

A Comparison of Word-Recognition Performances on the Auditec and VA Recorded Versions of Northwestern University Auditory Test No. 6 by Young Listeners with Normal Hearing and by Older Listeners with Sensorineural Hearing Loss Using a Randomized Presentation-Level Paradigm

DOI: 10.3766/jaaa.17135

Richard H. Wilson*

Abstract

Background: The Auditec of St. Louis and the Department of Veterans Affairs (VA) recorded versions of the Northwestern University Auditory Test No. 6 (NU-6) are in common usage. Data on young adults with normal hearing for pure tones (YNH) demonstrate equal recognition performances on the two versions when the VA version is presented 5 dB higher but similar data on older listeners with sensorineural hearing loss (OHL) are lacking.

Purpose: To compare word-recognition performances on the Auditec and VA versions of NU-6 presented at six presentation levels with YNH and OHL listeners.

Research Design: A quasi-experimental, repeated-measures design was used.

Study Sample: Twelve YNH ($M = 24.0$ years; PTA = 9.9-dB HL) and 36 OHL listeners ($M = 71.6$ years; PTA = 26.7-dB HL) participated in three, one-hour sessions.

Data Collection and Analyses: Each listener received 100 stimulus words that were randomized by 6 presentation levels for each of two speakers (YNH, -2 to 28-dB SL; OHL, -2 to 38-dB SL). The sessions were limited to 25 practice and 400 experimental words. Digital versions of the 16, 25-word tracks for each session were alternated between speakers.

Results: Each of the 48 listeners had higher recognition performances on the Auditec version of NU-6 than on the VA version. The respective overall recognition performances on the Auditec and VA versions were 71.4% and 64.1% (YNH) and 68.7% and 58.2% (OHL). At the highest presentation levels, recognition performances on the two versions differed by only 0.5% (YNH) and 3.3% (OHL). At the 50% correct point, performances on the Auditec version were 3.2 dB (YNH) and 6.1 dB (OHL) better than those on the VA version. The slopes at the 50% points on the mean functions for both speakers were about 4.9%/dB (YNH) and 3.0%/dB (OHL); however, the slopes evaluated from the individual listener data were steeper, 5.2 to 5.3%/dB (YNH) and 3.3 to 3.5%/dB (OHL). When the individual data were transformed from dB SL to dB HL, the differences between the two listener groups were emphasized. The four functions (2 speakers by 2 listener groups) were plotted for each of the 48 participants and each of the 200 words, which revealed the gamut of relations among the datasets. Examination of the data for each speaker across test sessions, in the traditional 50-word lists, and in the typically used 25-word lists of Randomization A revealed no differences of clinical concern. Finally, introspective reports from the listeners revealed that 91.7% and 83.3% of the YNH and OHL listeners, respectively, thought the Auditec speaker was easier to understand than the VA speaker. Recognition performances on each participant and on each word are presented.

*Arizona State University, Tempe, AZ

Corresponding author: Richard H. Wilson, Arizona State University, Tempe, AZ 85281; Email: wilsonr1943@gmail.com

This work was supported by the Rehabilitation Research and Development Service, Department of Veterans Affairs through the Auditory and Vestibular Dysfunction Research Enhancement Award Program (REAP) at Mountain Home. Additional support was provided by the Arizona State University Foundation.

The data were collected by Cayce N. Griffin, Madison P. Thode, and Ashley E. Light as part of their capstone projects at East Tennessee State University. The contents of this article do not represent the views of the Department of Veterans Affairs or the U.S. government.

Key Words: auditory perception, normal hearing for pure tones, sensorineural hearing loss, speech perception, word recognition

Abbreviations: CD = compact disc; CNC (CVC) = consonant, vowel nucleus, consonant; M = mean; m = slope; NU-6 = Northwestern University Auditory Test No. 6; OHL = older adults with sensorineural hearing loss; PTA = pure-tone average; rms = root-mean-square; SD = standard deviation; VA = Department of Veterans Affairs; YNH = young adults with normal hearing for pure tones

INTRODUCTION

The purpose of this project was to compare word-recognition performances by young adults with normal hearing for pure tones (YNH) and by older adults with sensorineural hearing loss (OHL) on the Auditec of St. Louis and the Department of Veterans Affairs (VA) recorded versions of the Northwestern University Auditory Test No. 6 (NU-6). The NU-6 (Tillman and Carhart, 1966), which recently reached its 50th year, was an expansion of Northwestern University Auditory Test No. 4 (NU-4; Tillman et al, 1963) from two lists to four lists of 50 words/list. Of the 200 NU-6 words, 185 were from the revised pool of 500 consonant, nucleus, consonant (CNC) words devised by Peterson and Lehiste (1962) from their original CNC lists (Lehiste and Peterson, 1959a), with the remaining 15 words drawn from the commonly occurring 1,263 CNC words reported by Thorndike and Lorge (1944). List 1 of NU-4 and NU-6 contained the identical words; List 2 of NU-6 deleted *faith*, *pun*, *sick*, and *towel* from NU-4 and added *fail*, *pick*, *south*, and *ton*; Lists 3 and 4 of NU-6 were composed of new words. NU-6 followed the principles set forth by previous word-recognition tests, including 50 words/list (Egan, 1948, 1957) and recorded materials (Hirsh et al, 1952).

Over the years, recorded versions of NU-6 have been studied in several investigations with YNH and OHL listeners, typically using mean data as the metric. Increasingly, during the 1960s, data generated from studies of word-recognition performance indicated that all other independent variables such as presentation level being somewhat equal, probably the most influential independent variable on word-recognition performance was the particular utterance by a particular speaker of the material under study, that is, the *speaker variable* (Kreul et al, 1968). Kreul et al (1969: p. 287) recognized the importance of the speaker variable when they stated, *Tests ought not be thought of as the written lists of words but as recordings of these words*. Recognition performance or intelligibility of a word can vary little or substantially when spoken by different speakers. Even the same word uttered by the same speaker can elicit a different recognition performance, albeit generally small (Brandy, 1966). When similar words from different lists are spoken by the same speaker but recorded during different sessions, small differences in recognition performances (2–4 dB) have been observed (Wilson

and Oyler, 1997). When similar words from different lists are spoken by the same speaker during the same recording session, differences in recognition performances among the common lists of different materials, for example, NU-6, CID W22s, and PB50s, all but disappear (Wilson et al, 2008).

Tom Tillman was the speaker on the initial recording of NU-6, but the use of that version waned, probably because of availability issues, in favor of the Auditec version (a male speaker) recorded in the 1970s and the VA version (a female speaker) recorded in the 1980s (Causey et al, 1983). Although male and female speakers have been the focus of study and speculation regarding word recognition, speaker characteristics should be thought of on a continuum based on many aspects of the uttered waveform including the fundamental frequency of the speaker and many information cues that are yet to be distilled.

Previous word-recognition studies have examined the performances by YNH and OHL listeners on NU-6 produced by different speakers with an occasional comparison to other word-recognition materials. Again, with the present project, the focus was on the NU-6 studies involving the Auditec speaker and the VA speaker. For both versions of NU-6, there was a substantial data based on recognition performances by YNH listeners but minimal data on performances by OHL listeners, who are ultimately the intended audience for the materials (Bilger, 1984). It is generally easy to compare word-recognition performances across studies when YNH listeners are involved, mainly because of their homogeneous audibility and their homogeneity of recognition performances, but it is difficult to compare performances across studies when OHL listeners are involved, mainly because of their heterogeneity of audibility and recognition performances.

Auditec Speaker with YNH and OHL Listeners

In the Wilson et al (1976) study, recognition performances by YNH and OHL listeners on the Auditec version of NU-6 can be compared. In their Experiment III, recognition performances on the NU-6 materials were determined on YNH listeners ($n = 16$; $M = 26$ years; speech-recognition threshold [SRT] = 0.6-dB HL) at eight sensation levels (SL, re: the SRT) from -3 to 32 dB in 5-dB steps. In Experiment IV, recognition performances were obtained with the same materials on

OHL listeners ($n = 12$; $M = 59$ years; $SRT = 32.8$ -dB HL) using presentation levels of 0- to 42-dB SL in 6-dB increments. The results are summarized in Figure 1 for the two groups of listeners, with recognition performance plotted as a function of presentation level (dB SL). The recognition performances at the 50% correct points calculated from the third-degree polynomial equation used to describe the data were 11.0-dB SL (YNH) and 13.1-dB SL (OHL), with slopes at the 50% points of 4.2%/dB and 2.8%/dB, respectively, which were calculated using the first derivative of the polynomial equation. As the presentation level increased above the 50% points, the slope of the OHL function became more gradual than the slope of the YNH function, with slopes at the 80% point of 1.6%/dB and 3.3%/dB, respectively. At the 80% correct points, the difference between the mean performances increased to 7.8 dB (YNH = 18.7-dB SL; OHL = 26.5-dB SL). Although not detailed here, equivalent recognition performances by the same groups of listeners required presentation levels that were 5 dB lower on the Tillman version of NU-6 than on the Auditec version; that is, the Tillman version was easier than the Auditec version. The data presented in Figure 1 from the Auditec speaker are recast in the upper panel of Figure 2 (open circles = YNH; filled circles = OHL) with the independent variable now being dB hearing level (HL). (Note: Specific percent points on the functions and the slopes of the functions at those points were calculated from the polynomials used to describe the data in Figure 2 and are listed in Tables 1 and 2, which include data from several studies). When plotted in HL, the recognition performance

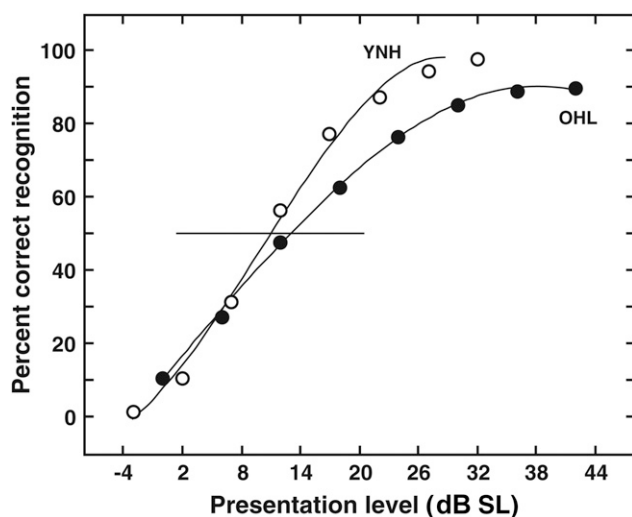


Figure 1. The word-recognition performances of 16 YNH listeners and 12 OHL listeners on the NU-6 lists spoken by the Auditec speaker in Experiments III and IV, respectively, of the Wilson et al (1976) study. Third-degree polynomials are used to describe the data. In this figure, the presentation level is in SL, re: the SRT. For comparison with data from other studies, the same data are expressed in terms of HL in Figure 2.

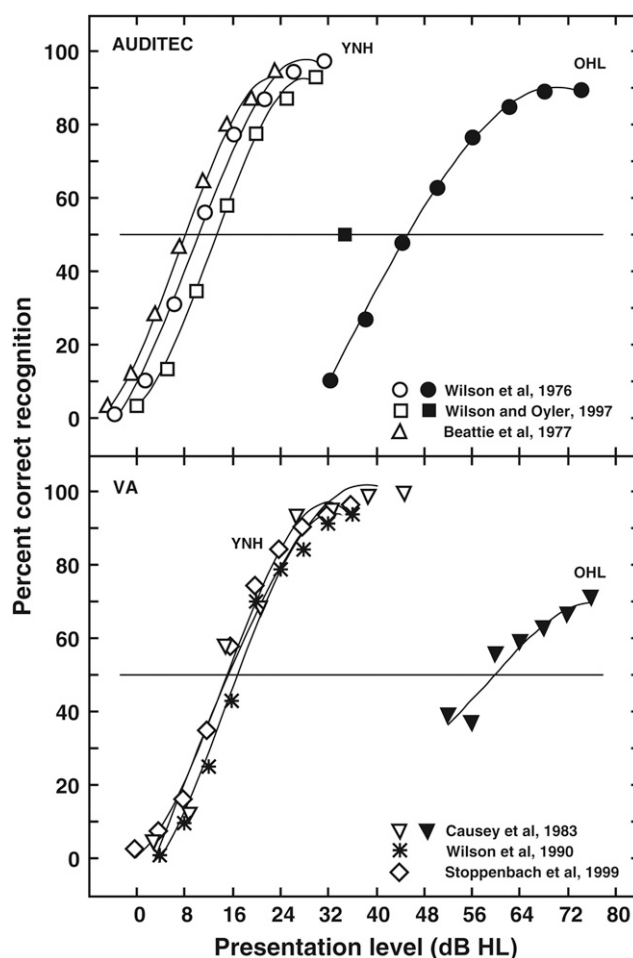


Figure 2. The psychometric functions from several studies of the NU-6 materials expressed in dB HL using the Auditec speaker (upper panel) and the VA speaker (lower panel) for YNH (open symbols) and OHL (filled symbols) listeners. Third-degree polynomials are used to describe the data. Specific points on the functions and the corresponding slopes at those points were calculated and are listed in Tables 1 and 2.

difference at the 50% point between the YNH and OHL groups increased to about 32.7 dB, which closely approximates the 30.5-dB SRT sensitivity difference between the two listener groups.

In a second investigation of the Auditec version of NU-6, 24 YNH listeners ($M = 23.1$ years) were studied by Beattie et al (1977) again using SL as the independent variable referenced to the SRTs, which were established with a monitored live-voice technique. Based on a third-degree polynomial fit to their data expressed in SL, the 50% point was at 16.8-dB SL, which was 5.8 dB higher than the 11.0-dB SL 50% point reported by Wilson et al (1976) for the same materials. This 5.8-dB discrepancy between the two datasets probably is attributable to the artificially low (11.1-dB SPL) SRT that was established. The Beattie et al data from their Table 1 (p. 63) were transformed from SL to HL using a mean SRT of 11.1-dB SPL and a 20-dB SPL reference

Table 1. The HLs (dB) at which the 20%, 50%, and 80% Recognition Performances Occurred in the Various Studies Depicted in Figure 1 for the Auditec Speaker

| Study | dB HL | | | | Slope (%/dB) | | | |
|-------------------------|-------|------|------|------|--------------|-----|-----|---------------------------|
| | n | 20% | 50% | 80% | 20% | 50% | 80% | linear _{20%–80%} |
| YNH | | | | | | | | |
| Wilson et al (1976) | 16 | 3.0 | 10.3 | 18.0 | 3.7 | 4.2 | 3.3 | 4.0 |
| Beattie et al (1977) | 24 | 1.2 | 8.2 | 15.6 | 3.8 | 4.5 | 3.2 | 4.2 |
| Wilson and Oyler (1997) | 24 | 3.0 | 13.4 | 20.9 | 3.7 | 4.5 | 3.2 | 3.6 |
| Mean | 3.5 | 10.6 | 18.2 | 3.7 | 4.4 | 3.2 | | |
| OHL | | | | | | | | |
| Wilson et al (1976) | 12 | 35.7 | 46.0 | 62.1 | 3.1 | 2.6 | 1.0 | |
| Wilson and Oyler (1997) | 24 | — | 34.8 | — | — | 2.3 | — | |

Note: Data are included for both YNH and for OHL. The slopes of the functions at those percent correct points also are listed along with the traditional linear slope between 20% and 80% correct (linear_{20%–80%}). See the Supplemental Materials for details regarding the conversions of the various presentation levels of the studies to HL.

threshold for speech. The data in this format are depicted in the upper panel of Figure 2 (triangles) and in Table 1. When transformed, the Beattie et al data are only about 2 dB easier than the data from Wilson et al, with slopes for both functions at the 50% points essentially the same, 4.2%/dB and 4.5%/dB.

