J Am Acad Audiol 18:813–825 (2007)

Intra- and Inter-session Test, Retest Reliability of the Words-in-Noise (WIN) Test

Richard H. Wilson*†
Rachel McArdle‡§

Abstract

Retest stability and retest reliability were assessed for the Words-in-Noise Test (WIN) in two experiments involving older listeners with sensorineural hearing loss. In Experiment 1, the 70-item WIN protocol was administered during two sessions 12 months apart to examine retest stability on a sample of 315 veterans from four VA Medical Centers. The mean 50% points on the WIN were 12.5- and 12.8-dB S/N for the two sessions with a critical difference of 3.5 dB and an intra-class correlation coefficient of 0.88. [Normal recognition performance on the WIN (50% point) is ≤6-dB S/N.] In Experiment 2, intra- and inter-session retest reliability was examined for the two 35-word WIN protocols on 96 veterans, 48 of whom had mild-to-severe hearing loss (Group 1) and 48 of whom had a moderate-to-severe hearing loss (Group 2). The mean 50% points on the WIN during the two sessions (separated by 40 days) were 13.0- and 13.4-dB S/N (Group 1) and 15.3- and 15.8-dB S/N (Group 2) with no significant intra-session differences. A 3.1-dB critical difference was calculated for the groups combined with intra-class correlations of 0.89 and 0.91 for Group 1 and Group 2, respectively.

Key Words: Auditory perception, hearing loss, retest reliability, retest stability, speech perception, word recognition in multitalker babble

Abbreviations: ANSI = American National Standards Institute; NU No. 6 = Northwestern University Auditory Test No. 6; SNR, S/N = signal-to-noise ratio; WIN = Words-in-Noise Test

Sumario

La estabilidad y la confiabilidad para el re-test se evaluaron para la Prueba de Palabras en Ruido (WIN) en dos experimentos que involucraron sujetos con hipoacusia sensorineural. En el experimento 1, se administró el protocolo WIN de 70 ítems en dos sesiones con 12 meses de separación, para examinar la estabilidad del re-test en una muestra de 315 veteranos de

*James H. Quillen VA Medical Center, Mountain Home, Tennessee; †Departments of Surgery and Communicative Disorders, East Tennessee State University, Johnson City, Tennessee; ‡VA Healthcare System, Bay Pines, Florida; §Department of Communication Sciences and Disorders, University of South Florida, Tampa, Florida

Richard H. Wilson, Ph.D., VA Medical Center, Audiology (126), Mountain Home, TN 37684; Phone: 423-979-3561; Fax: 423-979-3403; E-mail: RICHARD.WILSON2@VA.GOV

The Rehabilitation Research and Development Service, Department of Veterans Affairs supported this work through a Merit Review, the Auditory and Vestibular Dysfunction Research Enhancement Award Program (REAP), a Senior Research Career Scientist award to the first author, and a Research Career Development award to the second author.

nuestros Centros Médicos del VA. Los puntos medios 50% del WIN fueron 12.5 y 12.8 dB S/N para las dos sesiones, con una diferencia crítica de 3.5 dB y un coeficiente de correlación de intra-clase de 0.88. [El desempeño normal de reconocimiento en el WIN (punto 50%) es ≤ 6-dB S/N.] En el experimento 2, se examinó la confiabilidad del re-test tanto intra- como inter-sesiones para los 2 protocolos de 35 palabras del WIN, en 96 veteranos, 48 de los cuáles tenían una hipoacusia leve a severa (Grupo 1), y 48 tenían una hipoacusia moderada a severa (Grupo 2). Los puntos medios 50% del WIN durante las dos sesiones (separadas por 40 días) fueron de 13.0 y 13.4 dB S/N (Grupo 1) y 15.3 y 15.8 dB S/N (Grupo 2) sin diferencias significativas intra-sesión. Se calculó una diferencia crítica de 3.1 dB para los grupos, combinada con correlaciones intra-clase de 0.89 y 0.91 para el Grupo 1 y para el Grupo 2, respectivamente.

