1993

# The development of a computerized version of a revised W-1 word list for obtaining speech reception thresholds

Sara C. Meyers

# The Development of a Computerized Version of a Revised W-1 Word List for Obtaining Speech Reception Thresholds

An Independent Study by:

Sara C. Meyers

Supervisor:

Dr. Gerald R. Popelka

April 27, 1993

# INTRODUCTION

## Speech Reception Threshold Testing

Speech threshold testing has various forms in clinical audiology. Speech reception thresholds (SRTs) generally refer to the intensity level (dB HL) in which one can identify 50% of spondaic words. Spondaic words are two-syllable words with equal stress per syllable. Other forms of speech testing include speech detection threshold testing and speech discrimination testing.

Speech detection threshold is simply as the name suggests, as this is the level when speech is detected and not necessarily identified. Speech discrimination testing usually is a method of determining discrimination ability for out-of-context one-syllable words when presented at an amplified level.

Speech reception thresholds are generally used as a way to validate pure tone thresholds and provide a basis for determining discrimination presentation levels. Other terms used for SRT are spondaic word threshold or spondee threshold.

## History

Speech thresholds were initially obtained by speaking or whispering to the listener from measured distances. Inconsistent intensity measures made test results difficult to quantify. As a result, clinical speech threshold testing was developed to obtain "more controlled and reliable measures" (Hopkinson, 1978).

In 1926, Bell Telephone Laboratories developed a phonographic recording of spoken digits called the Western Electric 4A (now 4C) (Fletcher and Steinberg, 1929). Hudgins, et al (1947) later developed the Harvard's Psycho-Acoustic Laboratory (PAL) Auditory test #9 and #14 based on four characteristics: familiarity, phonetic dissimilarity, normal samples of English speech sounds, and homogeneity with respect to audibility. Homogeneous audibility is when each word is recognizable at the same intensity level presented.

Their goals were to resolve problems during construction of the auditory test, provide an accurate measure for all degrees of hearing loss, and differentiate between sloping and flat hearing losses.

Hudgins, et al, (1947) chose spondaic words. When relative intensity was compared among spondees, disyllables, and monosyllables, spondees required less intensity than the other two to obtain 100% correct recognition. The PAL Auditory Test #9 had 2 lists of 42 words with 6 scrambles per list. On the recording, each group of 6 words, after the first six, is attenuated 4 dB relative to the previous set of six. The PAL #14 words are presented at the same level across the list.

Development of the W-1 and W-2

Hirsh, et al, (1952) developed the Central Institute for the Deaf (CID) W-1 and W-2 to alleviate problems with the PAL #9 and #14. The tests were modified in terms of familiarity, phonetically

balanced samples, and via use of magnetic tape recordings. The CID W-1 and W-2 consisted of 6 scramblings of 36 spondaic words.

The W-1 words were recorded at the same level with each word 10 dB below the carrier phrase, "Say the word..." A 1000 Hertz tone is recorded at the beginning of the record so the audiometer may be calibrated. The tone is of equal loudness as the carrier phrase. The W-1 allows the audiologist to manually control the presentation level of each word.

The W-1 was created from 84 spondee words on the PAL list #9. Judges used a 3-point rating scale to determine familiarity of each word. After the most familiar words were selected and recorded, intensity levels were compared. Six listeners were tested using the PAL #9 and listened to 6 recordings of the W-1 at +4, +2, 0, -2, -4, and -6 dB relative to their PAL #9 thresholds.

Results showed some words were easier than others, as seen when presented at 4 and 6 dB *below* PAL thresholds yet identified correctly by 5 or more listeners. On the other hand, some words presented 2 and 4 dB above PAL thresholds were missed by all listeners in 5 or more trials (Hirsh, et al, 1952). These extreme easy and difficult words were eliminated, leaving 36 words.

Each word was recorded separately onto a record and then dubbed to tape. The carrier phrase was added with a six second block from carrier phrase to word to pause allowed for a response.

From the original recording, each word was spliced and redubbed to form six scrambles, A-F. Further tests revealed some words were still easier or more difficult than others by a range of ±2 dB. Due to these findings, the easier words were recorded 2 dB lower and the more difficult words were recorded 2 dB louder.

A repetition of the latter test proved each word to be homogeneous with respect to audibility and any additional variations would be ruled as chance.

