measure is poor, professionals cannot confidently infer the results from a single evaluation (Ostergard, 1983). The purpose of the present study was to examine test-retest reliability in WRS testing using CID W-22 word lists recorded at Brigham Young University in a varied population including a range of hearing loss levels, from normal hearing to severe hearing loss.

**Method**

**Participants**

The 40 adults who were selected for participation in this study were between the ages of 19 and 83 years, native English speakers, and reported no history of cognitive or language impairment. Participants signed an informed consent document which had been registered with and approved by the Brigham Young University Institutional Review Board (Appendix B). Basic ethical precautions were established prior to the beginning of the study to ensure the protection of all study participants.

Prior to selection for participation, a hearing screen was used to assess each potential subject. The hearing screen consisted of pure-tone testing, tympanometry, and otoscopy. Following the hearing screening, the participants were categorized according to their level of hearing loss. To increase statistical variance, the sample included participants with normal hearing (-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 (56-70 dB HL), and severe hearing loss (71-90 dB HL). These ranges were taken from Clark’s (1981)
modification of Goodman’s initial categories of impairment (Goodman, 1965). Clark’s scale included an additional level, slight hearing loss, from 16 to 25 dB HL. These levels, developed by Goodman and later modified by Clark, were based on average thresholds for the frequencies of 500, 1000, and 2000 Hz. Participants ranged from normal to severe sensorineural hearing loss, with a mean PTA of 31.2 dB HL and a range of -5.0 to 71.2 dB HL. For a complete distribution of participants by hearing loss category, see Table 1.

Table 1

_Hearing Classification of Participants Based on Pure-Tone Averages (N = 40)_

<table>
<thead>
<tr>
<th>Hearing Loss Category</th>
<th>Range dB HL</th>
<th>Number of Participants</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal hearing</td>
<td>-10 - 15</td>
<td>10</td>
</tr>
<tr>
<td>Slight hearing loss</td>
<td>16 - 25</td>
<td>6</td>
</tr>
<tr>
<td>Mild hearing loss</td>
<td>26 - 40</td>
<td>12</td>
</tr>
<tr>
<td>Moderate hearing loss</td>
<td>41 - 55</td>
<td>5</td>
</tr>
<tr>
<td>Moderately severe hearing loss</td>
<td>56 - 70</td>
<td>6</td>
</tr>
<tr>
<td>Severe hearing loss</td>
<td>71 - 90</td>
<td>1</td>
</tr>
</tbody>
</table>

Calibration

All WRS testing took place in a double-walled sound suite which met ANSI S3.1 (1991) standards for maximum permissible ambient noise levels for the ears uncovered condition. A Grason-Stadler GSI 61 (model 1761) clinical audiometer was used to obtain the word recognition scores. Immediately before the beginning of data collection for this study, the
audiometer was calibrated to ANSI S3.6 standards. Calibration was verified prior to, weekly
during, and at the conclusion of data collection.

Procedure

Each participant was tested in a single test session. Pure-tone testing, tympanometry, and
otoscopy were conducted to classify the participant into a category of hearing loss. This initial
evaluation helped ensure that the range of hearing loss was not homogenous. Following the
initial evaluation and classification of hearing loss, participants completed a word recognition
test. Prior to the administration of the WRS test, the participants were read the following
instructions in English:

You will hear one-syllable words at several different loudness levels. At the very soft
levels if may be difficult for you to hear the words. For each word, please listen carefully
and 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?

Next, each participant was presented with the four CID W-22 50-word lists at four
intensity levels. Randomized monosyllabic words were presented using TDH-50P headphones.
A listing of the Brigham Young University recordings of CID W-22 word lists is presented in
Appendix C. For each participant, one ear was randomly selected to be the test ear. A single
TDH-50P headphone was used for all WRS testing for all participants. Each of the four word
lists was randomly paired with a presentation level above the SRT; 10, 20, 30, or 40 dB sensation
level (SL). The intensity-word list pairings as well as the presentation order of the lists were
randomized for each participant. Each of the four 50-word lists was presented at its specified
intensity, and a tally was kept of correct and incorrect responses in order to calculate the
percentage correct for each list. After a 10 minute break, the same four intensity levels were
used to present the lists in the retest condition. The intensity-word list pairings, the presentation
order of the lists, and the order of the test items were randomized for the retest condition as well.

Results

The WRS data from the test condition, retest condition and difference (WRS test – WRS
retest) was statistically analyzed for each sensation level. The WRS test-retest descriptive
statistics are presented in Table 2.

A variability equation was adapted in order to analyze test-retest reliability of the recorded
WRS word lists. The original variability equation which was modified for this study is as
follows:

\[
P'_{xx} = \frac{\sigma_p^2}{\sigma_p^2 + \left[ \sigma_{pe}^2 \right]_{ni'}}
\]

Equation 1 was adapted to compare the variance within subjects to the variance between
subjects. Equation 2 was utilized to calculate the test-retest reliability of WRS measures at each
presentation level. A procvarcomp was the statistical model used to determine the estimated
variance. The SAS 9.3 software was used to perform the statistical analysis. Using equation 2,
the test-retest reliability was found to be .65 at 10 dB SL, .87 at 20 dB SL, .88 at 30 dB SL, and
.95 at 40 dB SL.