Palabras Clave: Percepción auditiva, pérdida auditiva, confiabilidad del re-test, estabilidad del re-test, percepción del lenguaje, reconocimiento de palabras en balbuceo de hablantes múltiples

Abreviaturas: ANSI = Instituto Nacional Americano de Estándares; NU No. 6: Prueba Auditiva No. 6 de la Universidad Northwestern; SNR, S/N = tasa señal-ruido; WIN = Prueba de Palabras en Ruido

The importance of evaluating the ability of listeners to understand speech in background noise has been emphasized for a number of years. For example, Carhart and Tillman (1970) observed that the level of noise that was "mildly disruptive" to the listener with normal hearing was in fact a "serious masker" for the listener with sensorineural hearing loss. Based on their experimental data, Carhart and Tillman suggested that defining hearing loss in terms of pure-tone thresholds and speech-recognition abilities in quiet was inadequate and that a measure of speech-in-noise should be added to the routine audiologic evaluation. Over the years, the issue of evaluating the ability of listeners to understand speech in background noise has been addressed by many investigators including Plomp (1978), Plomp and Duquesnoy (1982), Cox et al. (1987), Beattie (1989), Nilsson et al. (1994), and more recently Killion (2002) and Killion et al. (2004). A recent report by Strom (2006) indicated that less than half of the audiologists responding to a survey used a measure of speech-in-noise in their audiologic evaluations. Although the majority of audiologists are not using a speech-in-noise measure, the survey indicates that speech-in-noise testing is gradually becoming a recognized component of the audiologic evaluation.

The two traditional components of an audiologic evaluation, pure-tone thresholds and speech recognition in quiet, typically are expressed in decibels hearing loss and percent correct recognition, respectively. The various measures of speech-in-noise abilities like the Hearing in Noise Test (HINT, Nilsson et al., 1994), the QuickSIN (Killion et al, 2004), and the Words-in-Noise (WIN) are quantified in terms of the signal-to-noise ratio (SNR) at which 50% of the test items are correct, i.e., a SNR loss or a SNR hearing loss. Unfortunately, with one exception, the SNR hearing loss can not be predicted with any degree of certainty either from the pure-tone thresholds or from the speech-recognition score in quiet. The one exception is if the speech-recognition ability in quiet is poor, then poor speech-recognition in

noise is assured. As Killion (2002) indicated, if you want to know how a listener understands speech in background noise, then you must test that ability.

The WIN test was developed as an instrument that quantifies the ability of listeners to understand speech in background multitalker babble (Wilson, 2003). The original WIN paradigm involved the presentation of 10 words at each of 7 SNRs from 24 to 0 dB in 4-dB decrements. Subsequently for clinic use, the 70-word list was divided into two 35-word lists in which 5 words are presented at each of the 7 SNRs. The WIN uses a modified method of constants to establish the SNR at which 50% correct performance is achieved on the materials. The 50% point is computed with the Spearman-Kärber equation (Finney, 1952; Wilson et al. 1973). The 90th percentile on the WIN for young listeners with normal hearing is 6-dB S/N (Wilson et al., 2003), which is used to define the upper boundary of normal performance. Several studies from our laboratory indicate that listeners with high-frequency, sensorineural hearing loss typically have 50% points in the 10- to 16-dB S/N range, which translates to a 4- to 10-dB SNR hearing loss (McArdle et al., 2005; Wilson and Burks, 2005). The degree of hearing loss influences recognition performance on the WIN much more so than does the age of the listener (Wilson and Weakley, 2005). The purpose of this report was to examine both the short- and long-term test, retest characteristics of the WIN mainly to identify critical differences that can be used for identifying a true change in performance by an individual listener. Short-term test, retest differences in performance reflect, in part, measurement error and thus measure retest reliability whereas long-term fluctuations in performance can result from the passage of time in addition to measurement error thus reflecting retest stability (Demorest & Erdman, 1988). For the 70-word version of the WIN, retest stability was examined in two sessions over a 12 month period. For the two 35-word versions, retest reliability was examined within a session and between two sessions separated by 2-3 months. The 70- and 35-word versions of the WIN were examined because both versions can be used in the laboratory and in the clinic.