As part of another study on the Auditec version of NU-6 (Wilson and Oyler, 1997), recognition performance functions were measured on YNH listeners ($n = 24$; $M = 26.3$ years) and two-point functions separated by 10 dB were measured around the 50% point on OHL listeners ($n = 24$; $M = 58.0$ years; pure-tone average [PTA] = 24.7-dB HL). The results are plotted in the upper panel of Figure 2 (squares). For the YNH listeners, the 50% recognition point was at 13.4-dB HL, with a slope of 4.5%/dB. By contrast, the 50% point for the OHL listeners was at 34.8-dB HL, with a slope of 2.3%/dB. The ~21 dB difference between the 50% points for the two groups is probably an accurate reflection of the sensitivity differences between the YNH and OHL groups.

For practical purposes, the results of these three studies are in reasonable agreement considering the differences involved in their experimental protocols.

The data from these studies do, however, indicate that presentation levels more than a 30- to 40-dB SL range (re: the SRT) are sufficient to include most of the recognition performances expressed in percent correct. Whereas data on the Auditec version of NU-6 from YNH listeners are plentiful, corresponding data from OHL listeners are limited.

VA Speaker with YNH and OHL Listeners

Causey et al (1983) evaluated recognition performances on the VA recording of NU-6, which they referred to as the Maryland NU-6 (henceforth in this article referred to as the VA version of NU-6). Both YNH ($n = 40$; $M = 24$ years) and OHL ($n = 40$; $M = 53$ years; PTA = ~32-dB HL) listeners were studied using eight SLs, re: the SRT. The younger listeners were evaluated at 0- to 40-dB SL in 6-dB steps, with half the subjects receiving the four lists at 0-, 12-, 24-, and 36-dB SL and the other half of the subjects receiving the four lists at 6-, 18-, 30-, and 42-dB SL; the SRTs were not given. The same alternating design was used with the older listeners, with the range of SLs changing from 20 to 48 dB in 4-dB steps. Because the Causey et al

Table 2. The HLs (dB) at which the 20%, 50%, and 80% Recognition Performances Occurred in the Various Studies Depicted in Figure 1 for the VA Speaker

| Study | dB HL | | | | Slope (%/dB) | | | |
|--------------------------|-------|------|------|------|--------------|-----|-----|---------------------------|
| | n | 20% | 50% | 80% | 20% | 50% | 80% | linear _{20%–80%} |
| YNH | | | | | | | | |
| Causey et al (1983) | 40 | 7.9 | 15.3 | 24.0 | 4.2 | 3.8 | 2.9 | 3.7 |
| Wilson et al (1990) | 24 | 10.1 | 17.1 | 24.9 | 4.0 | 4.3 | 3.1 | 4.1 |
| Stoppenbach et al (1999) | 24 | 7.8 | 15.2 | 22.6 | 3.6 | 4.3 | 3.5 | 4.1 |
| Mean | 8.6 | 15.9 | 23.8 | 3.9 | 4.1 | 3.2 | 4.0 | |
| OHL | | | | | | | | |
| Causey et al (1983) | 40 | — | 59.6 | — | — | 1.3 | — | |

Note: Data are included for both YNH and for OHL. The slopes of the functions at those percent correct points also are listed along with the traditional linear_{20%–80%}. See the Supplemental Materials for details regarding the conversions of the various presentation levels of the studies to HL.

article provided no data on the SRTs for either group of listeners, for comparison purposes, the data were converted from SL to HL using the following logic. For the YNH group, a mean SRT of 3-dB HL was assumed to be reasonable and was used to determine the HLs. For the OHL group, the 500, 1000, and 2000 Hz PTA (from Figure 2, p. 65) was estimated to be 32.2-dB HL and was used as the estimated reference for SL. To evaluate the Causey et al findings, which are illustrated in the lower panel in Figure 2 (inverted triangles), the data from their Figures 1 and 3 (p. 64) were converted from SL to HL and fit with third-degree polynomials from which the 20%, 50%, and 80% points and slopes at those points were calculated (see Table 2). For the YNH listeners (open symbols), the 50% point was 15.3-dB HL with a slope of 3.8%/dB. The data for the OHL group (filled symbols), which obviously failed to encompass the complete range of recognition performances, yielded a 50% point at 59.6-dB HL, with a slope of 1.7%/dB. Thus, the 50% points in HL for the YNH and OHL listeners differed by about 44 dB with the VA speaker, which was somewhat different from the 36-dB difference observed between YHN and OHL listeners with the Auditec speaker (Wilson et al, 1976). The reasons for this discrepancy are unknown but certainly were probably attributable to procedural/speaker effects and audibility differences between the listener groups.

In a 1990 study, Wilson et al generated psychometric functions for the VA version of NU-6 under three listening conditions, quiet and two noise conditions. The focus here is on the quiet condition. The 24 YNH ($M = 23.2$ years) with pure-tone thresholds ≤ 15 -dB HL (ANSI, 1989) were given two of the NU-6 lists at each of ten presentation levels from 20- to 56-dB HL in 4-dB increments. The data were fit with third-degree polynomials, which are shown in the lower panel of Figure 2 (star-

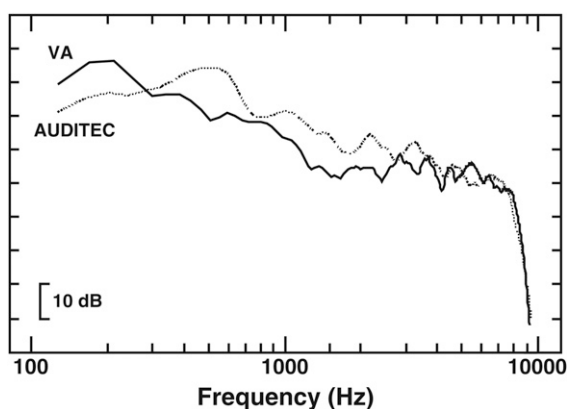


Figure 3. The long-term spectra of the 200 NU-6 words spoken by the Auditec and VA speakers and the difference between the two spectra. The carrier phrases were not included. Both signals were low-pass filtered and then set to the same overall rms before the frequency analyses were conducted.

bursts), from which the 20%, 50%, and 80% points and slopes at those points were calculated (see Table 2). The YNH data in both the graphic and numeric forms are in good agreement with the earlier YNH data from Causey et al (1983), with differences of about 2 dB.

Finally, Stoppenbach et al (1999) examined two versions of the NU-6 materials recorded by the VA speaker. Both versions were digitized at 20,000 samples/s from the same analog master audio tape, but with a few differences. The original digital version (D1.0) used a 12-bit digitizer and a 5000-Hz low-pass filter cutoff (115 dB/octave). The second digital version (D1.1) used a 16-bit digitizer and an 8800-Hz low-pass filter cutoff (96 dB/octave), which gave D1.1 better fidelity than D1.0. The D1.1 materials, which were used in all studies after 1989, were recorded 0.5 dB higher than the D1.0 materials. The comparison study by Stoppenbach et al involved 24 YNH listeners ($M = 23$ years). In the quiet condition, each listener was presented 1 of the 4 NU-6 lists at each of 12 presentation levels in 4-dB steps, ascending from 16- to 60-dB HL. Each list was presented three times to each listener and overall six times at each presentation level. The D1.1 data are plotted in the lower panel of Figure 2 (diamonds) with the 20%, 50%, and 80% points, and slopes at those points are listed in Table 2. Although not presented here, recognition performance on the D1.1 condition was 2.2 to 2.9 dB better than on the D1.0 condition, which may reflect the 0.5-dB level difference between the two recordings and the higher resolution obtained with the 16-bit digitizer and wider bandwidth. The results with the D1.1 version of NU-6 using YNH listeners demonstrate little difference from the results with the VA speaker reported by Causey et al (1983) and by Wilson et al (1990).

In summary, comparing the NU-6 functions produced by the Auditec and VA speakers and by two investigations (Wilson et al, 1976; Causey et al, 1983), two conclusions are apparent from the data in Tables 1 and 2 and Figures 1 and 2. First, for YNH listeners, the Auditec version of NU-6 appears to be about 5 dB easier at the 50% point than the VA version; that is, equivalent recognition performances would be achieved if the VA version were presented at a presentation level 5 dB higher than the presentation level of the Auditec version. Second, the difference between recognition performances by YNH and OHL listeners at the 50% point is about 10 dB greater on the VA version (see Table 2 from Causey et al; 59.6-dB HL – 15.3-dB HL = 44.3 dB) than on the Auditec version (see Table 1 from Wilson et al; 46.0-dB HL – 10.3-dB HL = 35.7 dB), but these differences, which are based on limited data, are probably attributable to the substantially greater hearing loss exhibited by the listeners in the Causey et al study than by the listeners in the Wilson et al study. Missing is a complete dataset in which the NU-6 words spoken by the Auditec speaker and by the VA speaker are

evaluated on the same YNH and OHL listeners, which is the focus of the present study. The design was such that with the same listeners, comparisons between the two speaker versions were made at multiple presentation levels for each of the individual 200 words that comprise the NU-6 lists.

Based on the NU-6 data reviewed using the Auditec and VA speakers, several hypotheses were proposed. First, YNH listeners would have better overall word-recognition performance at equal presentation levels on the Auditec version than on the VA version. Second, the overall performance difference between the two versions of NU-6 would be maintained, or perhaps exaggerated, on OHL listeners. Third, the performance differences between the two versions of NU-6 would be consistent on most of the individual words with the Auditec version being easier than the VA version but a minority of words would demonstrate a substantially reduced performance difference or even a reversal of the performance difference with the VA version being easier than the Auditec version. Because data were available on the individual words, the performances across the three test sessions, and the traditional 50-word and 25-word lists of NU-6, Randomization A was analyzed and reported. In addition, the words were ranked by overall recognition performances (i.e., ease/difficulty) for both speaker versions and listener groups. Clinically, the database of recognition performances established with the two versions of NU-6 provides a reference for evaluating word-recognition performances on both NU-6 versions by patients, typically at different clinics and/or at different points in time.

METHODS

Materials

The test materials consisted of the 200 NU-6 monosyllabic words and their associated carrier phrases spoken by the Auditec speaker and by the VA speaker. The proverbial elephant in the room with speech studies like the present one is the presentation level of the speech materials and the calibration of that level. A few points are helpful in appreciating this calibration issue. First, overall the acoustic speech signal is aperiodic with a substantial amplitude modulation component, which is totally unlike a pure-tone signal that acoustically is consistent and predictable. Second, speech signals (i.e., words) presented at overall equal root-mean-square (rms) levels do not necessarily result in equal word-recognition performances (Wilson, 2015, Figures 9 and 10). This tenuous relation between presentation level and intelligibility was recognized early on by Davis (1947) when he stated (p. 138), *More disturbing is the realization that all words, spoken naturally and in*

sequence, do not have the same physical power. Davis went on to say that, all words are not equally intelligible [understandable] and that Some [words] can be understood even when barely audible, whereas others [words] must be at a much higher level before even a practiced listener can identify them correctly (pp. 138–139). Finally, we do not know with any certainty what the various cues in a word (such as amplitude, frequency, and transitions) contribute singularly or in combination with the intelligibility of a unique word utterance, which may be complicated further by different listeners using different cues to understand the same utterance, especially listeners with sensorineural hearing loss. Although future methods/techniques may be developed to quantify precisely the amplitude component of various speech signals, the amplitude calibration of the materials used in this study follows the ANSI Standard (2010) for the carrier phrase and target word paradigm that is used universally in most audiology clinics that use recorded materials. Originally, the NU-6 materials spoken by the Auditec and VA speakers were produced in accordance with the ANSI standard for audiometers. That standard specified that the 1000-Hz calibration tone be equal to the *average peak deflection of the preliminary carrier phrase* on a vu meter that in turn may be considered representative of the *speech material immediately following [the carrier phrase] when the material is delivered in a natural manner at the same communication level as the carrier phrase* (ANSI, 2010: p. 18). As Wilson (2015) noted, this calibration protocol is a carryover from the early days of radio for which the vu meter was developed to standardize the target levels of speech materials for transmission line and radio broadcast purposes (Chinn et al, 1940).

Physically, both the long-term spectra of the Auditec and VA speakers and the waveforms of the same words spoken by the two speakers are in some ways similar, but at the same time, noticeably different. It is difficult, however, to know how the similarities and differences translate into any perceptual similarities and differences between the two versions of the same materials. To obtain the spectra, the words were edited from the carrier phrases and concatenated into a 200-word file with the Auditec and VA versions on the two channels that were low-pass filtered at 8800 Hz and then equated in rms. The spectra for the words spoken by the Auditec and VA speakers then were obtained (Adobe Systems, Inc., 2012) and are shown in Figure 3. The two spectra can be characterized as having the same general shape with frequency-specific differences. The waveforms (amplitude by time plots) for *Say the word book* spoken by the two speakers are illustrated in Figure 4 with a second example, *Say the word size*, shown in Supplemental Figure S1 (available with the online version of this article). These two waveform

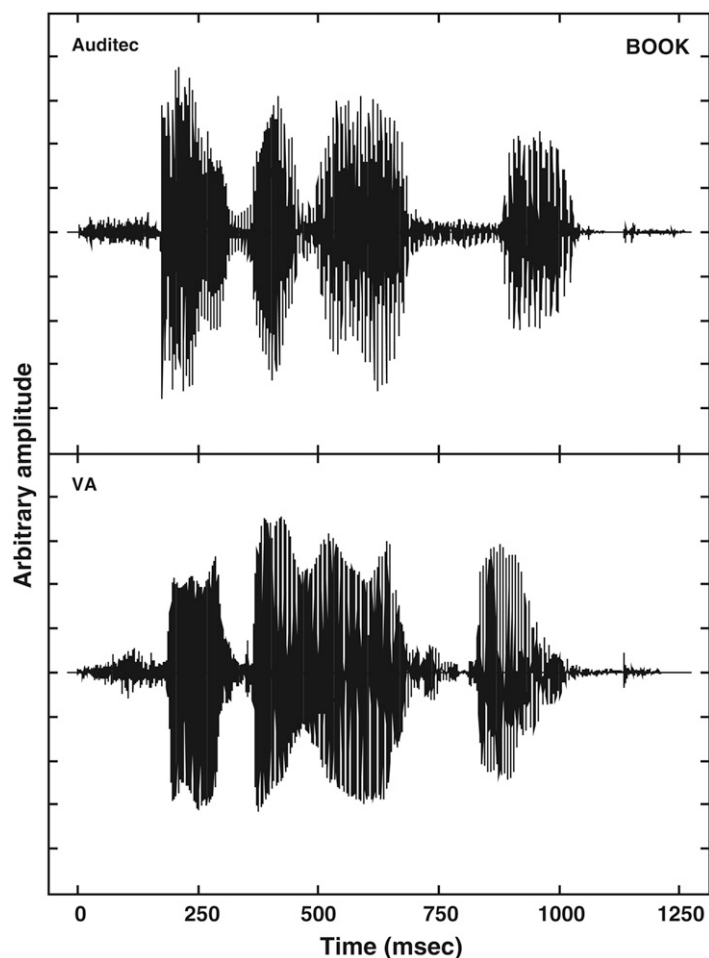


Figure 4. The waveforms of Say the word *book* spoken by the Auditec speaker (upper panel) and by the VA speaker (lower panel).

examples provide a general overview of the same words spoken by the two speakers, especially the upper and lower *temporal envelopes* of the signals. The temporal envelopes can be thought of as curving lines connecting the amplitude extremes (positive or upper; negative or lower) of the waveform. As an aside, close observation of the waveforms in Figure 4 and Supplemental Figure S1 reveals amplitude asymmetries, which often are observed with many sounds (Robjohns, 2013). Within the temporal envelope, the speech waveform is characterized by the *temporal fine structure*, which is randomly and/or systematically the signal modulations that occur from moment-to-moment throughout the course of the utterance. An example of an expanded temporal waveform is provided in Figure 5 in which both the temporal envelope and temporal fine structure of the word *neat* spoken by the Auditec and VA speakers can be observed. The differences between the waveforms in Figure 5 are striking. Viewed with respect to the temporal envelopes and temporal fine structures, the utterances of the two speakers depicted in the Figures 4 and 5 are similar in many ways and dissimilar in other ways. Ultimately, however, two utterances of the same script

from different speakers provide the same information to the listener with slightly varying degrees of efficiency that depends on the status of the auditory system (e.g., Grose et al, 2015). This component of auditory processing is in the growing literature on *speaker normalization* or *perceptual normalization of speech* (e.g., Bladon et al, 1984; Pisoni, 1997; Johnson, 2005), which like other auditory functions probably declines with age and increased hearing impairment.