\[
Reliability = 1 - \frac{\text{variance within subjects}}{\text{variance between subjects}}
\]
Table 2

Word Recognition Score (WRS) Descriptive Statistics in Percentage Correct (N = 40)

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>Minimum</th>
<th>Maximum</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>WRS test</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10 dB SL</td>
<td>19.05</td>
<td>0.00</td>
<td>72.75</td>
<td>14.05</td>
</tr>
<tr>
<td>20 dB SL</td>
<td>50.60</td>
<td>8.00</td>
<td>90.00</td>
<td>21.10</td>
</tr>
<tr>
<td>30 dB SL</td>
<td>72.75</td>
<td>20.00</td>
<td>100.00</td>
<td>19.82</td>
</tr>
<tr>
<td>40 dB SL</td>
<td>82.10</td>
<td>40.00</td>
<td>100.00</td>
<td>17.66</td>
</tr>
<tr>
<td>WRS retest</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10 dB SL</td>
<td>22.68</td>
<td>0.00</td>
<td>81.00</td>
<td>17.00</td>
</tr>
<tr>
<td>20 dB SL</td>
<td>52.25</td>
<td>6.00</td>
<td>92.00</td>
<td>20.47</td>
</tr>
<tr>
<td>30 dB SL</td>
<td>72.60</td>
<td>28.00</td>
<td>98.00</td>
<td>19.97</td>
</tr>
<tr>
<td>40 dB SL</td>
<td>82.05</td>
<td>46.00</td>
<td>100.00</td>
<td>17.33</td>
</tr>
<tr>
<td>Difference</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(WRS test-retest)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10 dB SL</td>
<td>-3.63</td>
<td>-34.00</td>
<td>16.00</td>
<td>10.82</td>
</tr>
<tr>
<td>20 dB SL</td>
<td>-1.65</td>
<td>-28.00</td>
<td>22.00</td>
<td>9.87</td>
</tr>
<tr>
<td>30 dB SL</td>
<td>.15</td>
<td>-20.00</td>
<td>28.00</td>
<td>9.36</td>
</tr>
<tr>
<td>40 dB SL</td>
<td>.05</td>
<td>-10.00</td>
<td>14.00</td>
<td>5.61</td>
</tr>
</tbody>
</table>

A mixed model analysis of variance (ANOVA) was calculated to examine the simple main effects of intensity (10, 20, 30, 40 dB SL), list (1, 2, 3, and 4), and time (test and retest). As expected, there were significant differences in WRS scores due to simple main effect of
intensity, $F(3,270) = 667.56, p < .001$. Also as expected, there were no significant differences in WRS scores due to the simple main effects of list, $F(3,270) = .21, p < .89$, or time, $F(1,270) = 1.54, p < .22$. However the interaction between time * list was significant, $F(3,270) = 2.85, p < .038$. The descriptive statistics related to the mixed model ANOVA are summarized in Table 3.

Table 3

*Word Recognition Score (WRS) Descriptive Statistics for Simple Main Effects (N = 40)*

<table>
<thead>
<tr>
<th>List</th>
<th>Mean</th>
<th>SD</th>
<th>Standard Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test 1</td>
<td>54.0</td>
<td>28.9</td>
<td>2.93</td>
</tr>
<tr>
<td>Test 2</td>
<td>58.6</td>
<td>32.0</td>
<td>2.92</td>
</tr>
<tr>
<td>Test 3</td>
<td>57.6</td>
<td>30.5</td>
<td>2.93</td>
</tr>
<tr>
<td>Test 4</td>
<td>54.3</td>
<td>31.0</td>
<td>2.94</td>
</tr>
<tr>
<td>Retest 1</td>
<td>54.0</td>
<td>27.1</td>
<td>2.93</td>
</tr>
<tr>
<td>Retest 2</td>
<td>53.9</td>
<td>31.2</td>
<td>2.93</td>
</tr>
<tr>
<td>Retest 3</td>
<td>64.3</td>
<td>28.3</td>
<td>2.94</td>
</tr>
<tr>
<td>Retest 4</td>
<td>57.6</td>
<td>30.9</td>
<td>2.93</td>
</tr>
</tbody>
</table>

*Discussion*

The purpose of the present study was to determine the test-retest reliability of WRS scores using updated digital recordings in a sample including a diverse range of hearing
impairment. As speech audiometry materials have been developed, professionals have researched and compared the two presentation methods: MLV and recorded materials. It has been found that while MLV offers greater adaptability with special populations, high-quality recorded materials provide more accurate and consistent speech audiometry test results (Brandy, 1966). By using recorded materials, variables and inconsistencies present in MLV presentation of stimuli were eliminated.

In order to make appropriate inferences about an individual’s hearing ability based upon speech audiometry test results it is necessary to understand the validity and reliability of the test (Ostergard, 1983). Jacobs (2012) discussed the test-retest reliability of SRT scores in a sample of normal hearing individuals. Jacobs (2012) found test-retest reliability for SRT in normal hearing participants to be poor (.47). Jacobs noted that the lack of variability of hearing impairment within her sample influenced the test-retest reliability. In order to increase reliability in the test-retest situation, a population with a diverse range of hearing abilities was used in the present study. Participants’ hearing ranged from normal to severe sensorineural hearing loss, with a mean PTA of 31.2 dB HL and a range of -5.0 to 71.2 dB HL. It is presumed that the diversity present in the current investigation’s sample contributed to the findings of comparatively increased reliability.