EXPERIMENT 1— RETEST STABILITY

Methods

Participants

Pure-tone thresholds, word-recognition data in quiet and WIN data from 315 listeners enrolled in a VA multi-center study (Abrams and Doyle., 2000) were evaluated. The mean age of the 311 males and 4 females with sensorineural hearing loss was 69.7 years old (SD = 7 years). Of the 315 listeners, >95% were Caucasian and >75% had a high school education. English was the first language for all listeners.

Multiple dependent *t*-tests were used to examine changes in pure-tone, air-conduction thresholds over the 12-month study period and to examine differences between right and left ear sensitivity. Five of the 16 right-ear and left-ear pure-tone thresholds decreased significantly between Session 1 and Session 2. The five differences, which were <1 dB or <20% of the 5-dB measuring interval, were not considered of any clinical significance. Only at 2000 Hz, was there a significant difference between ears with the mean thresholds being 51-dB HL (ANSI, 2004) and 48.4-dB HL for the right and left ears, respectively. This 2.6-dB difference is about half of the 5-dB measurement interval and is not considered noteworthy. For practical purposes, then, there were no differences between the pure-tone thresholds obtained in Session 1 and Session 2 or between the two ears. For these reasons, only the mean audiogram and standard deviations for the right ear obtained during Session 1 are shown in Figure 1. The hearing loss is characterized as a mild-to-severe, gently sloping sensorineural hearing loss.

Materials

In addition to the WIN materials, traditional speech-recognition measures in quiet

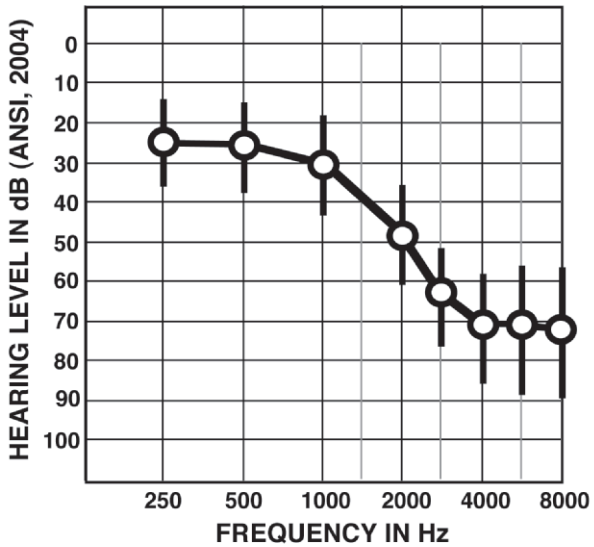

Figure 1. Mean audiogram for the right ears of the 315 participants. The vertical bars indicate ±1 standard deviation.

were made using Lists 1 and 2 of Northwestern University Auditory Test No. 6 (NU No. 6; Tillman and Carhart, 1966) spoken by a female speaker (Department of Veterans Affairs, 1998). Each 50-word list was ~4 minutes long. The WIN test had the following characteristics (Wilson, 2003): (1) 70 monosyllabic words from the same recorded version of NU No. 6 used for the recognition measures in quiet, which enabled the evaluation of recognition performance in quiet and in multitalker babble with the same materials spoken by the same speaker, (2) 10 unique words presented at each of 7 signal-to-noise ratios calibrated on a vu meter in 4-dB decrements from 24-dB S/N to 0-dB S/N, (3) each word was time-locked to a unique segment of multitalker babble, which reduced variability, (4) the level of the babble, which was presented continuously, was fixed and the level of the words varied, (5) the interval between words was 2.7 s, (6) the 50% point was quantifiable with the Spearman-Kärber equation, and (7) a stopping rule that terminated the test sequence when the 10 words at one level were incorrect.