The CID W-2 was created with the same words as the W-1. On each recording, the level of intensity decreases 3 dB per set of 3 words. The order of words are the same for each scramble as the W-1.

For the purpose of this Independent Study, the W-2 will not be referred to for the computer application sections. For further information of the W-2, please refer to Hirsh, et al., 1952.

Physical Characteristics of Speech

One factor that plays an important role when administering speech tests is the root-mean-square (RMS) sound pressure of each word. Fletcher (1929) showed when words are measured by individual speech sounds, the loudest vowel may be as much as 25 dB greater than the least intense consonant.

The sound pressure level of speech stimuli varies through both time and frequency. Since thresholds are dependent on SPL, it must

be measured correctly to obtain accurate results for speech thresholds.

The American National Standard Specifications for Audiometry (S3.6-1969) developed the current list of specifications for speech SPLs. It stated,

> "...the SPL of a speech signal at the earphone is defined as the rms SPL...of a 1000 Hertz signal adjusted so that the VU meter deflection produced by the 1000 Hertz signal is equal to the average peak VU meter deflection produced by the speech signal."

Unfortunately, due to the varying sound pressure levels in speech, the "average peak VU meter deflection" is not easy to find.

Presentation Mode

When speech tests are administered via monitored live voice (MLV), the speaker must monitor the VU meter as each word is presented. It was believed that with the VU meter averaging at 0, the list would be homogeneous with respect to audibility.

One advantage of MLV is the ability to adapt the presentation style to benefit the client. If the client was elderly or very young, the audiologist could modify the presentation rate (Popelka and Newby, 1992). However, it is difficult for one speaker to present a word at the same level as another speaker or for the same speaker to present the same word equivalently from one test to another. Another problem encountered was that the testers were choosing words randomly from the list, thus introducing tester bias.

Test-retest reliability and validity suffered due to the inconsistency of this self-monitoring method. For those reasons, tape recordings of the speech tests were developed and were preferred due to the pre-adjusted levels of speech according to a calibrated tone.

One advantage to the tape recorded test administrations include reduced tester bias, improved test-retest reliability, and homogeneity with respect to audibility. Unfortunately, as the tape was used repeatedly, the recording quality decreased and the tape became worn out. Furthermore, if the client could not keep up with the tape speed, or the tester wanted to select words for a child within his/her vocabulary, the tape could not be used.

Computerized speech

Recently, the W-1 was recorded on a compact disc (CD) which recorded analog signals to digital and played digital signals to analog form. This sound quality is better than tape, and there is no degradation over time. However, even a CD player does not allow one to choose words randomly.

Further digital advances enable any sound to be recorded onto a computer via live voice or direct audio input. The Macintosh computer system has such an application where sound is transmitted via a device called MacRecorder onto a software application called Sound Edit. The input can be altered by numerous functions such as filters or other general edit commands.

With the computer advances into the audiology field, computerized programs are becoming more prevalent. There are computer interfaces that enable audiologists to test, save and print information via the computer medium (Fournier and Margolis, 1992; Palmer, 1992).

## PURPOSE

The purpose of this independent study is to create a condensed list SRT application of the W-1 for the Macintosh to allow good test-retest reliability, manual selection of words while retaining homogeneity of audibility, and the beginning for audiometer interfacing through an audiometer (Grason-Stadler GSI-16).

## DEVELOPMENT OF TEST

### Selection of words

The CID W-1 and W-2 was created to ensure phonemic dissimilarity while preserving homogeneity of audibility. Several studies have shown that further reduction of the list size is possible while obtaining equivalent scores as obtained on a full list.

Bowling and Elpern (1961) studied the CID W-1 test words to determine the range of audibility, or the softest to loudest level at which each word is first understood. They found the range of intelligibility to be 10 dB when measuring the sensation level (SL) each word was first understood. This was not acceptable to them, and they suggested limiting the list to 22 words that were within a 3.5 dB range (Table 1).

Then in 1975(a), Beattie, Edgerton, and Svihovec also looked at the range of intelligibility and felt a range of 1.5 dB SL was more appropriate. They included 18 words and tested subjects' articulation functions when tested by MLV. Results showed the articulation functions with 18 words and with 36 words had equal slopes of 12% per dB. They believed these 18 words (Table 2) represented a more homogeneous group of words with respect to audibility.