The Brigham Young University digital recordings of CID W-22 word lists were found to have strong reliability at 20, 30, and 40 dB SL. Reliability was found to increase as the presentation intensity increased however, reliability was strong in all but the lowest presentation intensity. These test-retest reliability levels support the value of this measure as an audiological assessment tool.
The data reported in Table 2 revealed that the mean difference in WRS in percentage correct between test and retest decreased as sensation level increased. At 10 dB SL, mean difference in percentage correct was calculated to be -3.63%, decreasing to a mean difference of -1.65% at 20 dB SL, .15 % at 30 dB SL, and .05 % at 40 dB SL. The standard deviation of the difference between means accordingly decreased as the presentation level increased, from 10.82 at 10 dB SL, to 9.87 at 20 dB SL, to 9.36 at 30 dB SL, to 5.61 at 40 dB SL.

A mixed model analysis of variance (ANOVA) was calculated to analyze the simple main effects of list, intensity, and time, and to identify any interactions. While the findings in regards to these simple main effects were all expected: no significant difference in WRS scores due to list or time, but an expected significant difference in WRS scores due to intensity, an interaction was noted to be significant between time and re-test. The mean percentage correct in test and retest conditions for the four lists ranged from 53.9 to 58.6 with the exception of list 3 in the retest condition, which averaged 64.3 percent correct. The author speculates that perhaps the content of list 3 test items may have been more distinctive, or memorable, resulting in comparatively higher retest performance in WRS scores compared to the other three lists.

The findings of the present study correlate with the findings of an earlier study of the test-retest reliability of various audiological instruments. Hughes, Arthur, and Johnson (1979) examined the test-retest variability of word recognition score testing using CID W-22 words in a sample of 28 subjects with sensorineural hearing impairment, with PTA losses ranging from 5-23 dB HL. Hughes et al. found that to achieve the minimum acceptable clinical repeatability proposed by Engleberg (1968) of no greater than ± 11% change, it was necessary to perform 3 tests in the case of a SRT+20 dB presentation level, 3 tests at a SRT+30dB level, and 4 tests at a SRT+10 dB level (Hughes et al., 1979). The results of the present study were similar in that test-
retest reliability increased as a function of presentation level. The present study included a sample of individuals with diverse hearing abilities, ranging from a mean PTA of -5.00 to 71.67 dB HL. We speculate that this diversity contributed to the comparatively higher test-retest reliability (.95) at the 40 dB SL presentation level.

**Conclusions**

Variability is one of the shortcomings of speech audiometry testing, as discussed by Ostergard(1983). Examination of the reliability of these measures across time is essential to making appropriate inferences about the hearing system based on information from these tests.

From our findings we conclude that the Brigham Young University recordings are reliable speech stimuli for determining WRS scores. The examination of this measurement in a diverse sample of participants revealed it a reliable test, thus increasing support for its use in the clinical setting. Future research in this area might explore WRS test-retest reliability in a sample including subjects with conductive and mixed hearing losses, WRS test-retest reliability in noise, to simulate a more naturalistic communication environment, and the influence of male and female talkers in a diverse sample.
References


Appendix A

Annotated Bibliography


**Purpose of the work:** This standard outlines the American National Standards of Acoustics’ maximum permissible ambient noise levels (MPANLs) that are acceptable in sound booths and test rooms used for the purpose of audiological evaluations.

**Summary:** The article specifies the appropriate noise levels for the various test conditions, defines all terms, and describes techniques for the measurement of ambient noise and general considerations related to building a room for audiometric testing.

**Conclusions:** Compliance with the described standards is recommended to ensure that test results are not influenced by ambient noise.

**Relevance to the current work:** In today’s audiological practice, it is recommended that professionals comply with the current standards to ensure accurate test results in clinical and research settings.


**Purpose of the work:** The objective of the article is to outline and describe the national standard for manual pure-tone audiometry threshold testing.

**Summary:** This article defines related terms, and describes the national standards for placement and specification of testing equipment, general guidelines for verbal instructions to patients, interpretation of responses, criteria for determining threshold, frequencies to be tested, and guidelines for recording results on the audiogram.

**Conclusions:** While special populations may require modifications to the outlined protocol, compliance to pure-tone testing procedures is generally recommended to promote consistency in manual pure-tone testing.

**Relevance to current work:** When performing manual pure-tone tests, complying with the national standards will promote consistency in practice with audiologists throughout the country.


**Purpose of the work:** This article presents the American Speech Language Hearing Association’s recommendations for determining hearing thresholds.

**Summary:** The article explains requirements for test equipment, test equipment calibration, testing environment, infection-prevention measures, verbal instructions to the patient, and interpretation of responses, threshold determination procedures for air conduction and bone conduction testing, record-keeping guidelines, and possible modifications for special populations.
**Conclusions:** the article describes test conditions and procedures designed to reduce inter-tester and intersite variability. When standard procedures are modified, it is important to note what modifications were used.  

**Relevance to current work:** Abiding by national guidelines helps to promote consistency and decreased variability among speech and hearing professionals nationwide.


**Purpose of the work:** The purpose of these guidelines was to outline protocol for finding speech recognition thresholds (SRT).