The carrier phrases and words, which were ripped from available compact disc (CDs) (Auditec of St. Louis, courtesy of Bill Carver; Department of Veterans Affairs, 2010), were compiled into 400 files (two speakers by 200 words) with the signal on Channel A (left) designated the stimulus signal and the signal on Channel B (right) used as the monitor channel. The 1000-Hz calibration tones of the two versions differed by 0.67 dB, with the rms level of the VA recording higher than the rms level of the Auditec recording. This difference was corrected by adding 0.67 dB to both channels of the Auditec materials, which produced an overall common referent for the two sets of materials. Ideally, the goal with each subject was to capture as much of the

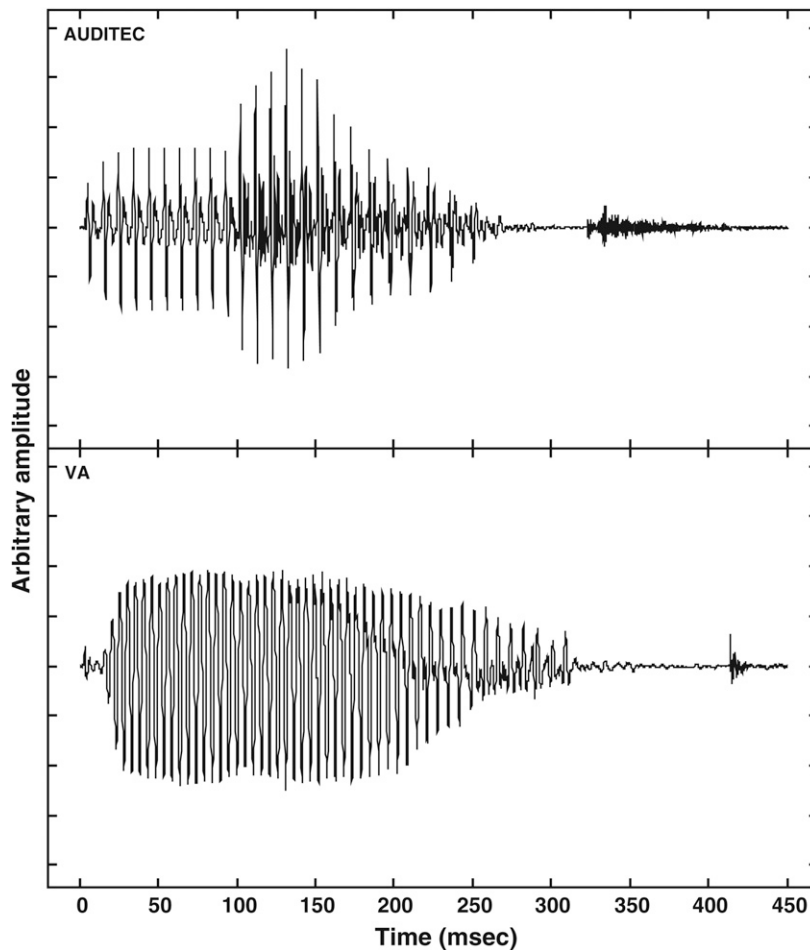


Figure 5. The waveform for the word *neat* is presented for the Auditec speaker (upper panel) and for the VA speaker (lower panel).

recognition psychometric function (0–100% correct) as possible. Because of the audibility variability associated with listeners with sensorineural hearing loss, it was necessary to use SL as the independent variable. The variability associated with previous word-recognition data (e.g., Wilson et al, 1976; Causey et al, 1983; Wilson et al, 1990) suggested that minimally a 30-dB range was required to define the psychometric functions for each subject and each word. Having captured most of the psychometric function from 0% to 100% correct, pilot data confirmed the 30-dB range for the YNH listeners but indicated a 40-dB range was required for the OHL listeners. The two ranges, both of which increased from –2-dB SL, were defined with six presentation levels that were in 6-dB steps for the YNH listeners and 8-dB steps for the OHL listeners. Thus, for each subject group, 2,400 stimulus words (2 speakers, 200 words, and 6 presentation levels) emerged. To generate the stimuli, first, the 400 carrier phrases and their companion target words were edited from the NU-6 lists and put into files that were assigned a unique five-digit file name that coded the speaker (1–2), the word list that contained the word (1–4), the word

number in the Randomization A list (01–50), and the level of the materials (0–5). Second, a batch-processing routine (Adobe Audition CS5) was used to apply the required attenuation to the various groups of 200 words on the left channel, leaving the monitoring materials on the right channel unaltered.

An issue in almost all studies involving performance measures on a task by the participants is how much “work” can be expected from a subject during a test session while maintaining a fairly consistent level of performance. Here, *work* is used to indicate the task of the subject, which in this study involved (a) listening to words being uttered at random presentation levels that made some words easy to recognize and some words difficult to recognize and (b) verbally recalling the words. The task required mental focus but in many ways was monotonous. To minimize the monotony and to avoid the fatigue effects often associated with long test sessions, the decision was made to limit each test session to 425 words, with the first 25 words used for practice to acquaint the listener with the experimental paradigm of random words presented at random presentation levels.

With 2,400 stimulus words for each group of listeners, the 400 test-word limit/session would have required 6 sessions that were deemed an unrealistic requirement, especially for the older participants. The necessity of 6 sessions prompted the decision to split the list of 200 words in half with each of two participants in a subject pair randomly assigned 100 words, thereby reducing the number of stimulus words to 1,200/listener (2 speakers, 100 words, and 6 presentation levels) and reducing the number of test sessions to three. Following this strategy, the list of 200 NU-6 words was randomized for each subject pair, with the first 100 words allocated to the odd numbered subject and the last 100 words allocated to the even numbered subject. To make each of the 100 words at each of the 6 presentation levels as independent as possible, the respective 600 stimulus words by each speaker for each subject were randomized and for convenience were assigned to blocks of 25 words. In an earlier study, Sommers et al (1994) demonstrated that randomly varying the presentation level from trial to trial had no overall effect on recognition performance. The mean distribution of the six presentation levels across the three test sessions was 32–34%/session. For the odd-numbered subject of the pair, a 25-word list by the VA speaker was presented first and subsequently the 25-word lists by the VA speaker were alternated with the 25-word lists by the Auditec speaker. Conversely, for the even-numbered subject, a 25-word lists by the Auditec speaker was presented first and subsequently alternated with the 25-word lists by the VA speaker. Thus, throughout the 1,200-word test protocol for each listener, the speaker changed every 25 words. Mullennix et al (1989) observed that changing speakers from trial to trial decreased perceptual performance, but this

effect was not thought to have a substantial influence on the current design in which there were only two speakers that alternated 25-word sets. An example of Channel A of one 25-word track for a YNH listener is depicted in Figure 6, in which the numbers just above the lower abscissa indicate the attenuation (in dB) of each stimulus. With this particular track, five words were at the maximum presentation level (0-dB attenuation), five words attenuated 6 dB, three words attenuated 12 dB, six words attenuated 18 dB, five words attenuated 24 dB, and one word attenuated 30 dB. In addition to the 100 test words for each participant, three practice lists of 25 words each were compiled for each participant. The 75 practice words were taken randomly from the 100-word corpus given to the companion listener of the subject pair. The practice list and 16 experimental lists for each of the three test sessions were recorded on audio CD for each participant. A 3.5-s interstimulus interval was used with all recordings and each 25-word list was 2 min.

Finally, to assess the introspective reports from the listeners about the ease/difficulty of understanding the two speakers, a three-question survey was developed. First, the listener was asked if one speaker was easier to understand than the other; if one speaker was identified as easier, then the listener was asked to identify the easier speaker. Second, the listener was asked to rate the easier speaker on a Likert scale from one (easy to understand) to ten (difficult to understand). Third, the same rating was then used with the more difficult speaker to understand. If on the first question the listener thought the speakers were equally easy/difficult to understand, then only one Likert scale was used with each speaker receiving the same rating.

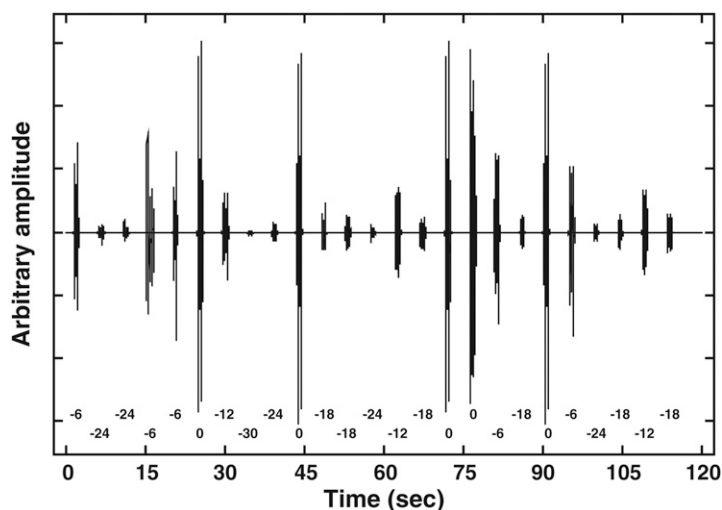


Figure 6. An example waveform of a randomly selected experimental track (left channel only) that includes 25 carrier phrases and their associated words at the various presentation levels (dB), re: the PTA. The numbers just above the lower abscissa represent the attenuation values (dB) used with the word files to achieve the presentation levels.

Participants

The 12 YNH listeners, nine of whom were females, were recruited from the local university community and ranged in age from 18 to 29 years ($M = 24.0$ years; standard deviation [SD] = 2.0 years). The younger adults had pure-tone thresholds at the octave frequencies <20 -dB HL (ANSI, 2010) with a 3-frequency, PTA (500, 1000, and 2000 Hz) of 9.9-dB HL (SD = 3.1 dB). The 36 OHL listeners with sensorineural hearing loss met the following inclusion criteria for the test ear: (a) 60–85 years of age, (b) English was their first language, (c) 500-Hz thresholds ≤ 30 -dB HL, (d) 1000-Hz thresholds ≤ 40 -dB HL, (e) a 3-frequency, PTA < 40 -dB HL, and (f) clinical word-recognition performance $\geq 60\%$ correct. These criteria were used with the OHL listeners because they were considered typical of the hearing loss/impairment associated with older veterans (Wilson and McArdle, 2013). The OHL subjects ($M = 71.6$ years; SD = 5.0 years) were recruited from the list of patients evaluated in the Audiology Clinic at Mountain Home who had consented to serve as research subjects in auditory/vestibular experiments. The mean PTA was 26.7-dB HL (SD = 7.1 dB). The mean pure-tone thresholds and SDs for the two groups of listeners are shown in Figure 7.

Procedures

Three 1-hour test sessions were conducted over a 1–100 day interval (YNH, $M = 6.5$ days, SD = 6.0 days; OHL, $M = 8.2$ days, SD = 12.8 days). In Session 1, the institutional review board consent forms were completed and pure-tone thresholds were established

with the Automated Method for Testing Auditory Sensitivity™ (Margolis et al, 2007, 2010) procedure using a tablet (Dell, Venue 10, Round Rock, TX) and Sennheiser HD280 Pro earphones (Hanover, Germany). The average pure-tone threshold of 500, 1000, and 2000 Hz in the test ear during Session 1, rounded to the nearest decibel HL, served as the reference for the presentation level of the test materials in all sessions. The test protocol then was explained, the protocol instructions were given, and questions answered. A mandatory break was provided following presentation of the eighth word list with other breaks provided as requested by the participant, which seldom occurred. In Sessions 2 and 3, the protocols were identical with an automated method for testing auditory sensitivity recheck of the 500-, 1000-, and 2000-Hz thresholds to monitor for potential changes in hearing sensitivity, followed by a review of the test protocol, instructions, a practice list, and the 16 experimental lists. At the end of Session 3, the three-question survey regarding the ease/difficulty of the listening task was administered.

The speech materials were reproduced by an audio CD player (Sony, Model CDP-CD375; Minato, Tokyo, Japan), fed through an audiometer (Grason-Stadler, Model 61; Eden Prairie, MN), and delivered to the test ear via a TDH-50P earphone (Farmingdale, NY) encased in a Type 51 cushion. Based on pilot data, presentation levels of -2 to 28-dB SL in 6-dB steps were used for the YNH listeners, whereas presentation levels of -2 to 38-dB SL in 8-dB steps were used for the OHL listeners. The nontest ear was covered with a dummy earphone. The testing was conducted in a sound booth and the verbal responses of the listeners were recorded in a spreadsheet. The participants were reimbursed each session for their travel expenses.

RESULTS AND PRELIMINARY DISCUSSION

The basic question of this investigation was do the Auditec and VA recorded versions of NU-6 produce the same or different word-recognition performances on primarily OHL listeners. As a reference, the same question was posed for YNH listeners. The individual subject data, the individual word data, and other relational data of interest are listed in tables and graphically detailed in the Supplemental Materials. The mean results for the two speakers and two subject groups are presented in terms of the mean data referenced to the independent variable (dB SL) supported by representative psychometric functions from the individual subjects and bivariate plots of the individual subject performances at each of the six presentation levels. To obtain an accurate estimate of the slopes of the psychometric functions, the data were evaluated across seven levels of recognition performances (dependent variable, 20–80% correct). The recognition

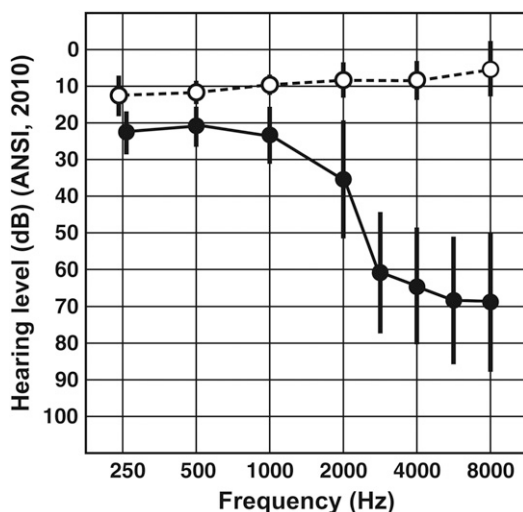


Figure 7. The mean test-ear pure-tone audiograms for the YNH listeners (open circles, $n = 12$) and OHL listeners (filled circles, $n = 36$) listeners involved in the study. The vertical lines represent ± 1 SD.

data then are presented with reference to the independent variable that was transformed from dB SL to dB HL, which provides another perspective on the relations between the performances by the two listener groups. Then, a second bivariate plot is provided for the recognition performances on the 200 words by the 2 speakers at each of the 6 presentation levels (1,200 points for each listener group). Representative word functions for the two speakers and two listener groups also are presented along with the introspective reports of the 48 listeners. Although not directly involved in the questions posed in this project, the data presented the opportunity to examine for the Auditec and VA versions of NU-6 (a) the issue of possible learning effects resulting from the repeated presentation of the same 100 words across sessions, (b) the traditional four, 50-word lists of NU-6, (c) the half-lists of the 50-word lists (Randomization A) that are often used clinically, and (d) the ranking of the words by recognition performance.