Later that year, these same researchers performed a very similar study on the Auditec of St. Louis recording. Again they proposed the use of a test consisting of 18 words rather than the original 36, explaining the 18 words were a more homogeneous group. The 18 words were not completely similar as their first study, but again they found the articulation functions equal when the SRT obtained with a complete list was compared to that using only 18 words (Beattie, Svihovec and Edgerton, 1975b)

Although each study suggested that a condensed list was capable of obtaining equivalent scores, none of these studies agreed on which words to use. In fact, from those three studies, only four words were used simultaneously.

Young, et. al. performed a research study to develop a shorter word list that would retain homogeneity of audibility and alleviate the limitations found with the previous studies(1982). Young, et. al.

believed the earlier studies were limited in that they did not account for speaker effects, the learning effects, or the rate each word became intelligible 50% of the time. The third factor accounted for comes into effect when a spondee has a particularly steep slope for the psychometric function. This would be affected in clinics where a 5 dB step size is utilized (1982).

Another differing factor in previous studies discussed was *how* the researchers chose what a homogeneous group of words was. Young, et. al. defined a homogeneous group of words as the words that "yielded thresholds and slopes with $\pm 1$ SD of the mean value" (1982). Using their criteria, a list of 15 spondees was developed that were homogeneous for both threshold intensities and slope of intelligibility function.

This condensed list of 15 words was approved and recommended by ASHA in the "Guidelines for determining threshold level for speech" (ASHA, 1988).

Creating the Computer Program

The 15 chosen spondees were played from the W-1 recording on a Technics Compact Disc player (SL-PG300) through direct audio input to the MacRecorder GZ53M5MR200A. This was recorded directly into the Sound Edit program of a Macintosh IIci computer. Just prior to recording, the peak levels of the recorded words were monitored in the Sound Edit program, and the volume control of the

MacRecorder was adjusted so no peak levels would be clipped, thus preventing distortion.

As each word was recorded, three functions were performed to create the word as its own "button" on the finished test. First, the word was isolated, involving the deletion of the silent gaps prior to and following each word. Second, each word was copied to Hypersound, a software application where waveforms can be recorded. This is the medial step to its final placement into Hypercard, a software application where waveforms can be played.

Once into Hypercard, the word is assigned to its own "button" that can be manipulated by shape, text or function. Each of the buttons for the 15 chosen spondees were named solely by the word and placed in three columns of five.

Six additional buttons were created to supply six scrambles of the fifteen words. A calibration button was made using the calibration tone on the compact disc. The recorded words on the compact disc were in calibration with the tone, and recorded for the program at the same recording level. A final button, "stop", was created to halt production of the calibration tone or words before automatic completion. (Table 3).

## RELIABILITY AND VALIDITY STUDY

Once the computerized application of the W-1 condensed list was created, reliability and validity data were obtained to ensure

equivalent scores would be obtained regardless of presentation mode.

## METHODS AND PROCEDURES

### ASHA Recommended Testing Procedures

ASHA recommends two testing methods for obtaining speech reception thresholds. One is the ascending method, and the second is the descending method. The method used for this study is the descending method. For complete description of the ascending method, the reader is referred to ASHA, 1979.

#### Descending method:

Prior to the main descent, a preliminary stage is performed to determine the starting intensity level. Beginning at a level approximately 30 dB above anticipated SRT level, the subject is given one word. After repeating that word correctly, the level is decreased 10 dB and a second word is given. This is done until the subject misses two words in a row. 10 dB is added to the ending level of the preliminary stage to obtain the starting level for the testing stage.

From the starting level, two words are given in decreasing 2 dB increments. The subject must correctly identify the first 5 of 6 words to continue. The test is continued until the subject misses 5 of 6 words. Using a formula, the SRT is obtained. This is also possible using 5dB increments.

Subjects

The subjects were 10 adults (4 males and 6 females), ranging in age from 24 - 27 years. Each subject was a graduate student in the Professional Education Program at Central Institute for the Deaf. All of the subjects exhibited pure tone thresholds of 20 dB HL or better at 250, 500, 1K, 2K, 4K, and 8K Hertz.