The multitalker babble was recorded by Causey (pers. comm., 1979) and consisted of three females and three males talking simultaneously about different topics (Sperry et al, 1997). Because 4.2-s segments of babble were compiled randomly in the test sequence, the babble was not intelligible. The word/babble segments were edited at the negative going zero crossings, which avoided clicks at the segment boundaries when the word/babble segments were concatenated to form the test lists. The NU No. 6 lists in quiet and the WIN lists were recorded on an audio compact disc (Hewlett Packard, Model DVD200i).

Procedures

Pure-tone thresholds (Carhart and Jerger, 1959) and word-recognition performances in quiet and performance on the WIN were measured for each participant during Session 1 and 12 months later during Session 2. For word-recognition performance in quiet, Lists 1 and 2 of the NU No. 6 (Department of Veterans Affairs, 1998) were presented at 60- and 80-dB HL, which were the levels that corresponded to the presentation levels of the words in the WIN paradigm at 0- and 20-dB S/N, respectively. Odd-numbered participants listened to List 1 at 80-dB HL followed by List 2 at 60-dB HL. The even-numbered participants listened to List 1 at 60-dB HL followed by List 2 at 80-dB HL. Following the two lists in quiet, each participant was presented the 70-word WIN list. The WIN words were presented in 4-dB decrements from 84- to 60-dB HL (24- to 0-dB S/N) with the level of the multitalker babble fixed at 60-dB HL. As per the protocol in the multi-center study, the presentation of both the NU No. 6 in quiet and the WIN materials was binaural under earphones in a double-wall sound booth. Previous data indicate that in a homophasic noise paradigm there is only a 0.5-dB binaural advantage over performance in the monaural better ear (Holma et al., 1997; Wilson, 2003).

Results and Discussion

Means and standard deviations for Session 1 and Session 2 of Experiment 1 are listed in Table 1. The mean 50% points for Sessions 1 and 2 were 12.5- and 12.8-dB S/N, respectively, with 3.6 dB standard deviations. A paired-sample *t*-test revealed that the 0.3-dB difference between the two means measured in Sessions 1 and 2 was significant (t [314] = -2.23, $p<.05$), suggesting that the difference was statistically reliable given a large n of 315 participants. The mean 50% points at 12.5- to 12.8-dB S/N are in good agreement with Wilson and Weakley (2004, Table 6) who reported a mean 50% point at 12.2-dB S/N for a group of 48 listeners with

Table 1. The mean data (and standard deviations) from Experiment 1 (n = 315) for the 50% points for the WIN, the two pure-tone averages, and the two speech in quiet presentation levels are shown for Session 1 and Session 2, which were separated by a 12-month hiatus.

Condition	Units	Session 1		Session 2	
WIN	dB S/N	12.5	(3.6)	12.8	(3.6)
PTA (500, 1000, 2000 Hz)	dB HL	35.2	(9.4)	35.9	(9.9)
PTA (1000, 2000, 4000 Hz)	dB HL	50.2	(9.1)	50.2	(9.7)
Speech in quiet (60-dB HL)	% Correct	69.0	(20.6)	66.0	(23.1)
Speech in quiet (80-dB HL)	% Correct	84.1	(13.8)	84.2	(14.8)

similar hearing loss (mean age = 63.5). Other studies with the WIN have observed similar results (Wilson and Weakley, 2005; McArdle et al., 2005; Wilson et al., 2007a,b).

A critical difference of 3.5 dB was calculated using the standard error of measurement (1.225) to find the standard error of the difference (1.73) and the 95% confidence interval for a true change. The critical difference suggests that the score of an individual from one test administration to another must change by 3.5 dB S/N to be concluded with 95% confidence that a true change has occurred between sessions.

The bivariate plot presented in Figure 2 gives a graphic analysis of the differences between the individual 50% points on the WIN that were calculated with the Spearman-Kärber equation during Session 1 (abscissa) and Session 2 (ordinate). The diagonal line represents equal performance between sessions with the means represented by the large filled circle. The numbers in parentheses indicate the percentage of 50% points that were above, on, and below the diagonal line. A little over half of the 315 listeners (52.4%) had better performance on the WIN in Session 2 than in Session 1; 39.7% of the listeners had better performance in Session 1 than Session 2 with the remaining listeners (7.9%) having equal performances in the two sessions. The individual data presented in this manner provide insight into the mean data that demonstrated little practical difference between the 50% points on the WIN obtained in the two sessions. The intra-class correlation coefficient for the WIN was 0.88, which suggests acceptable retest stability (Nunally and Bernstein, 1994).