Data in dB SL

Although averaging recognition performances across different presentation levels is a somewhat tenuous statistic, such a discretion provides an overall glimpse of the data. Accordingly, the average recognition performances on the Auditec and VA versions of NU-6 were (a) 71.4% and 64.1%, respectively, for the YNH listeners, and (b) 68.7% and 58.2%, respectively, for the OHL listeners. All subjects performed better on the Auditec version of NU-6 than on the VA version. The

mean data for the two speakers and the two listener groups evaluated across the presentation level (dB SL) are listed in Table 3 with the psychometric functions presented in Figure 8 (the individual subject data are plotted in this format in Supplemental Figure S2). In the figure, third-degree polynomials are used to describe the data. In general, the relations between the recognition performances by the two subject groups mirror the relations between similar subject groups with the Auditec version of NU-6 shown earlier in Figure 1. Basically, the OHL group requires a broader range of presentation levels to encompass the range of recognition performances. At the two highest presentation levels, which for the YNH and OHL listeners were offset by 10 dB, mean recognition performances by the two listener groups were for practical considerations the same on the two versions of NU-6. Specifically, from Table 3, maximum recognition performances, which antiquatedly often is termed *PB Max*, were reached (a) with the YNH listeners at 28-dB SL (98.1% and 97.6%) and at 22-dB SL (95.8% and 94.6%), which was only 2.3% and 3.0% lower, and (b) with the OHL listeners at 38-dB SL (95.1% and 91.8%) and at 30-dB SL (91.6% and 87.2%), which was only 3.5% and 4.6% lower. At the three lowest presentations levels, the mean performances by each listener group were better on the Auditec version than on the VA version by 8.0–20.0% (YNH) and by 12.6–18.7% (OHL). The 50% points on the mean functions in Figure 8 calculated from the polynomial equations for the Auditec version were 3.8-dB SL (YNH) and 6.4-dB SL (OHL) and for the VA version were 7.0-dB

Table 3. Five Overall Measures of Recognition Performance (in % correct) at the 6 Presentation Levels (dB SL) from the 12 YNH and 36 OHL on the NU-6 Materials Recorded by the Auditec Speaker and by the VA Speaker

| dB SL | Auditec Speaker | | | | | VA Speaker | | | | |
|---------------|-----------------|------|------|-----|-------|------------|------|------|-----|-------|
| | Mean | SD | Max | Min | Range | Mean | SD | Max | Min | Range |
| YNH Listeners | | | | | | | | | | |
| 28 | 98.1 | 1.8 | 100 | 93 | 7 | 97.6 | 2.2 | 100 | 93 | 7 |
| 22 | 95.8 | 2.8 | 99 | 90 | 9 | 94.6 | 2.4 | 98 | 90 | 8 |
| 16 | 91.5 | 3.1 | 96 | 85 | 11 | 88.0 | 6.0 | 97 | 76 | 21 |
| 10 | 76.8 | 9.4 | 85 | 54 | 31 | 68.8 | 12.9 | 81 | 39 | 42 |
| 4 | 48.9 | 13.0 | 62 | 23 | 39 | 28.9 | 14.5 | 51 | 7 | 44 |
| −2 | 17.2 | 10.0 | 36 | 3 | 33 | 6.5 | 6.3 | 20 | 0 | 20 |
| SK50% | 5.3 | 1.8 | 8.6 | 3.5 | 5.1 | 7.9 | 2.1 | 11.9 | 5.4 | 6.4 |
| OHL Listeners | | | | | | | | | | |
| 38 | 95.1 | 3.7 | 100 | 83 | 17 | 91.8 | 6.4 | 100 | 73 | 27 |
| 30 | 91.6 | 5.3 | 99 | 79 | 20 | 87.2 | 8.4 | 99 | 70 | 29 |
| 22 | 85.1 | 6.1 | 98 | 75 | 23 | 76.7 | 10.7 | 96 | 54 | 42 |
| 14 | 71.3 | 10.9 | 94 | 41 | 53 | 56.0 | 15.1 | 86 | 23 | 63 |
| 6 | 47.1 | 16.4 | 82 | 13 | 69 | 28.4 | 17.4 | 70 | 0 | 70 |
| −2 | 21.7 | 16.7 | 56 | 0 | 56 | 9.1 | 10.9 | 46 | 0 | 46 |
| SK50% | 9.0 | 3.3 | 15.4 | 0.6 | 14.9 | 14.1 | 3.7 | 20.8 | 5.4 | 15.4 |

Note: Also included are data on the 50% points calculated with the Spearman–Kärber equation from each individual set of data, which estimates the SL at which 50% correct performance was achieved. The data from the individual subjects are listed in Supplemental Tables S1 and S2 (available with the online version of this article)

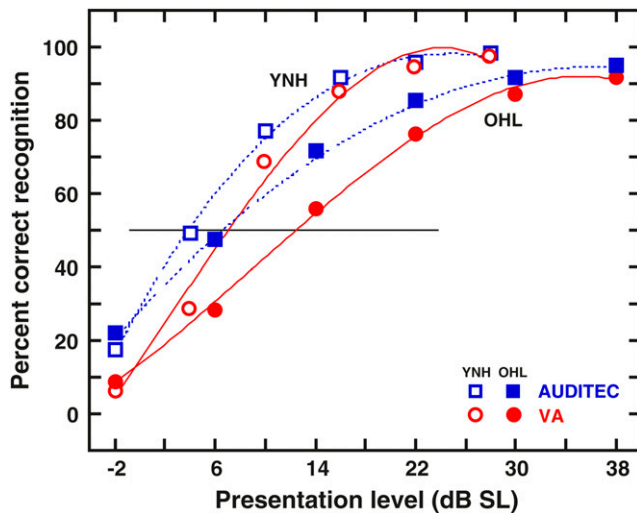


Figure 8. The mean recognition performances as a function of presentation level (dB SL) on the NU-6 words spoken by the Auditec speaker (blue squares and dotted lines) and by the VA speaker (red circles and solid lines) are shown for 12 YNH (open symbols) and for 36 OHL (filled symbols). The data were averaged across subjects and the independent variable. Third-degree polynomials are used to describe the data.

SL (YNH) and 12.5-dB SL (OHL). Thus, the performance differences in terms of decibels between the two versions of NU-6 were 3.2 dB for the YNH listeners increasing to 6.1 dB for the OHL listeners. As will be shown in the subsequent section, essentially all listeners performed better at the three lowest presentation levels on the Auditec version of NU-6 than on the VA version, which is ample support for the differences being considered valid. The 50% points for the two Auditec functions in the present study are about 7 dB lower than the 50% points obtained from the same materials in an earlier study (Wilson et al 1976), the results of which were depicted in Figure 1. A multitude of reasons can account for the discrepancy between the two results, for example, differences in subject criteria, calibration/SL references, and presentation protocol.

Traditionally, the slopes of word-recognition functions have been described as the slope (m) of the linear portion of the function between the 20% and 80% correct points (linear_{20%–80%}; $m = \Delta y / \Delta x$). For the YNH functions, the linear_{20%–80%} slopes were both 4.6%/dB, with the slopes at the 50% points calculated from the first derivative of the polynomials used to describe the data slightly steeper at 4.9%/dB (Auditec) and 4.8%/dB (VA). The linear_{20%–80%} slopes of the Auditec and VA speaker functions for the OHL listeners were 2.8%/dB and 2.9%/dB, respectively, with the slopes calculated at the 50% points essentially the same. Similar slopes with these materials have been reported for YNH and OHL listeners in other studies, for example, Wilson et al (1976) reported slopes of 4.2%/dB and 2.8%/dB, respectively, for the two types of listeners. The slopes

just reported are the slopes of the respective mean functions; the slopes of the recognition-performance data are considered further in a subsequent section.

Individual Subject Data

The relational patterns between the mean recognition performances on the two versions of NU-6 were exhibited by all subjects in each group of listeners, samplings of which from three YNH listeners and nine OHL listeners are depicted in Figure 9 (the data for each of the 48 listeners are presented in Supplemental Tables S1 and S2, and Supplemental Figures S3–S6 [available with the online version of this article]). The individual subjects in both listener groups demonstrated a variety of differences between recognition performances on the two versions of NU-6. With the YNH listeners, the differences in recognition performances at the 50% points on the functions ranged from 1.1 dB (YNH 4) to 4.8 dB (YNH 1) and with the OHL listeners the differences ranged from 1.2 dB (OHL 25) to 11.4 dB (OHL 15); in all cases, the differences indicated better performances on the Auditec version of NU-6. The mean differences between versions of NU-6 at the 50% points on the functions calculated from the polynomial equations for each subject were 3.0 dB (SD = 1.1 dB) for the YNH listeners and 5.8 dB (SD = 2.4 dB) for the OHL listeners, which are the same differences observed with the mean data in Figure 8 but with the addition of intersubject variability data.

The striking aspect of the individual psychometric functions is the intersubject variability both in terms of absolute values and differential values. As discussed in the previous section, the one near equality of recognition performances on the two versions of NU-6 is found at the highest presentation levels, which can be observed in Supplemental Tables S1 and S2 and Supplemental Figures S3–S6. For the YNH listeners at 28-dB SL, the mean recognition performance difference between the two versions of NU-6 was 0.5% with a range from –2% to 3%. Only slightly larger differences were observed at the next two highest presentation levels, 22- and 16-dB SL, at which the mean performance differences were 1.2% and 3.5%, respectively, again with better performance on the Auditec version. A slightly different picture emerged with the OHL listeners. At the highest presentation level, 38-dB SL, (a) the mean performance difference was 3.2% with the Auditec version of NU-6 being the easier, (b) the performance differences ranged from –5% (S16) to 16% (S14), and (c) 33 of the 36 listeners (92%) had recognition performance differences that were $\leq 10\%$. At the second highest level with the OHL listeners, 30-dB SL, the mean performance difference was 4.4%, again with the Auditec version the easier. Even at 30-dB SL, 30 of the 36 OHL listeners (83%) had differences between speakers

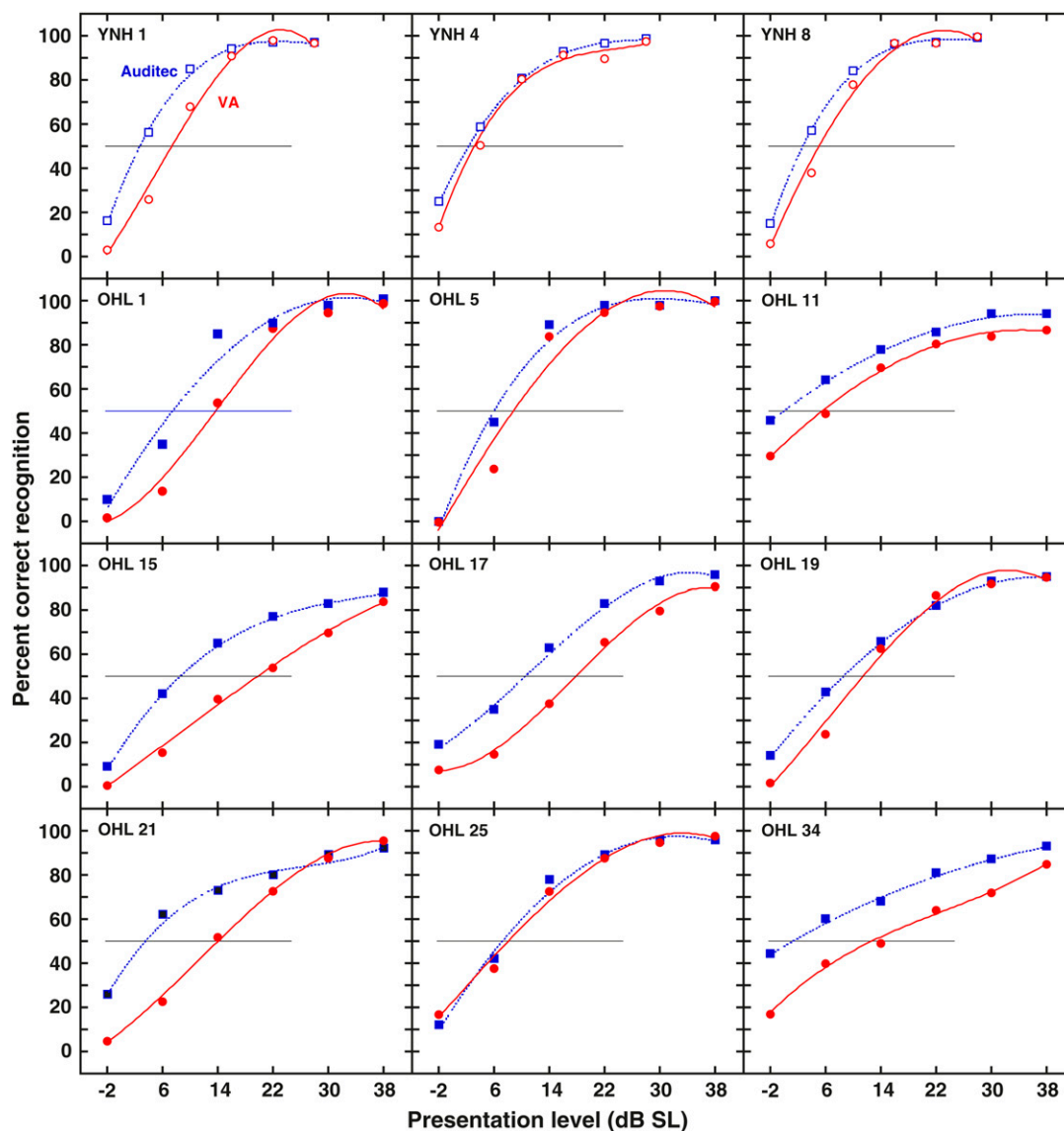


Figure 9. Representative psychometric functions for the Auditec (blue squares) and VA (red circles) versions of NU-6 obtained from three YNH listeners (top row) and nine OHL listeners (bottom three rows). Third-degree polynomials are used to describe the data. The psychometric functions for all of the 48 listeners are shown in Supplemental Figures S3–S6.

that were $\leq 10\%$. The implication here is that clinically, where word-recognition materials typically are presented at levels between 30- and 40-dB SL, word-recognition performances on the Auditec and VA versions of NU-6 are not that different for the clear majority of OHL listeners.

To examine in more detail, the recognition performances by each listener on the two versions of NU-6, bivariate plots were developed. These plots, which are shown in Figure 10, involve 72 datum points for the YNH listeners (12 subjects by 6 presentation levels; upper panel) and 216 datum points for the OHL listeners (36 subjects by 6 presentation levels; lower panel), with the Auditec recognition performances on the ordinate and the VA performances on the abscissa.

The data in the figure emphasize the extent to which performances were better on the Auditec version of NU-6 than on the VA version. Of the 72 mean performances by the YNH listeners, 54 (75.0%) were above the line of equality, indicating better performance on the Auditec version, 8 (11.1%) were on the line, indicating equal performances on the two versions, and 10 (13.9%) were below the line, indicating better performances on the VA version. Similarly, the OHL listeners performed better on the Auditec version with 182 of the 216 mean performances (84.3%) above the line of equality (better on the Auditec version), 11 (5.1%) on the line, and only 23 (10.6%) below the line (better on the VA version). The data in Figure 10 demonstrate most conclusively that the Auditec version of NU-6 was

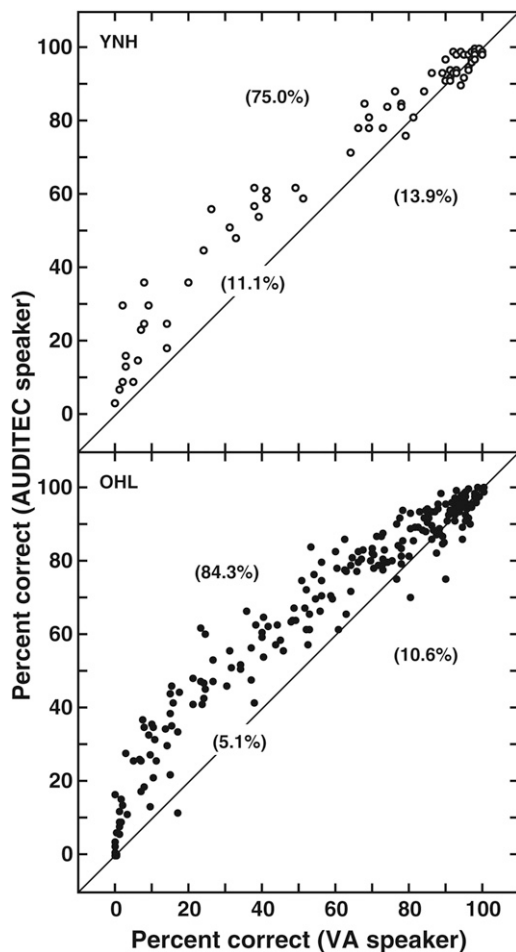


Figure 10. A bivariate plot of the recognition performances on the 200 NU-6 words spoken by the Auditec speaker (ordinate) and by the VA speaker (abscissa) at each of the six presentation levels for each of the 12 YNH listeners (upper panel, 72 datum points) and each of the 36 OHL listeners (lower panel, 216 datum points). The numbers in parentheses are the percent of datum points above, on, and below the line of equality. The data were jittered with a random additive algorithm from -0.4% to 0.4% in 0.1% steps.

slightly but consistently easier than the VA version with both YNH and OHL listeners.