Instructions:

One ear was tested per subject for this study. Prior to pure tone testing, each subject was instructed to raise his/her hand whenever he/she heard the beep. Once thresholds were found to be within normal limits across all frequencies, these instructions were given for the SRT test using the CD player:

> "I am going to read to you a list of words. For now, I just want you to listen to each word. If you don't understand any of the words, stop me. (Read list). Now you will hear a recording of a woman's voice saying the same words I just read to you, one at a time. The level of the words will get softer and softer. I want you to repeat each word back me and don't hesitate to guess."

After the CD player was used, I read the shorter list of words from the computer version and instructed him/her to do the same as he/she did previously.

RESULTS

Each subject scored an SRT within $\pm$4 dB between the CD and computerized versions of the test. This 4 dB difference is within the

±5 dB standard deviation for test-retest reliability, and is not believed to be a significant difference. (Table 4).

When each of the subject's scores were averaged together, the average score using the CD player was 6.5 dB. The average SRT score when tested with the computer program was 7.2 dB. This .7 dB difference is well within one standard deviation.

With a three frequency pure tone average, each subject's SRT scores were within ±5 dB from their PTA. (Tables 5 and 6).

These results indicate the computerized application of the condensed SRT list results in scores equivalent to those obtained with the conventional CD recorded list.

## CLINICAL USE AND APPLICATION

This computerized application can be interfaced with an audiometer and may save the clinician and patient time. With this test, the word list will take less time to familiarize with the subject. If working with an elderly patient or a patient with reduced vocabulary, this test will enable the clinician to manually adjust the rate of presentation or words to present.

Although results were not unanimous with one mode of recording scoring better than the other, some of the subjects mentioned the word intensities from the computerized recording were softer than the compact disc recording even when the tester

stated the tests were begun at the same intensity level. However, when the computerized recordings were made, input levels were monitored so that peak levels were as high as possible without being clipped. If the recording input was increased to attempt to equalize the level of output between the CD player and computer, distortion would be highly probable.

Since the differences between CD and computerized speech were within 5 dB, using a 5 dB step size would have resulted in the same results. If this computer application is to be interfaced with an audiometer, perhaps this intensity difference would be taken into account during calibration.

However, for the purpose of this study, it was determined that the computer application with the fifteen recommended spondee words provides SRT scores of equivalent value as using a CD player. In addition, it provides a shorter homogeneous list of spondee words to save time, provide accurate manual use, provide the best quality of recorded speech, and possible interfacing with an audiometer.

Table 1

Airplane
Armchair
Birthday
Cowboy
Eardrum
Farewell
Greyhound
Horseshoe
Hotdog
Iceberg
Mousetrap
Northwest
Oatmeal
Pancake
Playground
Railroad
Sidewalk
Stairway
Sunset
Toothbrush
Whitewash
Woodwork

22 word list exhibiting a 3.5 dB range of intelligibility proposed by Bowling and Elpern, 1961.

Table 2

Birthday
Cowboy
Farewell
Hardware
Headlight
Horseshoe
Iceberg
Inkwell
Mushroom
Northwest
Padlock
Playground
Railroad
Schoolboy
Stairway
Sunset
Toothbrush
Workshop

18 word list with greater homogeneity of audibility proposed by Beattie, Edgerton, and Svihovec, 1975a.

TABLE 3

# CENTRAL INSTITUTE FOR THE DEAF
# W-1 CONDENSED TEST LIST

| | | |
|---|---|---|
| BASEBALL | INKWELL | RAILROAD |
| DOORMAT | MOUSETRAP | SIDEWALK |
| DRAWBRIDGE | NORTHWEST | TOOTHBRUSH |
| EARDRUM | PADLOCK | WOODWORK |
| GRANDSON | PLAYGROUND | WORKSHOP |