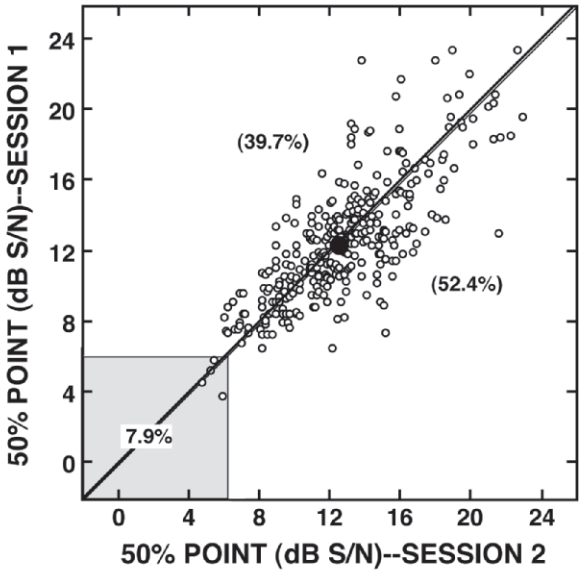

Figure 2. A bivariate plot of the 50% points calculated on the data for each participant with the Spearman-Kärber equation during Session 1 (abscissa) and Session 2 (ordinate). To minimize superimposed datum points, a multiplicative algorithm that randomly multiplied the 50% points by 0.975 to 1.025 in 0.005 increments was used to jitter the data. The numbers in parentheses indicate the percentage of the 50% points from the 315 listeners above, on, and below the diagonal line, which represents equal performance. The large, filled symbol represents the mean datum points and the shaded region defines performance by listeners with normal hearing (Wilson et al., 2003).

The psychometric performance functions for the WIN materials obtained during Sessions 1 and 2 are shown in Figure 3 as circles and squares, respectively. The vertical lines represent ±1 standard deviation and the line through the datum points is the best-fit, third-degree polynomial used to describe the data. Obviously, mean performance on the WIN materials during Sessions 1 and 2 were the same. The 50% points calculated from the polynomial equations and the slopes of the functions at the 50% points respectively were 11.6-dB S/N and 6.6%/dB (Session 1) and 11.8-dB S/N and 6.6%/dB (Session 2). For comparison with the WIN results, the recognition performances on the NU No. 6 materials presented at 60- and 80-dB HL in quiet also are shown as the filled symbols in Figure 3. Two relations between word-recognition performances in quiet and in multitalker babble are apparent. First, at the 80-dB HL presentation level performances in quiet and in multitalker babble were the same, which suggests that the babble did not interfere with

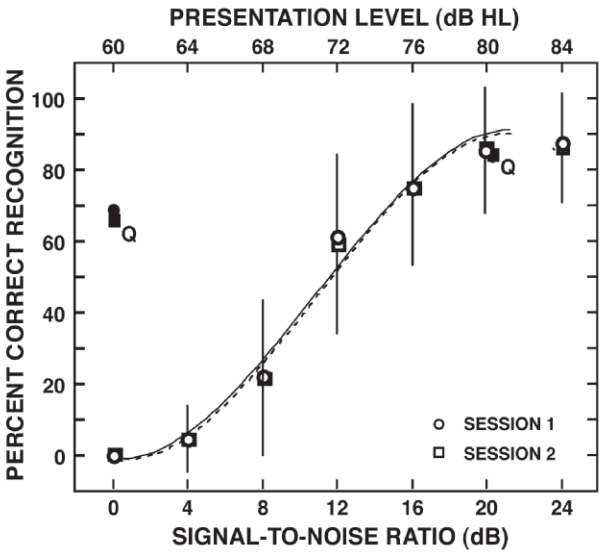

Figure 3. The psychometric functions for the WIN data obtained during Session 1 (circles) and during Session 2 (squares). The vertical lines indicate the standard deviations for the data from Session 1 and the lines through the datum points are the best-fit, third-degree polynomials used to describe the data. The filled symbols depict the data obtained with the NU No. 6 lists presented in quiet (Q) at 60- and 80-dB HL.