Slopes of the Functions

As Wilson and Margolis (1983: p. 86) discussed, the function of the mean data averaged across the independent variable, which here is presentation level, provides the best estimate of recognition performance (in % correct) by the individuals in the group. Except for unique circumstances, the slope of the mean function does not provide the most accurate indication of the slopes of the individual participant functions. The mean slope of a group of functions can be obtained by averaging across predetermined points of the dependent variable (percent correct in this case). To accomplish this, the

individual subject data for each of the two speakers were fit with third-degree polynomials that then were evaluated between 20% and 80% in 10% intervals for the corresponding SL (dB) values. The mean functions for each speaker were generated by averaging across the subject functions at these specified intervals in the dependent variable domain and are presented in Figure 11 and Table 4 for the YNH and OHL listeners (the individual data are listed in Supplemental Tables S11 and S12). From the mean functions in Figure 11, the slopes at the 50% points for the YNH listeners were both $5.1\%/dB$, whereas the slopes of the linear_{20%–80%} segments were $4.8\%/dB$ and $5.0\%/dB$ for the Auditec and VA versions of NU-6, respectively, which are about the same as the slopes of the functions for the YNH listeners depicted in Figure 8. For the OHL listeners, the slopes of the mean functions at the 50% point in Figure 11 were $3.2\%/dB$ and $3.0\%/dB$ for the Auditec and VA versions of NU-6, respectively, with linear_{20%–80%} slopes of $2.8\%/dB$ and $3.0\%/dB$, respectively. Again, as with the YNH listeners, the slopes of the functions for the OHL listeners were only slightly steeper than the slopes of the corresponding mean functions depicted in Figure 8. Perhaps, the most accurate mean slopes are calculated from the slopes of the functions from the individual listeners, which also enables quantification of the variability. For the YNH listeners, the mean slopes at the 50% points on the functions were $5.2\%/dB$ ($SD = 0.8\%/dB$) and $5.3\%/dB$ ($SD = 0.5\%/dB$) for the Auditec and VA versions of NU-6, respectively, with

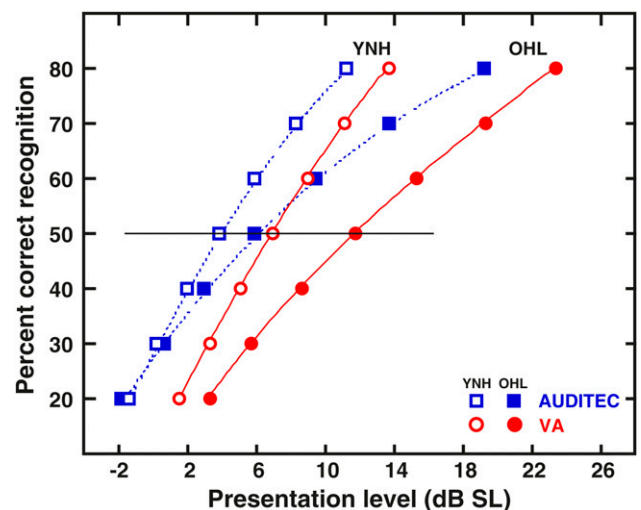


Figure 11. The mean recognition performances as a function of presentation level (dB SL) on the NU-6 words spoken by the Auditec speaker (blue squares) and by the VA speaker (red circles) are shown for 12 YNH listeners (open symbols and dotted lines) and for 36 OHL listeners (filled symbols and solid lines). The 20–80% points first were calculated from the polynomials used to describe the data for each subject and then were averaged across subjects and the dependent variable. Second-degree polynomials are used to describe the data.

Table 4. Five Measures of the Presentation Level (dB SL) at Which Seven Levels of Recognition Performance (in % correct) were Obtained from the 12 YNH and 36 OHL on the NU-6 Materials Recorded by the Auditec Speaker and by the VA Speaker

| | % Correct | Auditec Speaker | | | | | VA Speaker | | | | |
|---------------|-----------|-----------------|-----|------|-------|-------|------------|-----|------|-------|-------|
| | | Mean | SD | Max | Min | Range | Mean | SD | Max | Min | Range |
| YNH Listeners | | | | | | | | | | | |
| | 20% | −1.4 | 1.9 | 1.5 | −4.5 | 5.9 | 1.5 | 2.5 | 6.3 | −1.8 | 8.1 |
| | 30% | 0.2 | 2.0 | 3.3 | −2.9 | 6.1 | 3.2 | 2.6 | 8.3 | 0.1 | 8.2 |
| | 40% | 2.0 | 2.2 | 5.6 | −1.2 | 6.7 | 5.0 | 2.7 | 10.2 | 1.5 | 8.6 |
| | 50% | 3.9 | 2.3 | 7.9 | 0.8 | 7.1 | 6.9 | 2.6 | 11.9 | 3.1 | 8.7 |
| | 60% | 5.9 | 2.4 | 10.3 | 3.0 | 7.2 | 8.9 | 2.5 | 13.6 | 5.0 | 8.5 |
| | 70% | 8.3 | 2.4 | 12.8 | 5.7 | 7.0 | 11.1 | 2.4 | 15.4 | 7.4 | 8.0 |
| | 80% | 11.2 | 2.4 | 15.6 | 9.0 | 6.7 | 13.6 | 2.0 | 17.4 | 10.6 | 6.8 |
| OHL Listeners | | | | | | | | | | | |
| | 20% | −1.8 | 4.8 | 7.5 | −13.3 | 20.8 | 3.2 | 4.7 | 12.6 | −8.9 | 21.6 |
| | 30% | 0.7 | 4.8 | 10.3 | −9.4 | 19.7 | 5.6 | 5.5 | 14.9 | −12.0 | 26.9 |
| | 40% | 3.0 | 4.7 | 12.8 | −6.8 | 19.7 | 8.6 | 5.2 | 17.1 | −6.6 | 23.7 |
| | 50% | 5.9 | 4.6 | 15.3 | −3.9 | 19.2 | 11.7 | 4.8 | 19.6 | 1.0 | 18.6 |
| | 60% | 9.5 | 4.3 | 17.8 | −0.3 | 18.1 | 15.2 | 4.5 | 24.2 | 4.5 | 19.7 |
| | 70% | 13.7 | 4.6 | 21.6 | 2.9 | 18.7 | 19.2 | 5.1 | 29.5 | 6.7 | 22.7 |
| | 80% | 19.2 | 5.5 | 33.3 | 5.9 | 27.4 | 23.3 | 5.9 | 38.7 | 9.7 | 29.1 |

Note: The data were calculated from the polynomial equations used to describe the recognition performances by the Auditec and VA speakers for each individual listener.

corresponding linear_{20%–80%} slopes of 4.9%/dB (SD = 0.7%/dB) and 5.0%/dB (SD = 0.4%/dB). For the OHL listeners, the mean slopes at the 50% points on the functions were 3.3%/dB (SD = 1.1%/dB) and 3.5%/dB (SD = 1.0%/dB) for the Auditec and VA versions, respectively, with corresponding linear_{20%–80%} slopes of 3.1%/dB (SD = 0.8%/dB) and 3.4%/dB (SD = 1.1%/dB). When calculated using the individual psychometric functions, the slopes of the functions for both the YNH and OHL listeners were about 0.5%/dB steeper than comparable slopes calculated from the various forms of the mean data.

During sustained increases in the presentation level of a target word from inaudible to full recognition, the word is first detected, followed by incremental increases in the number of intelligibility cues available to the listener that eventually and collectively contribute to correct recognition of the word. This process, which involves different cues becoming audible at different presentation levels, is straightforward for most YNH listeners but is compounded by the effects of sensorineural hearing loss. With OHL listeners, the characteristics associated with hearing loss eliminate many of the intelligibility cues in words that are normally available to the YNH listener and/or delay on the presentation level continuum the available intelligibility cues in words until high presentation levels are attained. Because of audibility limitations imposed by a sloping sensorineural hearing loss with OHL listeners, the intelligibility cues are “recruited” over a wider range of presentations levels than they are with YNH listeners. That is, in all probability, the increments in in-

telligibility cues depend on many factors and occur at different rates for different listeners. These dynamics contribute to the slopes of the individual and group word-recognition functions of OHL listeners being more gradual than the slopes of the corresponding functions for YNH listeners.

Data in dB HL

Distortion aside, for the most part at suprathreshold levels, the ear with a sensorineural hearing loss performs (i.e., has similar perceptual capabilities) in a manner similar to the auditory behavior of a young normal ear at those same suprathreshold levels. Speech-recognition testing uses this suprathreshold effect by presenting test items at levels substantially above threshold, which in effect is a correction factor for the audibility differences between YNH and OHL listeners. In terms of speech-recognition performance, this correction factor is fairly effective except (a) a wider range of presentation levels is required by the OHL listeners to achieve maximum recognition performance and (b) maximum performances by OHL listeners never quite achieve the maximum performances attained by YNH listeners. When the SL correction factor is removed from the OHL listener data, then the differences between the two subject groups become exaggerated as illustrated in Figure 12, in which the independent variable (presentation level) has been transformed for each listener from SL to HL. The obvious differences between the recognition data for the individual listeners expressed in SL (Supplemental Figure S2) and in HL

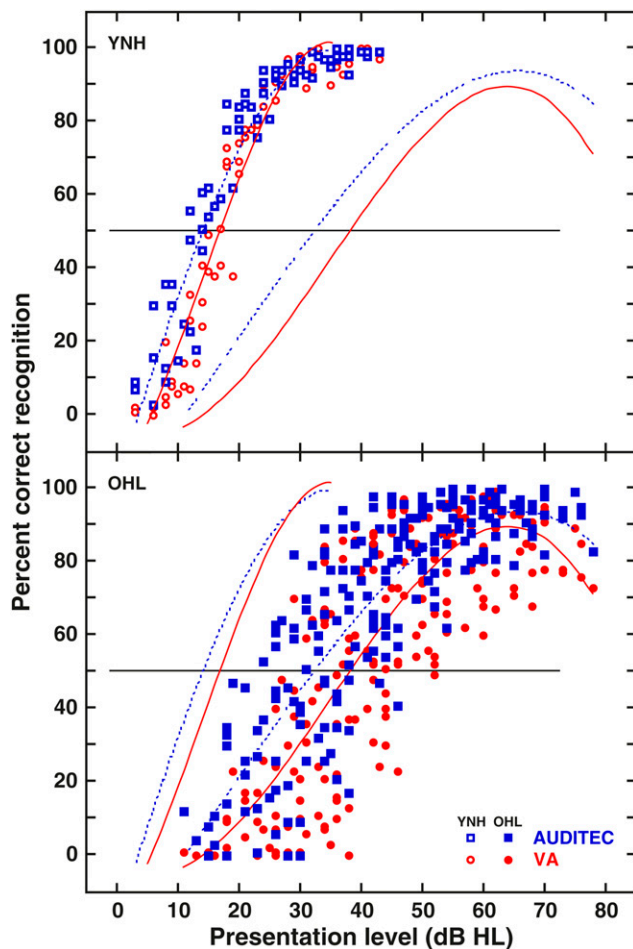


Figure 12. The individual subject recognition performances as a function of presentation level (dB HL) on the NU-6 words spoken by the Auditec speaker (blue squares) and by the VA speaker (red circles) are shown for 12 YNH listeners (top panel) and for 36 OHL listeners (bottom panel). The dotted functions in each panel serve as reference points for the other subject group. Third-degree polynomials are used to describe the data.

(Figure 12) are that in HL, the constraints imposed by SL are released that in effect produces more intersubject variability, especially in the OHL listeners who with respect to audibility are less homogeneous than are the YNH listeners. (*Note:* Figure 12 and Supplemental Figure S2 have the same decibel ranges on the abscissae, just different references). By definition in SL, the presentation levels are limited to ranges of 30 dB (YNH) and 40 dB (OHL), whereas in HL, the presentation levels for the YNH listeners range from 3- to 43-dB HL (40 dB) and for the OHL listeners range from 11- to 78-dB HL (67 dB).

The mean 50% points calculated from the polynomial equations for the Auditec and VA versions of NU-6 in Figure 12, respectively, were 13.7- and 17.0-dB HL (YNH) and 32.2- and 39.5-dB HL (OHL). The performance differences at the 50% points were slightly larger in HL (Figure 12) than in SL (Supplemental Figure S2),

which was an anticipated reflection of the increased intersubject variability when the data were scaled in HL. In sensation level and HL, the respective mean speaker differences at the 50% points were 3.1 dB and 3.3 dB (YNH listeners) and 5.7 dB and 7.3 dB (OHL listeners), again with the Auditec version being the easier version. Other mean differences between the data in Figures 8 and 12 are the comparisons between the two subject groups, that is, recognition performance by the OHL listeners minus performance by the YNH listeners, which in HL at the 50% points were 18.5 dB (Auditec speaker) and 22.5 dB (VA speaker). Again, because of the larger variability in HL, these differences were substantially larger than the corresponding 2.2 dB and 4.8 dB differences observed between the same variables in SL.

The slopes of the mean Auditec and VA functions in Figure 12, which were calculated from the first derivatives of the polynomial equations at the 50% points, were 4.2%/dB and 4.6%/dB (YNH) and 2.1%/dB and 2.2%/dB (OHL) for the Auditec and VA functions, respectively. For the YNH listener functions, the linear_{20%-80%} slopes were 4.0%/dB (Auditec) and 4.4%/dB (VA). The corresponding linear_{20%-80%} slopes for the OHL listeners were 2.0%/dB (Auditec) and 1.8%/dB (VA). These functions expressed in HL were 0.2–0.8%/dB more gradual than when the independent variable was SL, which again was attributed to the greater intersubject variability associated with the measures expressed in HL. All of these slope values based on the mean data are somewhat more gradual than the mean slopes of the individual listeners, which in HL would be the same as the previously described mean slopes of the individual listeners calculated in SL (Figure 8).

Both the locations of the functions (in HL) and the slopes of the functions in Figure 12 compare favorably with previous data. In the Wilson and Oyler (1997) study of the Auditec version of NU-6 with similar listener groups (Table 1), almost identical mean results to those of the present study were obtained at the 50% points, 13.4-dB HL (YNH, $m = 4.5\%/dB$) and 34.8-dB HL (OHL, $m = 2.3\%/dB$). Also, with the Auditec version, Wilson et al (1976) had mean 50% points of 10.3-dB HL ($m = 4.2\%/dB$) with YNH listeners and of 46.0-dB HL ($m = 4.2\%/dB$) with OHL listeners; the 46.0-dB HL value is somewhat understandable, given the PTA in the older study was about 6 dB higher than the PTA of the OHL listeners in the present study. Three studies of the VA version of NU-6 with YNH listeners show very good agreement with the current data. As shown in Table 2, the 50% points from three earlier studies ranged from 15.2- to 17.1-dB HL, with slopes from 3.8 to 4.3%/dB (Causey et al, 1983; Wilson et al, 1990; Stoppenbach et al, 1999), which compare favorably with the same metrics from the present study at the 50% points (17.0-dB HL and 4.6%/dB). Collectively, the agreement between the current data and the data

from these earlier investigations attest to the concurrent validity of the results in this report.