**Summary:** The article summarizes the history of speech audiometry, the purpose of speech recognition testing, and its correlation with the individual’s pure-tone average. The testing materials for SRT are homogeneous spondee words. The required response is generally a repetition of the stimulus word, but the response type may be modified for non-verbal individuals, or other special populations. The article also discussed the advantages and disadvantages of monitored live voice stimulus presentation and recorded stimulus materials. The authors describe the procedure for administering the test and calculating the threshold.

**Conclusions:** The guidelines presented are consistent with the current research related to speech audiometry.

**Relevance to the current work:** Adhering to these updated recommendations allows professionals to provide best practice and consistent and accurate test results for clinical and research purposes.


**Purpose of the study:** The purpose of the study was to examine the effects of various acoustically dissimilar recordings on speech audiometry test performance.

**Method:** 25 words were chosen at random from list 3 of the CID Auditory test W-22. These words were randomized three times to create 3 test forms. An adult male recorded these lists on different days, under the same conditions, with the same instructions. Some of the recordings were acoustically altered and controlled to reduce variability among test items. 24 adults were used in the study as listeners. Half of the group was presented with recorded lists, and half with monitored live voice, and asked to write down the words they heard.

**Results:** The data revealed that the results of the live presentation of stimulus materials varied by 9.76 percentage points, while the recorded materials condition yielded variability of only 3.3 percentage points.

**Conclusions:** The authors concluded that the data demonstrated increased consistency in speech audiometry testing with the use of recorded materials.

**Relevance to current work:** Consistency and reliability are necessary characteristics in a sound audiological instrument. It is imperative to control as many variables as possible in the testing environment so that the results are as reliable as possible. A greater
understanding of the benefits of different stimulus presentation methods will aid professionals in providing best practice to their clients.


**Purpose of the work:** The purpose of the chapter is to provide a concise history and overview of speech audiometry stimulus materials, test procedures and important considerations in speech audiometry testing.  
**Summary:** Brandy’s chapter chronicles the history of speech audiometry. Various speech audiometry test procedures are described; among these are speech recognition threshold testing, speech detection threshold testing, word recognition score testing, and sentence recognition testing. Factors affecting these tests are also discussed including presentation method, selection of stimulus materials, and response type.  
**Conclusions:** The chapter provides a summary of audiological procedures included in the category of speech audiometry.  
**Relevance to current work:** This work clearly and concisely describes procedures and considerations of word recognition score testing which is the subject of the present investigation.


**Purpose of the work:** The purpose of the article is to examine the relationship between pure-tone thresholds and speech recognition thresholds.  
**Summary:** The article is a review of numerous studies examining relationships between pure-tone and speech audiometry. Fletcher’s average, an alternative to the traditional pure-tone-average is discussed. Other topics of discussion included the Articulation Index, using regression equations from collected data on the thresholds of various populations, and correction constants.  
**Conclusions:** The data presented in the review demonstrates that speech recognition thresholds are valuable as an approximation of hearing ability for clinical purposes. It is essential to carefully consider the correction factor used. Additionally, the 2-frequency pure-tone-average may be considered as a preferable alternative to the 3-frequency pure-tone average in certain clinical situations.  
**Relevance to the current work:** As the topic of speech audiometry is investigated, it is essential to examine existing studies of these tests. The presented data reflects that speech audiometry can be a valuable clinical tool.


**Purpose of the work:** The purpose of this article is to detail past and present systems of classifying hearing loss, and when each system is best used.  
**Summary:** Several systems have been developed to help describe hearing loss severity. Some of the classification systems have included using percentages, and others have used descriptive words as category classes. While both systems demonstrate some flaws,
professionals have tended to use adjective descriptors to name categories of hearing impairment. The greatest issue currently is the lack of standardization in using and interpreting hearing classifications.

**Conclusions:** While classification systems are concise and familiar descriptors, it is best to describe a person’s hearing in multiple ways to allow for a comprehensive understanding of the hearing system.

**Relevance to current work:** The classification of hearing loss (Scale of Hearing Impairment) described in this work was used in the present investigation to classify participants into categories of hearing impairment to ensure a diverse sample.


**Purpose of the study:** This study sought to examine the test-retest reliability of 2 speech recognition assessments: the speech perception in noise (SPIN) test and the dichotic sentence identification (DSI) test in the elderly population.

**Method:** 17 subjects, ages 63-82 years old with symmetrical sensorineural hearing loss and pure-tone averages (PTA) ranging from 18-52 dB HL participated in the study. The SPIN stimulus materials used were tapes of the revised spin test provided by Dr. Robert Bilger of the University of Illinois. The DSI materials were provided by Auditec of St. Louis. Both tests were administered at three presentation levels: 65, 75, and 85 dB SPL. During the administration of the SPIN tests, babble was introduced at 8 dB signal to noise ratio. The first and second administrations of testing were given with a 1 hour break in between trials. The third testing was performed 1-2 weeks later. All testing was conducted in a sound booth meeting specifications for permissible ambient noise.

**Results:** The reliability of SPIN measures was reported through calculating the standard error of measurement and the standard error of the difference between the repeated measures, and an analysis of variance for repeated measures. The DSI scores were analyzed using the same statistical model as described for the SPIN test. The statistical analysis revealed to significant correlations between age and performance on speech audiometry tests across presentation levels and test trials. Significant correlations were noted between pure-tone sensitivity and performance on DSI and SPIN scores. Despite a strong correlation between left and right ear PTA, the performance on testing was not highly coordinated between ears. Significant test-retest variability was found on all tests, particularly between the first and second test trials. The difference in scores within individuals was great enough to change the diagnostic category of their impairment.