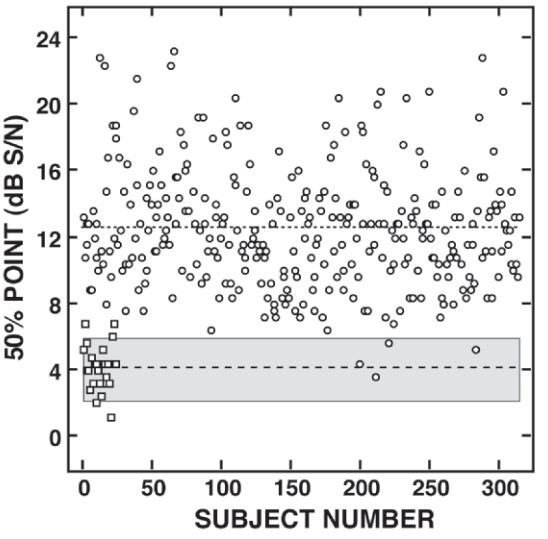

Figure 4. The 50% points calculated with the Spearman-Kärber equation for the individual WIN data obtained from the 315 listeners with hearing loss during Session 1 (circles). The squares depict similar data obtained from 24 listeners with normal hearing (Wilson et al., 2003). The shaded region defines normal performance on the WIN task. The dotted line represents the mean performance by the listeners with hearing loss and the dashed line represents the mean performance by the listeners with normal hearing.

speech recognition at 20-dB S/N. Second, although at 60-dB HL the recognition performance in quiet dropped slightly (15-18%) as compared to 80-dB HL (i.e., from 84% to 66-69%), recognition performance on the WIN dropped 84% over this same range (i.e., from 84% to 0%). The dramatic difference between the performances in quiet and in babble at 60-dB HL can be attributed almost entirely to the increased difficulty of the listening task that was introduced by the competing multitalker babble. Audibility was not the major contributor to the reduced performance obtained at the 60-dB HL presentation level in babble.

To illustrate the variability in recognition performance on the WIN, individual datum points from Session 1 are plotted in Figure 4 for each of the 315 participants in the current study (open circles) as well as individual 50% points for 24 listeners with normal hearing (open squares) reported by Wilson et al (2003). The shaded area represents the 10th to 90th percentile recognition performances for the listeners with normal hearing and the dotted line represents the mean performance of the 315 listeners with hearing loss. The following two points are apparent from the data in the figure: (1) the data are not very homogeneous with a random distribution among the listeners with hearing loss, and (2) only four listeners with hearing loss were within the (shaded) range of normal performance. The distributions of the WIN performances at the 50% point are detailed further in Table 2 in 3.6-dB S/N steps, which are equivalent to 10 words in the 70-word paradigm. The distributions for both sessions are essentially identical and can be characterized as normal. Although almost all of the listeners with (sensitivity) hearing loss have difficulty understanding speech in background noise, an impressive aspect of the data in Table 2 are the number of listeners (93 or 30%) whose 50% point on the WIN were >14-dB S/N, which translates into a signal-to-noise hearing loss of >8-dB (i.e., 14-dB S/N minus 6-dB S/N, which is

Table 2. The distributions (%) of the 50% points on the WIN calculated with the Spearman-Kärber equation are shown for the 315 listeners with hearing loss in Experiment 1. Data from both sessions are listed.

Range (dB S/N)	Session 1	Session 2
≤6.1	1.6	2.2
6.5 - 10.1	25.1	24.4
10.5 - 14.1	47.3	47.3
14.5 - 18.1	17.5	17.5
18.5 - 22.1	7.0	7.0
22.5 - 26.1	1.6	1.6

the 90th percentile on the WIN for listeners with normal hearing). As has been pointed out by others (e.g., Plomp and Duquesnoy, 1982), an 8-dB S/N hearing loss is a substantial impairment in communication.