PLAY ALPHABETICALLY

STOP

| | | |
|---|---|---|
| LIST 1 | LIST 3 | LIST 5 |
| LIST 2 | LIST 4 | LIST 6 |

CALIBRATION TONE

TABLE 4

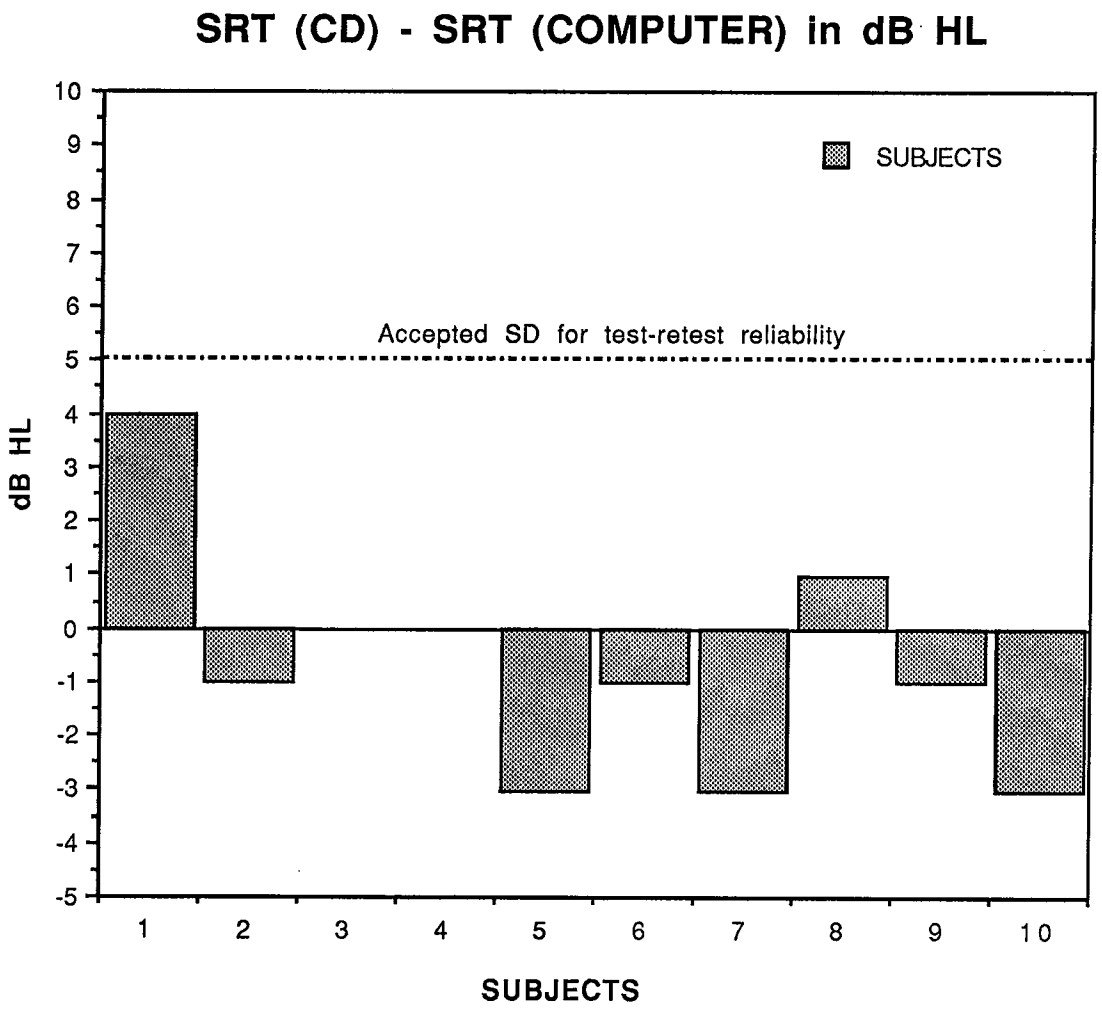

SRT (CD) - SRT (COMPUTER) in dB HL
SUBJECTS
Accepted SD for test-retest reliability
dB HL
10
9
8
7
6
5
4
3
2
1
0
-1
-2
-3
-4
-5
1
2
3
4
5
6
7
8
9
10
SUBJECTS