**Conclusions:** The significant test-retest variability in the examined speech audiometry tests questions the usefulness of these tests in classifying hearing impairment in the elderly population. Variability and reliability of other speech audiometry tests should be examined in this population to see if another test could better diagnose impairment.

**Relevance to the current work:** It is essential to understand the variability and reliability of audiometric evaluation tools to ensure that patient are properly treated. The results of the described study promote further research into the test-retest reliability of speech audiometry tools.

Purpose of the study: The objective of Egan’s work is to develop lists of balanced words which are appropriate for speech and hearing tests.

Summary: This article discusses important variables influencing articulation testing including types of test items, speech sounds included, speech lists representing the sound inventory and frequency of sounds in the language, and reliability of test lists.

Method: Words selected were closely matched to Dewey’s vowel counts. In each list, there are about 10 compound consonants in the initial position, and about 10 in the final position. Words were read 11 times to a panel of 11 listeners to measure the approximate relative difficulty of each word. Non-discriminating words, and uncommon words were eliminated. Preliminary lists were read to a panel of 23 listeners and words were rated on a scale of 1-3 with 1 being familiar, 2 being somewhat familiar, and 3 being very unfamiliar. Words with a combined rating of 35 or greater were unused because of unfamiliarity. The remaining words were compiled into 6, 50-word lists.

Conclusion: There are numerous variables related to articulation testing and speech audiometry. Developing phonetically balanced word lists and controlling the test environment can increase the validity and reliability of these tests.

Relevance to the current work: Identifying and controlling variables in audiometric testing is important to providing sound audiological care. Evaluating reliability of such tests is an important area of study.


Purpose of the study: The purpose of this study was to investigate test-retest reliability for in individuals with predominantly conductive hearing loss and those with predominantly sensorineural hearing loss.

Method: Participants for this study included 121 veterans between the ages of 21 through 69 years old. Each participant received an otological and audiological evaluation. All subjects poorest ear had an SRT of no greater than 58 dB HL, allowing subjects to be presented with speech stimuli at suprathreshold levels of up to 40 dB SL. The audiological evaluations consisted of speech recognition threshold testing (SRT), pure-tone testing (both air conduction and bone conduction), and speech discrimination testing. All speech materials were presented by monitored live voice. SRTs were obtained using the 2 dB step descending method. For speech discrimination testing, each participant was presented with one of the 50-word CID W-22 lists. No individual list was presented to a participant more than one time. An attempt was made to use all lists an equal number of times. Each list was presented at 40 dB above the SRT, or at the most comfortable listening level.

Results: The data revealed that the normal hearing participants displayed the smallest test-retest variability. The conductive loss test-group had the next smallest variability, and the sensorineural test group had the greatest test-retest variability. The normal hearing and the conductive loss test groups were more similar to each other than either was to the sensorineural test group. There were no consistent trends in first or second performance being better, thus the authors ruled out the variable of fatigue as a significant concern.
Conclusions: For individuals with a conductive loss, the test-retest variability is small enough that an accurate discrimination score can be obtained from 1 testing, and still fall within the 85 maximum test-retest difference recommendation. The variability of the sensorineural group is large enough that the authors concluded that more than one discrimination score is needed to provide an accurate measure. At least two scores should be obtained. If the difference between scores is greater than 10%, the second score should be used. If the difference is less than 10%, the two scores should be averaged to obtain a representative measure.

Relevance to the current work: The focus of Engleberg’s study is test-retest reliability of speech discrimination, or word recognition score testing, which is similar to the objective of the present investigation. Several differences exist between the two studies, namely the presentation method of the stimuli and the participants used in the study. Examining these differences as well as the results of the two investigations can aid audiologist and speech-language professionals in utilizing speech audiometry instruments appropriately, thus providing best care to their clients.


Purpose of the article: The purpose of the article was to describe new standards, for pure-tone audiometer calibration; the 1964 ISO reference zero.

Summary: This article described the new standards recommended for calibrating pure-tone audiometers, as well as how these new standards could affect clinical practice, namely the classification of hearing impairment of individuals tested. The new system for classifying hearing impairment was outlined in the article.

Conclusions: These new standards in calibration necessitated a change in the classification system for hearing impairment.

Relevance to the current work: Clinically, it is important to have current standards for calibration and consistent systems for classification. This results in consistent assessment as interpretation of test results. In the present investigation, ensuring a diverse sample of hearing abilities was critical to the design of the study.


Purpose of the work: The purpose of Hamid and Brookler’s article is to provide a brief overview about the procedures and purposes of speech audiometry.

Summary: The article briefly describes the 2 most common speech audiometry measures; speech recognition threshold testing and word recognition score testing. Also described are the general procedures for these measures, common stimulus materials used, and the purpose of each of these assessment tools.

Relevance to current work: It is essential to have a basic understanding of these speech audiometry measures and their value as a clinical tool.

Purpose of the work: The purpose of the article is to discuss modifications to auditory tests, and the additional clinical features available in these new assessment tools.