Speech-recognition performance in quiet at 60- and 80-dB HL also was measured during both sessions. At 60-dB HL in quiet, which corresponded to the 0-dB S/N on the WIN, mean correct recognition performances of 69% (SD = 20.6%) and 66% (SD = 23.1%) were obtained during Session 1 and Session 2, respectively. At 80-dB HL in quiet, which corresponded to 20-dB S/N on the WIN, mean recognition performances of 84% were obtained in both sessions with standard deviations of 13.8% (Session 1) and 14.8% (Session 2). Figure 5 is a two-panel, bivariate plot of the individual word-recognition scores in quiet obtained in Session 1 (ordinate) and in Session 2 (abscissa). The data for the 60- and 80-dB HL conditions are shown in the top and bottom panels, respectively. The diagonal line in each panel represents equal performance on the two trials with the large filled symbols depicting the mean data. It is obvious from the figure that the individual recognition performances at 60-dB HL were worse than the performances at 80-dB HL with more inter-subject variability, which follows the well established principle that the closer the data are to the 50% point on the psychometric function, the larger the variability will be (Thornton and Raffin, 1978). For both presentation levels, the distributions of the data from Session 1 and Session 2 around the diagonal line are similar. The data in Figure 5 show for the 60-dB HL condition that 123 (39.0%) of the listeners performed better on Session 2, 165 (52.4%) performed better on Session 1, and 27 (8.6%) of the listeners had equal performances. The same evaluation of the data obtained at 80-dB HL indicates a similar distribution.

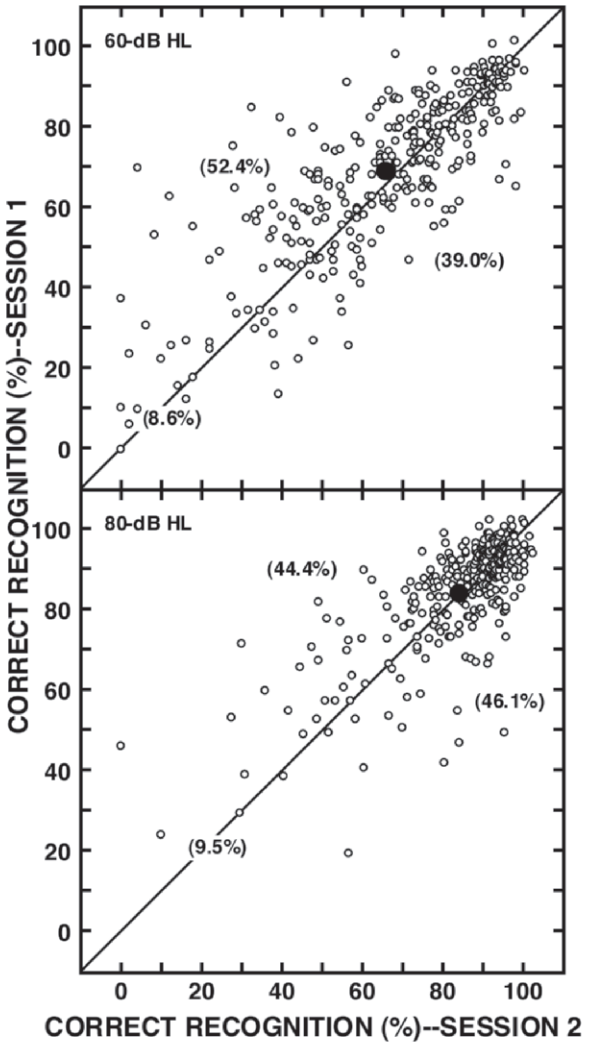

Figure 5. The percent correct recognition obtained by the 315 listeners on the NU No. 6 presented in quiet at 60- (top panel) and 80-dB HL (bottom panel). The datum points were multiplied by a random algorithm that used 0.975 to 1.025 in 0.005 increments to jitter the data. The numbers in parentheses indicate the percentage of recognition performances from the 315 listeners above, on, and below the diagonal line, which represents equal performance.