Individual Word Data

To this point in the results, the dependent variable was the average performance on the words at each presentation level for each subject, which provided information on the subjects including intersubject variability data, but no information about the recognition performances on each of the 200 NU-6 words. To examine the recognition performances on the individual words, especially with respect to interword variability, the data were recast so the dependent variable was the average performance by the subjects at each presentation level for each word. The data in this form (200 words by 6 levels) are shown in Figure 13 for the two subject groups as bivariate plots with the performance from the Auditec version on the ordinate and from the VA version on the abscissa. For graphic clarity, the data were jittered with an additive random algorithm, which produced the clustering of responses at each percent correct bin (*Note*: the number of response clusters reflects the different sizes of the two subject groups). First, with the YNH listeners of the 1,200 word comparisons (200 words by 6 presentation levels), 561 comparisons (46.8%) had equal recognition performances on the materials spoken by the two speakers, 444 (37.0%) had better performances with the Auditec speaker, and 195 (16.3%) had better performances with the VA speaker. The vast majority of the 561 equal performances were at either 100% correct recognition, ($n = 389$ or 69.3%), or at 0% correct recognition ($n = 68$ or 12.8%), which reflect ceiling and floor effects, respectively. Second, with the OHL listeners, 772 (64.3%) of the 1,200 comparisons had better performances with the Auditec speaker, 234 (19.5%) were better with the VA speaker, and 194 (16.2%) had equal performances with the two speakers. Excluding the equal performances, with both listener groups, better recognition performances were 2.3 times (YNH, 37.0%/16.3%) and 3.3 times (OHL, 64.3%/19.5%) more prevalent with the words spoken by the Auditec speaker than by the VA speaker. As with the individual listeners, most of the individual NU-6 words were more intelligible when spoken by the Auditec speaker than when spoken by the VA speaker, which supports the third hypothesis that recognition performances on most of the words would be better when spoken by the Auditec speaker than when spoken by the VA speaker but a minority of words would exhibit better performances when spoken by the VA speaker.

Figure 14 depicts representative examples of the psychometric functions for ten individual words with each panel containing data from the two speakers and the two subject groups. The data in this format for each of the 200 NU-6 words are in Supplemental Figures

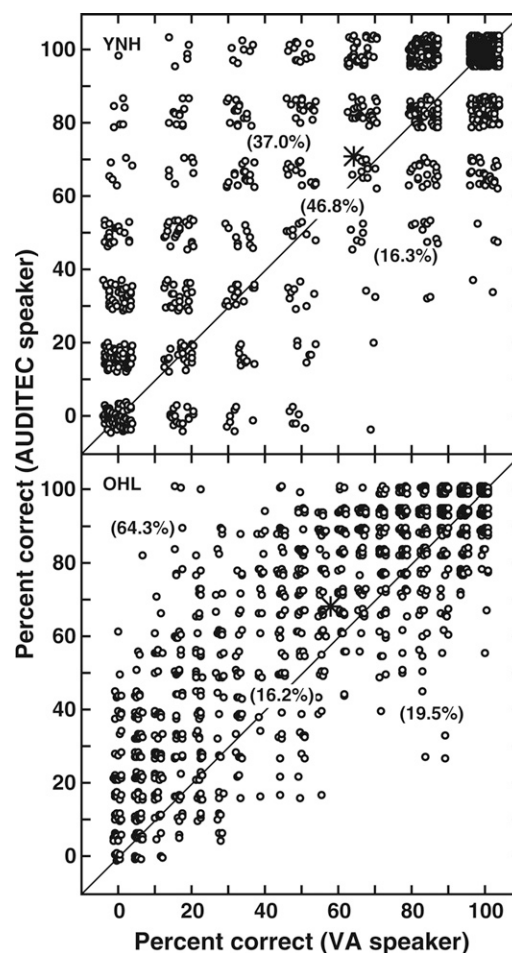


Figure 13. A bivariate plot of the recognition performances on the 200 NU-6 words spoken by the Auditec speaker (ordinate) and by the VA speaker (abscissa) at each of the seven possible percent correct categories for the 12 YNH listeners (upper panel, circles) and each of the 19 possible percent correct categories for the OHL listeners (lower panel, squares). The starburst (need starburst) represents the mean performances in each panel. The numbers in parentheses are the percent of datum points above, on, and below the line of equality. The data were jittered with a random additive algorithm $\pm 4\%$ in 0.25% steps (YNH) and $\pm 1.0\%$ in 0.2% steps (OHL).

S7–S26 (listed in Supplemental Tables S3–S10; available with the online version of this article). As a reference, remember the systematic nature of the mean functions shown in Figure 8. The functions for the individual words demonstrate how similar intelligibility-wise some words were for both the two speakers and the two listener groups, how totally dissimilar other words were for the speakers and listener groups, and everything in-between. In the first Figure 14 example, *chief*, the lower segments of the four functions demonstrate similar recognition performances but above the 50% points the performances on the materials spoken by both speakers become unique for the two listener groups with performances substantially better by the YNH group. Slightly, different relations were

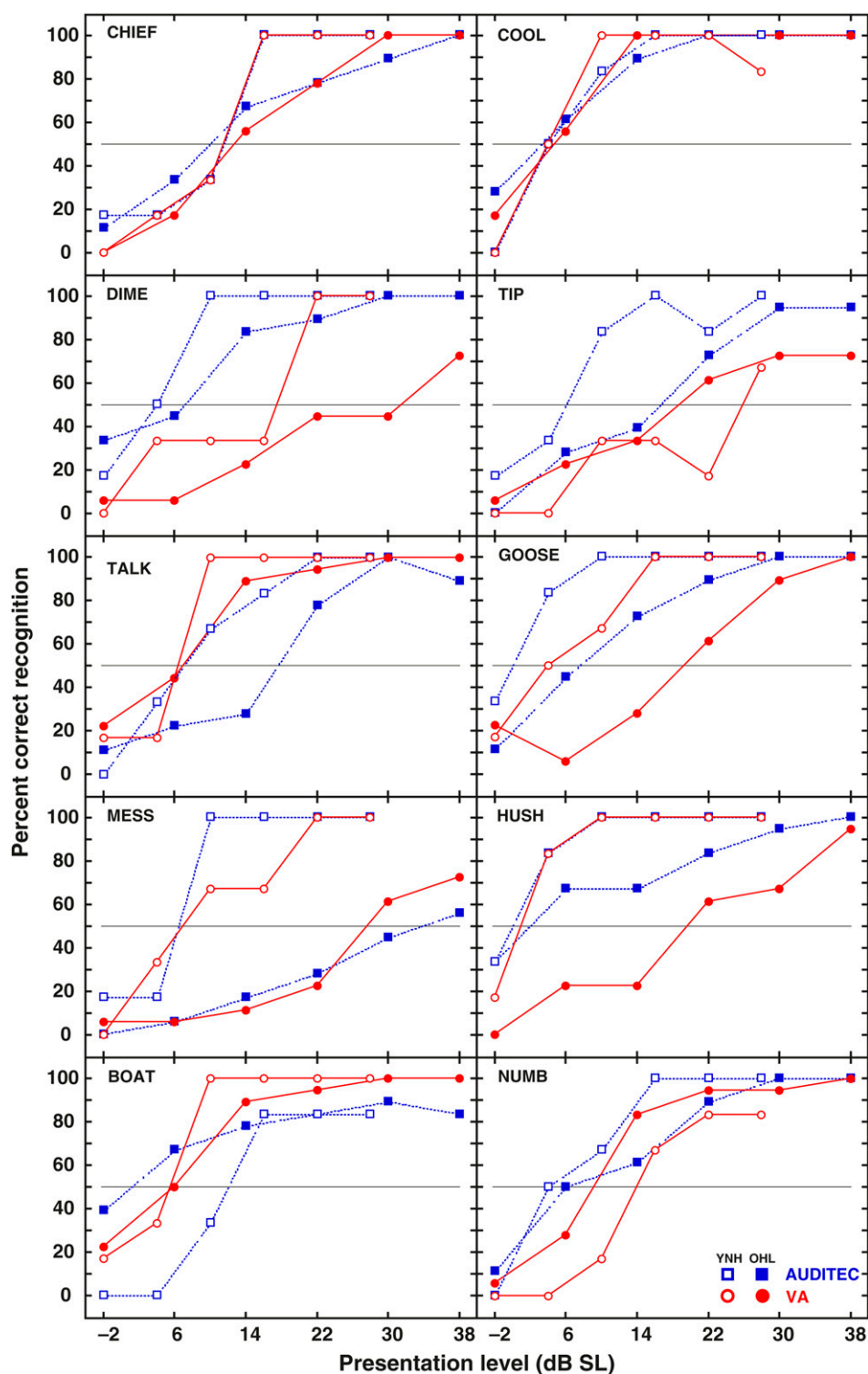


Figure 14. Psychometric functions for ten representative NU-6 words spoken by the Auditec (squares) and VA (circles) speakers obtained from YNH listeners (open symbols) and OHL listeners (filled symbols). The psychometric functions for the individual 200 words are shown in Supplemental Figures S7–S26.

demonstrated by the word *cool* for which the four functions were pretty much the same, except at -2 -dB SL. With *dime*, large overall differences between speakers emerged with Auditec version being overall easier

by 27.8% (YNH) and 42.6% (OHL). The word *tip* presents an interesting set of data, again with the Auditec version being easier than the VA version, but this time, the overall differences were 44.4% for the YNH

listeners and only 10.2% for the OHL listeners. The data from the second two words in Figure 14 (*dime* and *tip*) clearly indicate better performances with the Auditec speaker than with the VA speaker. The functions for the fifth word in the figure, *talk*, indicate just the opposite relation between speakers with better overall performances of 8.3% (YNH) and 20.4% (OHL) with the VA speaker than with the Auditec speaker. The data for *goose* start to approach the mean data shown in Figure 8 in that both groups of listeners performed better overall on the Auditec version by 13.9% (YNH) and 18.5% (OHL) than on the VA version and the YNH listeners performed better than the OHL listeners. The word *mess* produced substantial group differences for both speakers with the YNH listeners 31.5% (VA speaker) and 47.2% (Auditec) better than the OHL listeners. In all probability, the performance differences reflect the difficulties the OHL listeners had with the final consonant, /s/, in the target word. A slightly different finding was observed with *hush*. With *hush*, the functions for the two speakers are essentially the same for the YNH listeners, with a 2.8% overall difference, whereas with the OHL listeners there was a substantial disparity between performances with both between speakers (29.6%), which again was probably related to the final consonant in *hush*, /ʃ/. *Boat* demonstrates better overall performance by 27.8% throughout the range of presentation levels by the YNH listeners on the VA version than on the Auditec version. By contrast, with the OHL listeners, the functions for the two speakers intersected with recognition performance on the VA speaker version better than on the Auditec version at the four highest presentation levels and poorer at the two lowest levels. The final exemplary word in Figure 14 is *numb*, which on the YNH listeners showed a distinct 27.8% better overall performance on the Auditec version than on the VA version, whereas the two speaker functions with the OHL listeners were intertwined equivalent overall performances of 67.6% (VA speaker) and 68.5% (Auditec speaker). Although the data depicted in Figure 14 and expounded on in Supplemental Figures S7–S26 appear at times to be chaotic and at times not systematic, the locations and shapes of the functions and the relations among the functions are for the most part representative of the two types of listeners studied. Substantial increases in the numbers of listeners would improve the systematics of the individual functions but would not substantially alter the locations and shapes of the functions and the relations among the functions.

Introspective Reports from the Subjects

Following data collection in Session 3, each participant was asked if one speaker was easier to understand than the other speaker. Eleven of the 12 YNH listeners

(91.7%) and 30 of the 36 OHL listeners (83.3%) indicated that the Auditec speaker was easier to understand than the VA speaker. Although all subjects in both groups had better overall recognition performance on the NU-6 spoken by the Auditec speaker, with the OHL listeners, two (5.6%) listeners thought that the VA speaker was easier to understand and four (11.1%) thought the two speakers were equally understandable. When the ease/difficulty of the speakers were rated on a Likert scale from one (easy to understand) to ten (difficult to understand), interestingly, both subject groups gave similar values. The Auditec speaker received mean ratings of 4.3 (SD = 2.0) and 4.2 (SD = 1.6) by the YNH and OHL groups, respectively, whereas the VA speaker ratings were 5.8 (SD = 1.7) and 5.9 (SD = 1.5) for the respective groups. These observations from both groups of listeners support from a subjective point of view the objective psychometric data developed in the experiment that the Auditec speaker was easier to understand than the VA speaker.

Three Test Sessions

Concern is often expressed that repeated exposure to the materials used in word-recognition tasks are subject to improved recognition performances across the multiple presentations owing to learning effects like increased familiarity with the target words or learning to listen in a complex or unfamiliar listening environment such as in a degraded speech task. Because the current experiment involved (a) a simple listening environment in quiet, (b) presentation levels that were random, and (c) a practice list in each session to acquaint/refresh the listener to the listening and response paradigms, learning effects were expected to be minimal. Again, the recognition data were recast to provide psychometric functions for each of the three test sessions by speaker (Auditec and VA) and by subject group (YNH and OHL) and are illustrated in Figure 15. The individual listener data for the three sessions are shown in Supplemental Figures S27–S30. Visual inspection of the data in the figure suggests little change in recognition performance occurred across the three sessions, the only exception being a 1-dB improvement in Session 3 with the Auditec speaker by the OHL listeners (Q1 in Figure 15). For the YNH listeners, the 50% correct recognition points for the three sessions, which were calculated from the polynomial equations used to describe the data, were at 4.0-, 3.7-, and 4.1-dB SL (Auditec speaker) and 7.1-, 7.6-, and 6.5-dB SL (VA speaker) with all slopes at the 50% points in the 3.7%/dB to 4.7%/dB range. At 28-dB SL, the mean performances across the three trials of the YNH listeners were 98.5%, 98.4%, and 99.1% (Auditec speaker) and 98.2%,

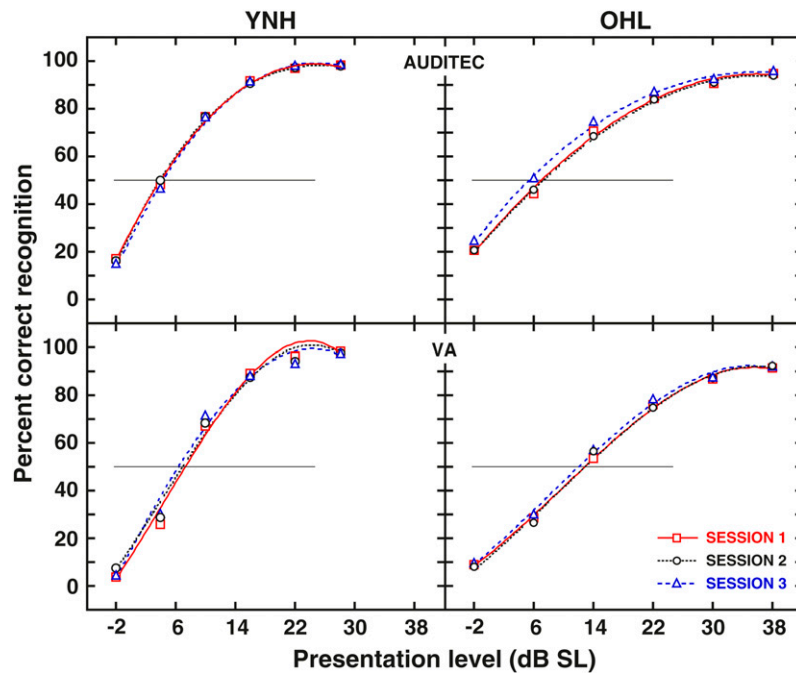


Figure 15. The psychometric functions from the three test sessions are shown for the YNH listeners (left panels) and for the OHL listeners (right panels) in response to the Auditec speaker (upper panels) and the VA speaker (lower panels). Third-degree polynomials are used to describe the data.