TABLE 5

## SRT (COMPUTER) - PTA in dB HL

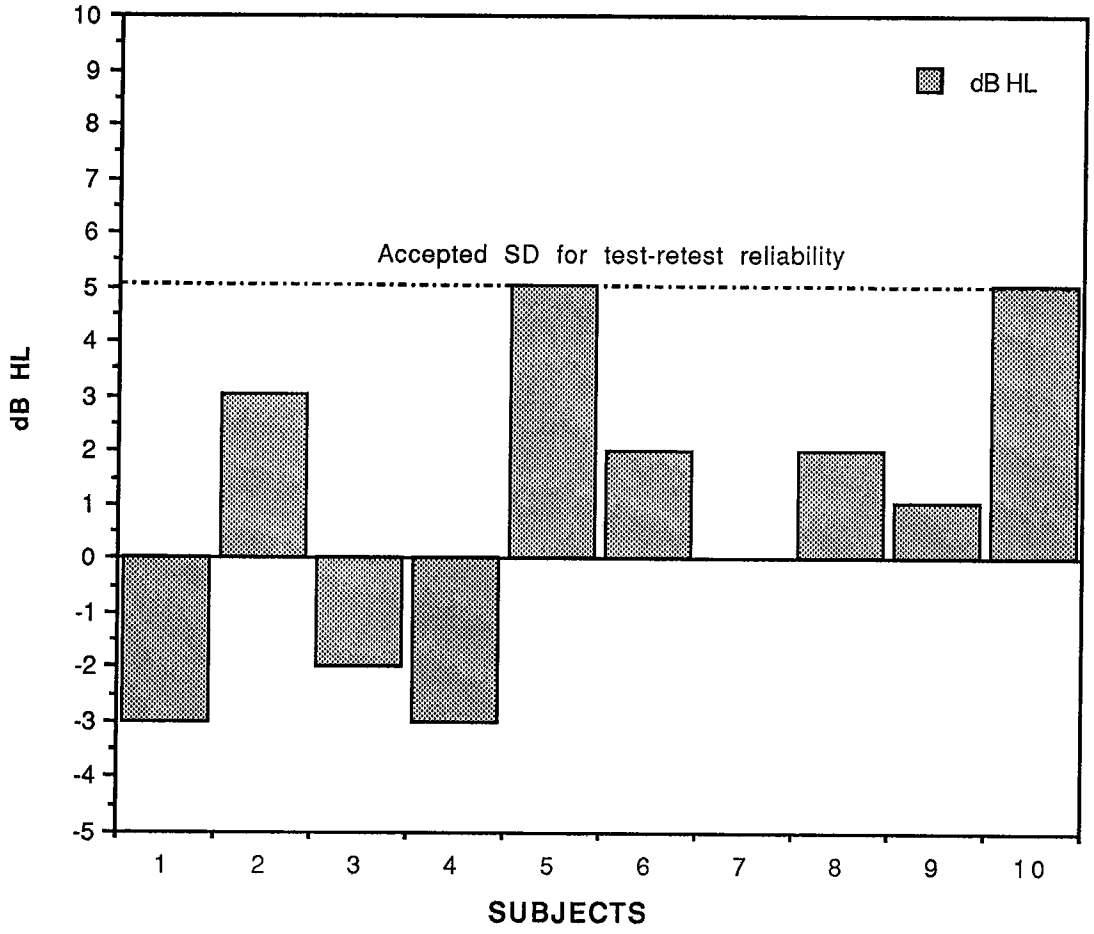
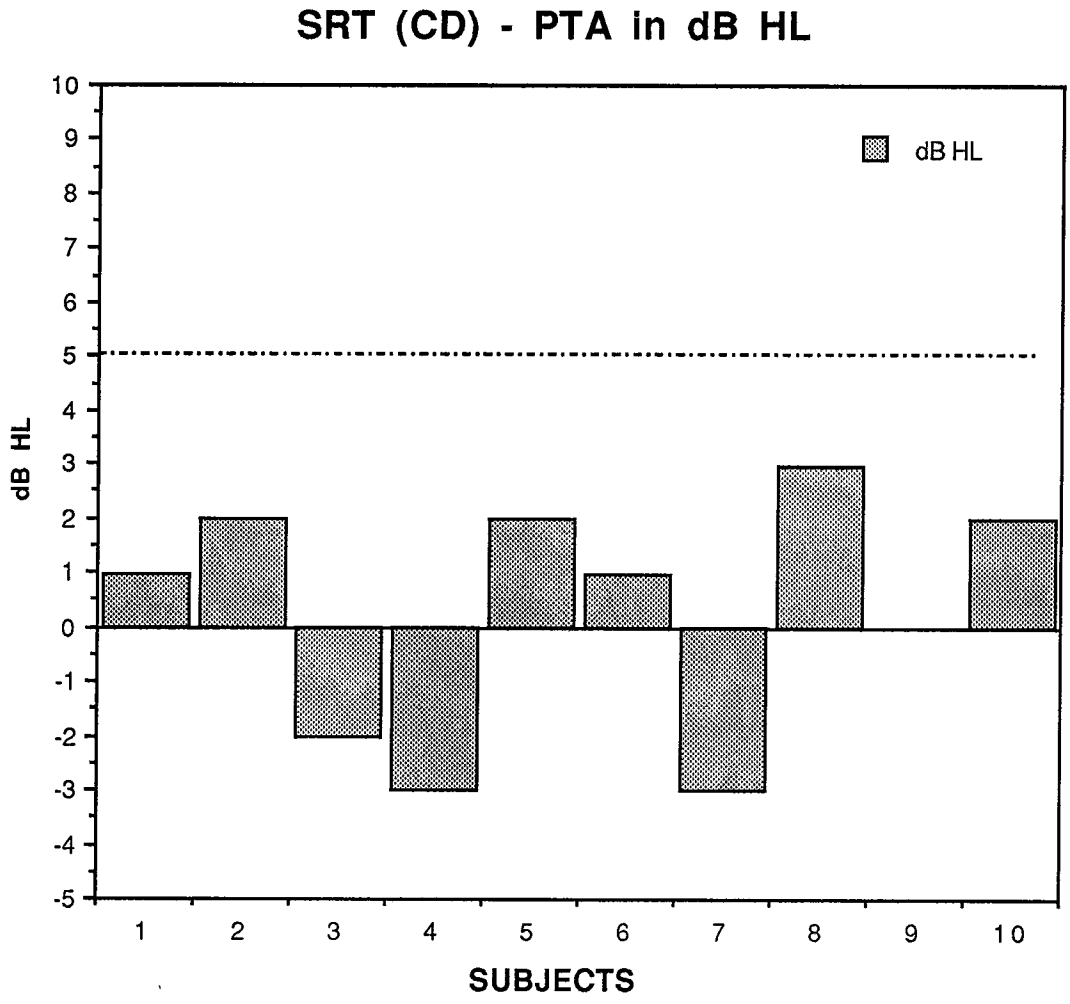