Summary: Certain audiometric tests were the results of developments for military communications. Some of the major frailties of the original speech audiometry tests were included poor list agreement between PAL Auditory tests 9 and 12, and an overwhelming vocabulary required to be successful in tests using Egan’s 20 phonetically balanced (PB) lists. To improve these tests, the quality of recordings was improved, and the vocabulary for the tests using PB lists was restricted, to ensure listener familiarity with the words. With regard to the CID Auditory test W-22, test items were found to vary in difficulty. This list was revised to include 36 equally difficult spondee words. These revisions results in a much more sensitive measure. The CID Auditory test W-2 was also discussed. The CID auditory test W-22, which consists of 200 monosyllabic words divided into 2 50-word lists, was recorded with improved technology for the time, and the lists were checked for equivalency. This resulted in a steeper function and a test with greater sensitivity.

Conclusions: The improvements made to the speech audiometry tests discussed resulted in greater sensitivity and increased equivalency among test items and between lists.

Relevance to current work: Improvements in technology and recording quality can have a marked impact on the usefulness of speech audiometry instruments. The examination of updated recordings promotes best clinical practice.


Purpose of the study: Hughes et al. examined the variability of speech discrimination scores in noise and in quiet conditions.

Method: 50 participants from the ages of 10-82 years old, were chosen with sensorineural hearing loss. The subjects’ PTAs ranged from 10-90 dB HL. Bekesy audiometry was performed in a sound booth with stimuli presented from tape recordings via headphones. 50-word lists from the Ear Research Institute were presented to the subjects three different times. Each participant was presented with the same list each time to reduce variability caused by randomization of lists. The average minimum time between testing sessions was 10 days, so as to eliminate learning effects.

Results: The authors found that the PI functions of normal hearing subjects had a steeper slope than the functions from the results of the participants with sensorineural loss. While the first test session consistently yielded lower test scores, no significant difference was found between the second and third test sessions. The variability found between the test sessions would deem it necessary to perform 3 tests in the case of a SRT+20 dB presentation level, 3 tests at a SRT+30dB level, and 4 tests at a SRT+10 dB level.

Conclusions: Clinically, the authors recommend that at least 3 separate speech discrimination test are necessary to approach appropriate reliability levels, and four tests are needed at low presentation levels.

Relevance to the current work: The work of Hughes et al. emphasizes the need for testing protocols that achieve acceptable clinical reliability. This is necessary to accurately measure the function of the hearing system and effectiveness of intervention. Establishing acceptable reliability is also the focus of the current investigation.

**Purpose of the study:** The purpose of Jacob’s study was to determine the test-retest reliability of updated recordings of speech recognition threshold testing materials.

**Method:** Forty adults between the ages of 18 and 35 were selected for participation in the study. Each participant passed a hearing screening, with thresholds at no greater than 15 dB HL at all test frequencies, tympanometry with a type A tympanogram, and acoustic immittance measures within the normal range. The speech stimuli used in the speech recognition threshold (SRT) testing were digital recordings of the CID W-1 spondaic word list. All participants were familiarized with the stimulus materials by presenting them in written form and aurally at a comfortable listening level prior to testing. The SRTs were obtained using the ASHA recommended 2 dB step descending method. A randomized block design was used to assign the presentation order of the words, the gender of the talker, and the test ear for each participant. Four SRTs were found for each participant in the test condition, then after a short break, four more SRTs were found for the retest condition.

**Results:** Jacobs found the test-retest reliability to be poor (.47), with slightly increased performance overall in the retest condition. Jacobs attributed this to a learning effect, and did not consider it significant. Jacobs found no clinically significant difference in performance with a male talker compared to a female talker.

**Conclusions:** Jacobs attributed the poor test-retest reliability to the low variability in between subjects. While the discrepancy between test and retest thresholds falls between the margins of clinical error, Jacobs suggests evaluating the test retest reliability of the recordings with a sample of participants including a wide variety of hearing impairment.

**Relevance to the current work:** Jacob’s suggestion to evaluate the test-retest of the speech stimuli with a more varied population is the objective of the current work. Comparing the results of Jacob’s study to the present investigation reveals the importance of using a diverse sample in such studies.


**Purpose of the study:** The purpose of the study was to review audiological practices in the United States.

**Method:** An 8-page survey was sent to 500 randomly selected audiologists from the American Auditory Society. Each question was followed by objective response options, as well as the option to provide an alternative response.

**Results:** The information from the 258 eligible respondants revealed that 98% of respondents administer speech recognition threshold testing, 53% familiarize their patients with stimulus materials prior to testing, and 90% used monitored live voice to present stimuli. Word recognition score (WRS) tests were administered regularly by 99% of respondants, 53% using the CID auditory test W-22, 42% using the NU-6 lists, and 805 using monitored live voice to present stimuli. 70% reported that they performed WRS at 40 dB SL with a range of 10 to 80 dB SL. 60% of respondants always used half-lists of 25 words, and only 3% use 50 word lists at all times. The article also details
participant responses regarding pure-tone testing, masking in pure-tone testing, immitance testing, central auditory processing testing, electrophysiological tests, hearing aids, and counseling.

**Conclusions:** The study indicates that while professionals agree that recorded stimulus materials for speech promote consistency, monitored live voice is still used clinically. Also, audiologists tend to use half lists more frequently than the full 50-word lists.