97.9%, and 97.4% (VA speaker). For the OHL listeners, the 50% correct recognition points for the three sessions were at 7.0-, 7.1-, and 5.1-dB SL (Auditec speaker) and 12.8-, 12.8-, and 11.9-dB SL (VA speaker) with all slopes in the 2.9%/dB to 3.1%/dB range. At 38-dB SL, the mean performances across the three trials of the OHL listeners were 95.0%, 94.1%, and 96.0% (Auditec speaker) and 91.4%, 92.4%, and 92.2% (VA speaker). These mean data across the three test sessions support the contention that practice/learning effects had no appreciable effect on the present data that would be of experimental or clinical concern.

Traditional 50-Word Lists and 25-Word Half Lists

The traditional word-recognition lists, including the PB-50s (Egan, 1948), the CID W-22s (Hirsh et al, 1952), and NU-6, comprised 50 words, which was the number of words in a list that Egan found necessary to achieve a semblance of phonetic balance. The recognition performance data from the YNH and OHL listener groups in the present study were compiled into the original four NU-6 lists and are listed in Supplemental Table S13 and depicted in Supplemental Figure S31 (Auditec speaker) and Supplemental Figure S32 (VA speaker). A couple of relations are apparent from the two datasets. First, as with the overall result of the present study discussed earlier, better recognition performances were obtained by the YNH listeners than by the OHL listeners at each of the eight list comparisons

(two speakers by four lists). At the 50% points on the functions, mean recognition performances for the YNH listeners among the four 50-word lists ranged 1.1 dB for the Auditec speaker from 3.2- to 4.3-dB SL and 1.5 dB for the VA speaker from 6.2- to 7.7-dB SL. The same performance measures for the OHL listeners among the four lists ranged 3.3 dB from 4.9- to 8.2-dB SL (Auditec) and 2.3 dB from 11.6- to 13.9-dB SL (VA). Considering the slopes of the functions at the 50% points of about 5%/dB (YNH) and 3%/dB (OHL), the differences between performances translate to 5–7% for YNH listeners and 7–10% for the OHL listeners. These slight differences among the four NU-6 lists should not be of clinical concern, especially when consideration is given, the presentation levels involved, which were all below 15-dB SL at the 50% points. Second, word-recognition materials typically in the clinic are presented in quiet at 30- to 40-dB SL, which are levels comparable to the two highest presentation levels used in the present study with the OHL listeners. At 30- and 38-dB SL differences among recognition performances on the four NU-6 lists in the present study differed by 1 dB or less. At 30-dB SL, the recognition performances on the NU-6, 50-word lists for the OHL listeners ranged 1.3% from 91.1% to 92.4% (Auditec) and 0.7% from 86.8% to 87.5% (VA). At 38-dB SL, slightly higher performances were obtained for the OHL listeners that ranged 0.8% from 94.6 to 95.4% (Auditec) and 2.9% from 90.2% to 93.1% (VA). For comparison, with the YNH listeners, all mean recognition

performances on the four, NU-6 50-word lists at the two highest presentation levels (22- and 28-dB SL) for both speakers were >94% correct.

The clinical use of 25 words instead of 50 words for word-recognition testing has been discussed and studied since the 1960s (Elpern, 1961; Resnick, 1962; Grubb, 1963; Campbell, 1965), mainly as a time-saving measure (Wiley et al, 1995). Although statistical questions have been raised regarding these shortened version word-recognition lists (e.g., Thornton and Raffin, 1978), the use of these so-called *half-lists* is probably more common than the use of full lists, especially when multiple presentation levels of word-recognition materials are a component of the clinic protocol (Martin et al, 1998; DeBow and Green, 2000). The current data afforded an examination of the equivalency of the 25-word lists. (*Note:* as used in this context, *equivalency* should be used with caution because lists of words that produce equivalent recognition performances for one group of listeners may not necessarily produce equivalent performances for a different group of listeners). The mean word-recognition data for each of the 16, 25-word lists of NU-6, Randomization A (eight lists by two speakers) are shown in Supplemental Figure S33 (YNH) and Supplemental Figure S34 (OHL). In each figure, the upper four panels display the data for the Auditec speaker and the lower four panels illustrate the data for the VA speaker. Visual inspection of each figure indicates little difference between the functions for the first 25-words of each list and the second 25-words of each list. In fact, for each speaker and each listener group, the respective eight half-list functions are pretty much the same. With the Auditec speaker, the eight, 50% points on the functions (a) for the YHN listeners ranged 2.6 dB from 2.8-dB SL (List 2b) to 5.4-dB SL (List 3a), with a mean of 3.9-dB SL, and (b) for the OHL listeners ranged 4.5 dB from 4.2-dB SL (List 2b) to 8.7-dB SL (List 3a), with a mean of 6.5-dB SL. For the VA speaker, the eight, 50% points (a) for the YHN listeners ranged 2.4 dB from 5.7-dB SL (List 3b) to 8.1-dB SL (List 4b), with a mean of 3.9-dB SL, and (b) for the OHL listeners ranged 3.3 dB from 10.9-dB SL (List 2b) to 14.2-dB SL (List 4a), with a mean of 6.5-dB SL. As with the 50-word lists, the slopes of the functions were consistently around 5%/dB (YNH) and 3%/dB (OHL) and at the maximum presentation levels the recognition performances on the eight sets of materials were essentially the same.

It is not surprising that the various 50- and 25-word lists produce the same mean results within each listener group, even though as shown in Figure 14 (and Supplemental Figures S7–S26) equal intelligibility is not obtained with each word at equal presentation levels. The performance equality among the lists, especially at the highest presentation levels, is attributable to the 200 CNC NU-6 words being simple words familiar to most

listeners that were recorded at similar levels. For the YNH listeners at 28-dB SL, 93.5% (Auditec) and 89.0% (VA) of the words were correct 100% of the time. As shown in Figure 16, for the OHL listeners, 90% (Auditec) and 77.0% (VA) of the words were correct $\geq 88\%$ of the time at 38-dB SL. The data in Figure 16 also indicate that there were a few words in each recorded version of the materials that produce maximum intelligibility that was >80% correct. These few so-called *difficult words* intelligibility-wise appear in most recorded versions of monosyllabic word lists (Wilson and McArdle, 2015). When the data from the Auditec and VA versions of the NU-6 words were randomized into lists of 25 or 50 words for the OHL listeners, the lists produce percent correct data that were very similar, especially when the materials were presented at a high SL. To demonstrate this effect, the current percent correct data from the OHL listeners at 38-dB SL were randomized into 100, 50-word lists and into 200, 25-word lists. For the Auditec version of NU-6, the 100 mean 50-word list

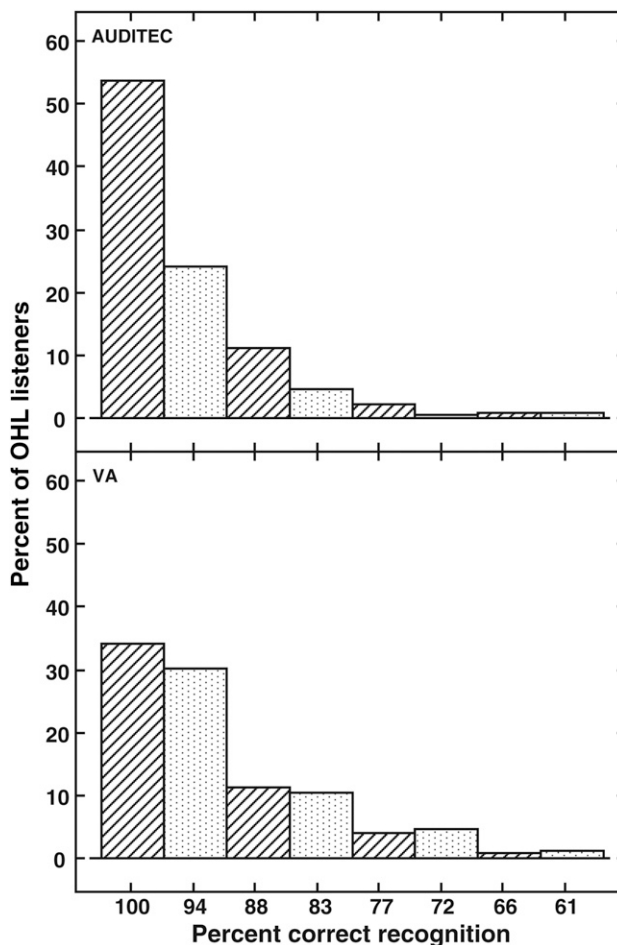


Figure 16. The percent of OHL listeners at the respective percent correct recognition intervals when the NU-6 lists were presented at 38-dB SL. The data from the Auditec speaker (top panel) and VA speaker (bottom panel) are shown. The abscissa values have been rounded to the nearest whole number.

performances ranged 5.0% from 92.2% to 97.2% and the 200 mean 25-word performances ranged 8.4% from 90.4% to 98.9%. For the VA version, the 100 mean 50-word list performances ranged 4.6% from 89.7% to 94.2% and the 200 mean 25-word performances ranged 8.7% from 86.9% to 95.6%. Similar findings were reported by Wilson et al (2008) in a study involving 540 monosyllabic words spoken by the same speaker, the psychometric functions for the 12 organized, 50-word lists (PB-50, CID W-22, and NU-6) were essentially the same as 12 functions generated by randomly selecting from a pool of 500 words the 50 words for each. Collectively, these data indicate that there is little need to restrict the NU-6 words used to evaluate word-recognition abilities to the formalized four-word lists that were compiled initially by the test developers. The only requirements are recorded materials and normative data of the psychometric characteristics of the materials from both YNH and OHL listeners.

Another line of thinking regarding word-recognition testing is the use of words (a) that are homogeneous regarding the ease/difficulty of understanding the words or (b) that are on the difficult end of the understanding continuum. An example of the first instance is Margolis and Millin (1971) who devised two 25-word lists from the Hirsh recording of the W-22s by discarding words on which recognition performance was at or near 0% and 100% correct. Although the use of 25-word lists to evaluate word-recognition abilities is common in audiology practice, some investigators have suggested the use of word lists even shorter than 25 words. The most common suggestion is 10-word lists composed of the words from a set of materials that are the most difficult to understand. The Hurley and Sells (2003) 10-word lists of the Auditec version of NU-6 are an example of this approach. As a convenience for those interested in compiling word-recognition lists based on the ease/difficult criteria, the overall percent correct performances for each of the 200 NU-6 words spoken by the Auditec and VA speakers are listed in Supplemental Tables S14 and S15 (YNH) and in Supplemental Tables S16 and S17 (OHL). Any alterations in the word lists must be based on data from that particular recorded version of the lists derived from the appropriate group of listeners. As demonstrated earlier, arbitrarily shortening the word lists from 50 words, to 25 words, to ten words, or whatever number of words will result in increased intersubject variability.

DISCUSSION

Collectively, the recognition data from the two recorded versions of NU-6 indicate for YNH listeners that overall performance on the Auditec version of NU-6 (71.4%) was 7.3% better than performance on the VA version (64.1%). With the OHL listeners, the

overall difference was even greater (10.4%), 68.6% (Auditec) and 58.2% (VA). These overall results plus the results from the individual participants confirm the first two hypotheses (a) that recognition performances with the Auditec speaker would be better than the performances with the VA speaker and (b) that in comparison to the YNH listeners, the performance differences between the two versions of NU-6 would be maintained or exaggerated by the OHL listeners. The disparities between the two versions could easily be attributed to differences in presentation levels of the two sets of materials. Recall that both recordings of NU-6 were made using the ANSI standard that specified that the carrier phrase should be monitored to a target level on a vu meter with the target word being uttered naturally following the carrier phrase. These two versions of NU-6 were matched amplitude-wise in accordance with the way they were recorded and the levels of their calibration tones (carrier phrase and vu meter), neither of which guarantees that the target words have equivalent amplitudes. The calibration issues involve if and how the amplitudes of words can be evaluated equitably given the amplitude modulation that characterizes word waveforms. Perhaps, Stevens et al (1947: p. 771) expressed it best when they stated, *Speech is such a complex phenomenon, varying continuously in time, in frequency, and in intensity, that single measures become of necessity merely rough approximations; and multiple measures, if they try to be complete and definitive, bog down in complication.* To paraphrase Davis (1947), there is no rule stating that equal amplitudes of words produce equal intelligibility. It is doubtful that the calibration issue associated with speech signals will ever be resolved with the confidence associated with pure-tone calibration, but that should not impede attempts to seek reasonable solutions.

The vu meter (ASA, 1954), which was never intended to make precise amplitude measures of speech signals, integrates energy over time and can be considered a mechanical averager with an ~ 300 -ms time constant, which is the time it takes the meter needle to go from -20 vu to 0 vu. (Note: Current audiometers use a *monitoring meter* in place of a vu meter. The monitoring meter typically is a light bar that has a time constant of 350 ms, ± 10 ms [ANSI, 2010]). Most CNC monosyllabic words are 500–600 ms (Wilson, 2015), with the vowel segment half that duration (House, 1961). Vowels have a maximum, sustained amplitude that is usually < 100 ms. This limitation and the difficulty reading a moving needle (Levitt and Bricker, 1970; Lobdell and Allen, 2007) preclude the vu meter as an appropriate instrument for the accurate measurement of speech-signal amplitudes. With digital technology, the rms of a signal is a precise way to quantify the amplitude. In the Wilson report (2015), the mean rms of the two recorded versions of the 200 NU-6 words used in the

present study were -19.9 dB (Auditec; SD = 1.6 dB) and -18.2 dB (VA; SD = 1.7 dB), re: the maximum digitization range of the recordings. The mean rms values of the carrier phrases were much closer, -14.8 dB (Auditec; SD = 0.8 dB) and -14.1 dB (VA; SD = 0.5 dB). (*Note:* Because these rms values were referenced to the maximum digitization range, they indicate that the amplitudes of the carrier phrases and target words of the VA version of NU-6 were at higher levels than the Auditec version.) An alternative method to quantify the level of individual CNC words involves calculating the rms of the 50-ms vowel segment with highest amplitude. With these 400 NU-6 words, the amplitude of the vowel in each word was higher than the amplitude of either consonant in each word, which is a relationship that is not necessarily true in all speech signals. Although the amplitudes of vowels vary from vowel to vowel and are influenced by the phonemic neighborhood (House and Fairbanks, 1953; Lehiste and Peterson, 1959b; Jacewicz and Fox, 2008), the vowel segment calibration method is easy to implement and replicate. The mean maximum 50-ms vowel segments of the CNC words used in the present study were 3.8 dB higher for the VA speaker ($M = -15.9$ dB, SD = 1.2 dB) than for the Auditec speaker ($M = -19.7$, SD = 2.0 dB). So, if this vowel segment calibration method were used, then the performance differences between the two versions of NU-6 would have been 3.8 dB greater than those reported using the traditional calibration technique with the vu meter. There is no doubt that the presentation level (audibility) is the major contributor to the intelligibility of an utterance of a speech signal and is the easiest parameter to manipulate. Cognitive factors aside, there are other cues, characteristics, or parameters of the speech waveform, which are more difficult to manipulate than presentation level, that also contribute to the intelligibility of the speech signal.