TABLE 6


SRT (CD) - PTA in dB HL
dB HL
10
9
8
7
6
5
4
3
2
1
0
-1
-2
-3
-4
-5
dB HL
1
2
3
4
5
6
7
8
9
10
SUBJECTS

## REFERENCES


ANSI. American National Standard Specification for Audiometers. ANSI S3.6-1969. American National Standards Institute. New York: Acoustical Society of America.

ASHA. Guidelines for determining threshold level for speech. ASHA, 1988, 30, 85-89.

ASHA. Guidelines for determining the threshold level for speech. ASHA, 1979, 21, 353-355.

Beattie, R.C., Edgerton, B.J., and Svihovec, D.V. An investigation of the Auditec of St. Louis recordings of the Central Institute for the Deaf spondees. Journal of the American Audiology Society, 1975a, 1(3), 97-101.

Beattie, R.C., Svihovec, D.V., and Edgerton, B.J. Relative intelligibility of the CID spondees as presented via monitored live-voice. Journal of Speech and Hearing Disorders, 1975b, 40, 84-91.

Bowling, L.S., and Elpern, B.S. Relative intelligibility of items on CID Auditory Test W-1. Journal of Auditory Research, 1961, 1, 152-157.

Fletcher, H. Speech and Hearing. New York: Van Nostrand, 1929.

Fletcher, H. and Steinberg, J.C. Articulation testing methods. Bell Telephone Systems Technical Publications, 1929, 8, 806-854.

Hopkinson, N.T. Speech Reception Threshold. In J. Katz (Ed.), Handbook of Clinical Audiology (2nd Ed.). Baltimore, MD: Williams and Wilkins Co., 1978.

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

Newby, H.A., and Popekla, G.R. Audiology (6th Ed.). Englewood Cliffs, N. J.: Prentice Hall, 1992.

Young, L.L., Dudley, B., and Gunter, M.B. Thresholds and Psychometric Functions of the Individual Spondaic Words. Journal of Speech and Hearing Research, 1982, 25, 586-593.