**Relevance to current work:** As test-retest reliability is examined, it is valuable to consider the commonplace clinical practices in regards to speech audiometry. Ideally, common practice will align with what research purports to be best practice.


**Purpose of the work:** The purpose of the article is to identify and discuss factors related to the validity and reliability of speech audiometry materials and testing.

**Summary:** When an individual is given a test multiple times, scores will be somewhat different due to statistical variation, however controlling test variables and parameters could result in more consistent testing. Purposes of audiological tests are reviewed, measurement instruments are discussed, and content validity, criterion-related validity, and construct validity are defined and discussed. Reliability is discussed as a necessary quality of audiological tests, so that subsequent improvements in scores can be attributed to treatment or amplification in a clinical setting. Errors in testing are discussed, false positives and false negatives, which affect the proper identification of individuals that need intervention to assist their hearing. While false positives negatively affect the efficiency of testing, false negatives, failing to identify those who failed the test, is worse. Thornton and Raffin’s binomial model for speech discrimination testing is discussed. The later portion of the article lists several items to consider when attempting to control variability. Speech reception thresholds can be influenced by word list, the list item, and how many list items are administered. While not all variables can be controlled, test conditions and test materials can be improved to increased reliability. Using recorded materials when appropriate, standardizing word familiarity, and using longer lists are some of the practical considerations suggested to improve reliability.

**Conclusions:** Speech audiometry tests are not as precise as the historically have been viewed. Achieving the perfect speech audiometry test is unlikely, however efforts should focus on developing and improving tests keeping in mind the specific purpose of the measurement. Reliability and validity should be carefully examined.

**Relevance to the current work:** Understanding the variability and reliability of speech audiometry tests is critical to providing sound clinical care. Examining such characteristics of tests is a subject worthy of investigation.


**Purpose of the work:** The purpose of the article was to report confidence levels for differences in speech discrimination scores.
Summary: The article consists of several tables reporting the probabilities of chance differences between any two scores for tests of any length. Conclusions: These tables provide a good reference for audiologists seeking to determine the probability that differences in between speech discrimination scores are due to chance alone. Relevance to the current work: In the interest of test-retest reliability, it is valuable to note under which circumstances differences in scores can be attributed to chance, and when the are significant.


**Purpose of the study:** The purpose of the study was to examine test-retest variability of pure tone audiometric testing across age groups using three types of insert phones. **Method:** A 30 normal hearing participants were assigned a group according to age: group 1: 6-9 years old (10 subjects), group 2: 10-23 years old (10 subjects), and group 3: young adults (10 subjects). Subjects were screened to ensure they met the criteria of thresholds no higher than 20 dB HL in at least one ear between the frequencies of 250 and 8,000 Hz. Testing was conducted in a double-walled sound booth with a clinical audiometer, both pieces of equipment meeting national specifications for ambient noise in test booths, and audiometer calibration. 6 frequencies were tested (250, 500, 1,000, 2,000, 4,000, and 8,000 Hz). **Results:** A two-way analysis of variance (ANOVA) was calculated for each test frequency across age and earphone condition. The results revealed no significant differences between age groups or earphone conditions for test-retest in pure-tone testing. The variability of test-retest was noted to be greater in extreme low and extreme high frequencies. This finding aligns with results of other studies addressing test-retest reliability with supra aural earphones. Also, there were no significant interactions between age groups and earphone condition for each test frequency. **Conclusions:** This suggests that test-retest variability in pure-tone testing is not affected by age nor insert-phone condition. Two factors were identified as contributors to variability in test-retest reliability: subject variance (motivation, behavior), and instrument variance. **Relevance to the current work:** While no significant effects were seen relating age or earphone condition to test-retest variability in pure-tone testing, it is possible that equipment and procedural limitations prevent researchers from being able to identify such interactions. Future research could examine the possibility of these interactions under more stringent standards and more conservative threshold procedures.


**Purpose of the study:** The purpose of the article is to test a simple, binomial model for describing the variability across forms of speech discrimination tests. **Method:** The binomial model were compared to the 4,120 test results from hearing impaired participants using the CID auditory test W-22 test for speech discrimination
testing. Each participant was presented with the 4 lists at 40 dB above the speech recognition threshold, when possible. The standard deviation of binomial variables was compared to the standard deviations of the scores on subtests of the 50-word lists. Knowing about the form of the distribution allows inferences to be made about the population from a smaller sample. This is expressed in the study as a confidence interval.

**Results:** The binomial model considers the variability between test forms on both the basis of the number of test items, and the individual’s true score. Determining the appropriate number of test items is accomplished best by examining the estimated true score in the population and considering the balance between test length and the increased variability that can result from decreased test length.

**Conclusions:** For the sake of efficiency, it would be desirable to reduce test length, however the authors conclude that this cannot be accomplished without significant changes to test structure and content. This binomial model of examining assessment tools may perhaps be applied to other tests in the field of audiology, provided that they meet certain criteria. This would be especially useful in tests used to classify patients into categories or measure progress.

**Relevance to current work:** In considering the length of list to use for speech audiology testing it is critical to understand the differences in variability resulting from the list length.


**Purpose of the study:** the purpose of the article is to present and describe N.U. Auditory test No. 6 (an expanded version of N.U. Auditory test No. 4). High equivalence between lists and strong reliability characterized the N.U. Auditory test No. 4, however the corpus of monosyllabic stimulus words was deemed too restrictive.