From the mean data in Table 3, at 22- and 28-dB SL with the YNH listeners, recognition performances on the two versions of NU-6 for practical purposes were the same and can be considered maximum performance (94.6–98.1%), meaning that further increases in the presentation level would not produce increased recognition performances. At 16-dB SL, differences between the mean performances on the two versions emerged with 3.5% better performance on the Auditec version; the difference increased to 20% at 4-dB SL. For the OHL listeners, maximum recognition performances (95.1% and 91.8%) were obtained at 38-dB SL, which were 3.5–4.6% poorer than the maxima attained by the YNH listeners 10 dB lower at 28-dB SL. Even at 38-dB SL, mean performance was 3.3% better on the Auditec version of NU-6 than on the VA version. The recognition performances at these higher presentation levels are representative of the SLs typically used clinically. As can be seen in Tables 3 and 4 and in Figures 8

and 11, throughout the 20–80% correct range, better recognition performance was achieved by both YNH and OHL listeners on the Auditec version of NU-6 than on the VA version. This difference ranged from 8% at 10-dB SL to 20% at 4-dB SL (YNH) and from 8.4% at 22-dB SL to 18.7% at 6-dB SL (OHL). All of the OHL listeners in the present study had maximum recognition performances $>80\%$ (Auditec) and $>70\%$ (VA). The real gray area of comparison on the two versions of NU-6 is with listeners whose maximum recognition performance is $<70\%$. Take, for example, an individual with maximum word-recognition performance on either version of NU-6 of 54% correct. Based on the current data, no doubt this person would perform better on the Auditec version of NU-6, but it would not be appropriate to extrapolate linearly the 54% correct to the VA version. The reason is the 54% is a maximum value that is reflecting an auditory behavior not included in the current OHL listeners. To evaluate accurately the transfer function for the two versions of NU-6 for use on individuals with lower maximum word-recognition performances, OHL listeners with that lower maximum word-recognition characteristic need to be studied. For such a study, one would predict that the differences between the two versions of NU-6 would be in the same direction as those observed in the present study. This is reasonably because of some yet-to-be defined distortion, for lack of a better word, in the VA version of NU-6, a hint of which is seen even at the highest presentation levels at which slightly poorer performances were obtained by both listener groups on the VA version. Perhaps, it is this slight difference in recognition performances that is being reflected in the subjective ratings by the listeners that indicated overwhelmingly that the Auditec speaker was easier to understand than the VA speaker.

All the mean functions for the 12 YHN listeners (Supplemental Figure S3) had minimum recognition performances less than 40% correct at the lowest presentation level, -2 -dB SL. Likewise, a substantial majority of the mean functions for the 36 OHL listeners (83.3%) (Supplemental Tables S1 and S2; Supplemental Figures S4–S6) achieved minimum performances less than 40% at -2 -dB SL. There were, however, exceptions, viz., S2, S11, S14, S31, S34, and S35, all of whom had minimum recognition performances at -2 -dB SL $>40\%$ correct, in particular on the Auditec materials. The question was, why were these six performances at -2 -dB SL ($M = 47.5\%$ correct) appreciably better than the performances ($M = 16.5\%$) demonstrated by the remaining 30 OHL listeners at -2 -dB SL? There are possibly two reasons. First, these six OHL listeners for unknown reasons had unique listening skills and auditory processing abilities that enable them to achieve better recognition performances at these seemingly low audibility levels than were achievable even by the YNH listeners. This line of reasoning is difficult to support, much less to

understand. The second reason for these apparently remarkable recognition performances by the six OHL listeners at the lowest presentation levels is related to the use of SL, re: the three-frequency, PTA, as the presentation level of the target signals. Typically, the presentation level variable is a signal attribute that is tied to a physical standard such as sound-pressure level or HL (ANSI, 2010). The use of SL interjects a subjective component, in this case threshold measures, into the presentation level variable that essentially makes presentation level a quasi-independent variable. This subjective component was a vulnerability that can produce erroneous and undesirable effects on the measurements being made. For example, in the present study, the pure-tone thresholds of the six OHL listeners were quite possibly 5–10 dB higher than their “true” thresholds. A 5- to 10-dB inflation in the PTA was tantamount to one step on the presentation-level scale (8 dB), which if implemented would be enough to depress recognition performances less than 40% at –2-dB SL becoming more akin to the performances by the other listeners whose data were based on their true pure-tone thresholds. Except for S35 who had a 5 dB lower PTA in Sessions 2 and 3 than in Session 1, the remaining five OHL listeners produced consistent PTAs across the three sessions ($M = 35.0$ -, 36.0 -, and 37.0 -dB HL); hence, they were reliable measures of the pure-tone thresholds but perhaps just not particularly valid measures. This threshold issue involves how the *response criteria* of these six listeners, which is a subjective matter and may have been extremely conservative, might have differed from the response criteria of the remaining YNH and OHL listeners. Substantiation of these possibly differing response criteria associated with pure-tone thresholds awaits further study. Related to the SL issue is the possibility that the PTA as used in this project may not be a good reference of the presentation levels to use with some subjects to encompass the desired range of word-recognition performances. The PTA was established with a detection task, whereas the task in the present study was a recognition task. Perhaps, with the listeners whose recognition performances at –2-dB SL were >40% correct would have been better served with the SRT as the reference for the presentation levels. There has certainly been enough written on the relation between the PTA and SRT, but in the context of the present study, the concern expressed here about 5–8 dB is about individual listeners not groups of listeners. Overall, however, the PTA was a good reference for the vast majority of listeners in providing nearly the complete range of recognition performances.

SUMMARY

The data from the present study with the NU-6 materials indicate that word-recognition performance both by YNH and OHL listeners overall was better on

the Auditec recording than on the VA recording, which was a relation demonstrated by each of the 48 listeners. The overall performance differences between NU-6 versions were small with the YNH listeners (2–3 dB) and somewhat larger with the OHL listeners (7–9 dB). At the highest presentation levels, recognition performances on the two versions were essentially the same within each group of listeners. At the three lowest presentation levels, recognition performances were better on the Auditec version for essentially all YNH and OHL listeners. The mean slopes of the individual listener functions were 5.2%/dB (Auditec, $SD = 0.8\%/dB$) and 5.3%/dB (VA, $SD = 0.5\%/dB$) for the YNH listeners and 3.3%/dB (Auditec, $SD = 1.1\%/dB$) and 3.5%/dB (VA, $SD = 1.0\%/dB$) for the OHL listeners, which were about 0.5%/dB steeper than the slopes of the respective mean functions. Slight, unspecified calibration and presentation level differences were offered as possible contributors to the differences between the two versions of NU-6, but there are a multitude of other variables that probably also contribute. Recognition performances on most of the words were better with the Auditec version but there was a minority of words that produced better performances on the VA version. Subjectively, 91.7% of the YNH listeners and 83.3% of the OHL listeners indicated that the Auditec version was easier to understand than the VA version. Clinically, at the typical 30–40-dB SL presentation levels used to administer word-recognition tests, the results produced by the Auditec and VA versions of NU-6 are the same for listeners like those included in the present study. Although this article is focused on word recognition in a quiet listening environment, a thorough insight into the ability of a patient with hearing impairment to understand speech must include a measure of speech-recognition performance on a degraded speech task like speech-in-noise.

Acknowledgments. The author thanks Sherri Smith for the administrative guidance she provided at the Mountain Home VA Medical Center. Appreciation is expressed to Auditec of St. Louis and Bill Carver for the Auditec materials used in the study.

REFERENCES

- Adobe Systems, Inc. (2012) *Adobe Audition CS6*. San Jose, CA.
- American National Standards Institute (ANSI). (1989) *Specification for Audiometers*. ANSI S3.6-1989. New York, NY: ANSI.
- American National Standards Institute (ANSI). (2010) *Specification for Audiometers*. ANSI S3.6-2010. New York, NY: ANSI.
- American Standards Association (ASA). (1954) *American Recommended Practice for Volume Measurements of Electrical Speech and Program Waves*. C16.5-1954. New York, NY: ASA.
- Beattie RC, Edgerton BJ, Svihovec DV. (1977) A comparison of the Auditec of St. Louis cassette recordings of NU-6 and CID W-22 on a normal-hearing population. *J Speech Hear Disord* 42(1):60–64.

- Bilger RC. (1984) Speech recognition test development. *Asha Rep* 14:2–15.
- Bladon RAW, Henton CG, Pickering JB. (1984) Towards an auditory theory of speaker normalization. *Lang Commun* 4(1): 59–69.
- Brandy WT. (1966) Reliability of voice tests of speech discrimination. *J Speech Hear Res* 9(3):461–465.
- Campbell RA. (1965) Discrimination test word difficulty. *J Speech Hear Res* 8(1):13–22.
- Causesy GD, Hermanson CL, Hood LJ, Bowling LS. (1983) A comparative evaluation of the Maryland NU 6 auditory test. *J Speech Hear Disord* 48(1):62–69.
- Chinn HA, Gannett DK, Morris RM. (1940) A new standard volume indicator and reference level. *Bell Labs Tech J* 19(1):94–137.
- Davis H. (1947) Tests of hearing. In: Davis H, ed. *Hearing and Deafness: A Guide for Laymen*. New York, NY: Rinehart Books, 125–160.
- DeBow A, Green WB. (2000) A survey of Canadian audiological practices: pure tone and speech audiometry. *J Speech Lang Audiol* 24(4):153–161.
- Department of Veterans Affairs. (2010) *Speech Recognition and Identification Materials. Disc 4.0*. Mountain Home, TN: VA Medical Center.
- Egan JP. (1948) Articulation testing methods. *Laryngoscope* 58(9): 955–991.
- Egan JP. (1957) Remarks on rare PB words. *J Acoust Soc Am* 29(6):751.
- Elpern B. (1961) The relative stability of half list and full list discrimination tests. *Laryngoscope* 71(1):30–36.
- Grose JA, Mamo SK, Buss E, Hall JW. (2015) Temporal processing deficits in middle age. *Am J Audiol* 24(2):91–93.
- Grubb PA. (1963) Considerations in the use of half-list speech discrimination tests. *J Speech Hear Res* 6(3):294–297.
- Hirsh IJ, Davis H, Silverman SR, Reynolds EG, Eldert E, Benson RW. (1952) Development of materials for speech audiometry. *J Speech Hear Disord* 17(3):321–337.
- House AS. (1961) On vowel duration in English. *J Acoust Soc Am* 33(9):1174–1178.
- House AS, Fairbanks G. (1953) The influence of consonant environment upon the secondary acoustical characteristics of vowels. *J Acoust Soc Am* 25(1):105–113.
- Hurley RM, Sells JP. (2003) An abbreviated word recognition protocol based on item difficulty. *Ear Hear* 24(2):111–118.
- Jacewicz E, Fox RA. (2008) Amplitude variations in coarticulated vowels. *J Acoust Soc Am* 123(5):2750–2768.
- Johnson K. (2005) Speaker normalization in speech perception. In: Pisoni DB, Remez RE, eds. *The Handbook of Speech Perception*. Malden, MA: Blackwell Publishing, 363–389.
- Lehiste I, Peterson GE. (1959a) Linguistic considerations in the study of speech intelligibility. *J Acoust Soc Am* 31(3):280–286.
- Lehiste I, Peterson GE. (1959b) Vowel amplitude and phonemic stress in American english. *J Acoust Soc Am* 31(4):428–435.
- Levitt H, Bricker PD. (1970) Reduction of observer bias in reading speech levels with a VU meter. *J Acoust Soc Am* 47(6B): 1583–1587.
- Lobdell BE, Allen JB. (2007) A model of the VU (volume-unit) meter, with speech applications. *J Acoust Soc Am* 121(1):279–285.
- Kreul EJ, Bell DW, Nixon JC. (1969) Factors affecting speech discrimination test difficulty. *J Speech Hear Res* 12(2):281–287.
- Kreul EJ, Nixon JC, Kryter KD, Bell DW, Lang JS, Schubert ED. (1968) A proposed clinical test of speech discrimination. *J Speech Hear Res* 11(3):536–552.
- Margolis RH, Glasberg BR, Creeke S, Moore BC. (2010) AMTAS®: automated method for testing auditory sensitivity: validation studies. *Int J Audiol* 49(3):185–194.
- Margolis RH, Millin JP. (1971) An item-difficulty based speech discrimination test. *J Speech Hear Res* 14(4):865–873.
- Margolis RH, Saly GL, Le C, Laurence J. (2007) Qualind™: a method for assessing the accuracy of automated tests. *J Am Acad Audiol* 18(1):78–89.
- Martin FN, Champlin CA, Chambers JA. (1998) Seventh survey of audiometric practices in the United States. *J Am Acad Audiol* 9(2):95–104.
- Mullennix JM, Pisoni DB, Martin CS. (1989) Some effects of talker variability on spoken word recognition. *J Acoust Soc Am* 85(1): 365–378.
- Peterson GE, Lehiste I. (1962) Revised CNC lists for auditory tests. *J Speech Hear Disord* 27(1):62–70.
- Pisoni DB. (1997) Some thoughts on “normalization” in speech perception. In: Johnson K, Mullennix JW, eds. *Talker Variability in Speech Processing*. San Diego, CA: Academic Press, 9–32.
- Resnick DM. (1962) Reliability of the twenty-five word phonetically balanced lists. *J Aud Res* 2(1):5–12.
- Robjohns H. (2013) Why do waveforms sometimes look lop-sided? *Sound on Sound* 28(7):224.
- Sommers MS, Nygaard LC, Pisoni DB. (1994) Stimulus variability and spoken word recognition: I. Effects of variability in speaking rate and overall amplitude. *J Acoust Soc Am* 96(3):1314–1324.
- Stevens SS, Egan JP, Miller GA. (1947) Methods of measuring speech spectra. *J Acoust Soc Am* 19(5):771–780.
- Stoppenbach DT, Craig JM, Wiley TL, Wilson RH. (1999) Word recognition performance for northwestern university auditory test No. 6 word lists in quiet. *J Am Acad Audiol* 10(8):429–435.
- Thorndike EL, Lorge I. (1944) *The Teacher’s Word Book of 30,000 Words*. New York, NY: Teachers College, Columbia University.
- Thornton A, Raffin MJM. (1978) Speech discrimination scores modeled as a binomial variable. *J Speech Hear Res* 21(3): 507–518.
- Tillman TW, Carhart R. (1966) An Expanded Test for Speech Discrimination Utilizing CNC Monosyllabic Words. Northwestern University Auditory Test No. 6. Brooks Air Force Base, TX USAF School of Aerospace Medicine Technical Report No. SAM-TR-66-55.
- Tillman TW, Carhart R, Wilber LA. (1963) A Test for Speech Discrimination Composed of CNC Monosyllabic Words (N. U.

Auditory Test No. 4). Brooks Air Force Base, TX USAF School of Aerospace Medicine Technical Report No. SAM-TDR-135.

Wiley TL, Stoppenbach DT, Feldhake LJ, Moss KA, Thordardottir ET. (1995) Audiologic practices: what is popular versus what is supported by evidence. *Am J Audiol* 4(1):26–34.

Wilson RH. (2015) Amplitude (vu and rms) and temporal (ms) measures of two northwestern university auditory test No. 6 recordings. *J Am Acad Audiol* 26(4):346–354.

Wilson RH, Coley KE, Haenel JL, Browning KM. (1976) Northwestern university auditory test No. 6: normative and comparative intelligibility functions. *J Am Audiol Soc* 1(5):221–228.

Wilson RH, Margolis RH. (1983) Measurements of auditory thresholds for speech stimuli. In: Konkle DF, Rintelmann WF, eds. *Principles of Speech Audiometry*. Baltimore, MD: University Park Press, 79–126.

Wilson RH, McArdle R. (2013) Characteristics of the audiometric 4,000 Hz notch (744,553 veterans) and the 3,000, 4,000, and 6,000

Hz notches (539,932 veterans). *J Rehab Res Develop* 50(1): 111–132.

Wilson RH, McArdle R. (2015) The homogeneity with respect to intelligibility of recorded word-recognition materials. *J Am Acad Audiol* 26(4):331–345.

Wilson RH, McArdle R, Roberts H. (2008) A comparison of recognition performances in speech-spectrum noise by listeners with normal hearing on PB-50, CID W-22, NU—6, W-1 spondaic words, and monosyllabic digits spoken by the same speaker. *J Am Acad Audiol* 19(6):496–506.

Wilson RH, Oyler AL. (1997) Psychometric functions for the CID W-22 and NU Auditory Test No. 6. Materials spoken by the same speaker. *Ear Hear* 18(5):430–433.

Wilson RH, Zizz CA, Shanks J, Causey GD. (1990) Normative data in quiet, broadband noise, and competing message for Northwestern University Auditory Test No. 6 by a female speaker. *J Speech Hear Disord* 55(4):771–778.