**Method:** The article describes the process of creating the 4 phonemically balanced lists of monosyllabic words. The phonemic distribution established by Lehiste and Peterson was used, and the first step was to select 4 independent groups of words that preserved this phonemic distribution. Each word list was randomized four times resulting in four test forms. The relative familiarity of the words was also evaluated and determined to be similar to the Peterson-Lehiste revised CNC lists. These new materials were evaluated for list equivalency, test-retest reliability and other characteristics using a sample of 36 individuals; 24 with normal hearing and 12 with sensorineural loss. Stimuli were then recorded on magnetic tape. Participants were presented with lists at the following sensation levels above the speech recognition threshold: -4,0,8,16,24, and 32 dB SL.

**Results:** The results of participants with normal hearing yielded equivalent slopes between lists, and little change was noted between test and retest, indicating good list equivalence and strong test-retest reliability. Like the previous test (N.U. Auditory test 4) the curve of the slopes reaches an asymptote where improvement slows with presentation increases, until perfect discrimination is approximated. In the previous test, the plateau point was reached at 24 dB SL, while in the present investigation, the plateau was not reached until 32 dB SL. The standard deviations revealed a trend towards increased variability in scores in the middle of the slopes and decreased variability at the extreme
ends. These findings were very similar to evaluations of the previous test. In the group tested with sensorineural loss, list equivalency was found to be strong. The sensorineural loss test group demonstrated several differences however, from the normal hearing test group. The slope of the performance-intensity curves, while equivalent were not as steep, variability of scores about the mean score was increased at each presentation level. While the normal hearing participants demonstrated little difference in performance from list to list, the hearing loss test group demonstrated greater variability in performance in response to list, indicating list equivalency may not be preserved when presented to individuals with hearing loss. In both populations, performance of retest tended to be higher than test. The test-retest differences were not significant enough to cause concern.

**Conclusions:** The authors conclude that the N.U. Auditory test 6 is comparable in list equivalency and test-retest reliability to the previously developed N.U. Auditory test 4. The newer test, however contains twice the vocabulary present in the older test, providing speech and hearing professionals with a broader corpus of words to use in assessing patients.

**Relevance to the current work:** Studying the reliability and list equivalency of other speech discrimination materials is pertinent as the current work is seeking to establish similar measures for updated recordings of another set of stimuli.
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 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
## Appendix C

### CID W-22 Word Lists

<table>
<thead>
<tr>
<th>List 1</th>
<th>List 2</th>
<th>List 3</th>
<th>List 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>ace</td>
<td>me</td>
<td>move</td>
<td>add</td>
</tr>
<tr>
<td>ache</td>
<td>mew</td>
<td>air</td>
<td>new</td>
</tr>
<tr>
<td>an</td>
<td>none</td>
<td>and</td>
<td>now</td>
</tr>
<tr>
<td>as</td>
<td>not</td>
<td>bin</td>
<td>oak</td>
</tr>
<tr>
<td>bathe</td>
<td>or</td>
<td>by</td>
<td>odd</td>
</tr>
<tr>
<td>bells</td>
<td>owl</td>
<td>cap</td>
<td>off</td>
</tr>
<tr>
<td>carve</td>
<td>poor</td>
<td>cars</td>
<td>one</td>
</tr>
<tr>
<td>chew</td>
<td>ran</td>
<td>chest</td>
<td>own</td>
</tr>
<tr>
<td>could</td>
<td>see</td>
<td>die</td>
<td>pew</td>
</tr>
<tr>
<td>dad</td>
<td>she</td>
<td>does</td>
<td>rooms</td>
</tr>
<tr>
<td>day</td>
<td>skin</td>
<td>dumb</td>
<td>send</td>
</tr>
<tr>
<td>deaf</td>
<td>stove</td>
<td>ease</td>
<td>show</td>
</tr>
<tr>
<td>earn</td>
<td>them</td>
<td>eat</td>
<td>smart</td>
</tr>
<tr>
<td>east</td>
<td>there</td>
<td>else</td>
<td>star</td>
</tr>
<tr>
<td>felt</td>
<td>thing</td>
<td>flat</td>
<td>tare</td>
</tr>
<tr>
<td>give</td>
<td>toe</td>
<td>give</td>
<td>that</td>
</tr>
<tr>
<td>high</td>
<td>true</td>
<td>ham</td>
<td>then</td>
</tr>
<tr>
<td>him</td>
<td>twins</td>
<td>hit</td>
<td>thin</td>
</tr>
<tr>
<td>hunt</td>
<td>up</td>
<td>hurt</td>
<td>too</td>
</tr>
<tr>
<td>isle</td>
<td>us</td>
<td>ice</td>
<td>tree</td>
</tr>
<tr>
<td>it</td>
<td>wet</td>
<td>ill</td>
<td>way</td>
</tr>
<tr>
<td>jam</td>
<td>what</td>
<td>jaw</td>
<td>well</td>
</tr>
<tr>
<td>knees</td>
<td>wire</td>
<td>key</td>
<td>with</td>
</tr>
<tr>
<td>law</td>
<td>yard</td>
<td>knee</td>
<td>yore</td>
</tr>
<tr>
<td>low</td>
<td>you</td>
<td>live</td>
<td>young</td>
</tr>
</tbody>
</table>