EXPERIMENT 2— RETEST RELIABILITY

Methods

In this experiment, the test, retest characteristics of the two, 35-word WIN lists (Wilson and Burks, 2005) were evaluated both within each of two sessions and between the two sessions. Two groups of 48 listeners each with sensorineural hearing loss were studied. Data from previous WIN studies on listeners with specified degrees of hearing loss indicate that the mean results and associated variability stabilize when the group size exceeds 36 listeners (Wilson and Burks, 2005; Wilson et al., 2007). Group 1 had mild-to-severe hearing loss and Group 2 had moderate-to-severe hearing loss. As points of reference, pure-tone thresholds and word-recognition in quiet were evaluated during each of the two sessions.

Participants

The participants were recruited sequentially from the Audiology Clinics at Mountain Home. Based on pure-tone thresholds, two groups of listeners were evaluated. General inclusion criteria for both groups were the following for one ear: (1) a word-recognition score of >28% on NU No. 6 (Department of Veterans Affairs, 2006), (2) audiometric results consistent with a sensorineural hearing loss, i.e., no signs of a conductive or retrocochlear hearing loss, and (3) a return appointment within the subsequent 4 months.

Group 1 pure-tone inclusion criteria included: (1) 500-Hz threshold ≥30-dB HL, (2) 1000-Hz threshold ≥40-dB HL, and (3) a pure-tone average (PTA) at 500, 1000, and 2000 Hz between 10- and 40-dB HL. Because the pure-tone thresholds were the same for the two sessions (±2.5 dB) only the mean audiogram and standard deviations for Session 1 are shown in Figure 6 as open circles with the vertical bar representing 1 standard deviation. The mean PTA in Session 1 was 28.2-dB HL (SD = 9.1 dB) with 81.4% (SD = 14.4%) word recognition in quiet. Group 1 ranged in age from 54 to 86 years with a mean age of 66.9 years (SD = 9.1 years). The two sessions differed by 14 to 89 days with a mean difference of 39.5 days (SD = 14.2 days).

Group 2 pure-tone inclusion criterion was a PTA between 41.6- and 60-dB HL. Because the pure-tone thresholds were the same for the two sessions (±3.0 dB) only the mean audiogram and standard deviations for Session 1 are shown in Figure 6 as filled circles with the vertical bar representing -1 standard deviation. The mean PTA in Session 1 was 45.8-dB HL (SD = 4.1 dB) with 69.0% (SD = 17.0%) word recognition in quiet. Group 2 ranged in age from 52 to 87 years with a mean age of 71.9 years (SD = 10.1 years). The two sessions differed by 21 to 130 days with a mean difference of 42.1 days (SD = 20.8 days).

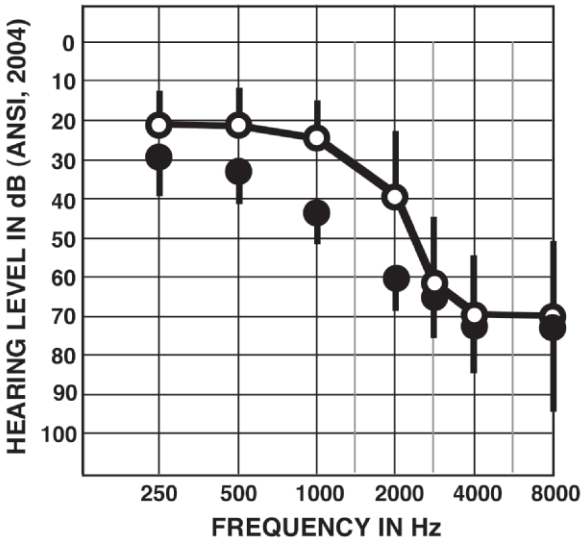

Figure 6. Mean audiogram for the right ears of the 48 listeners with mild-to-severe hearing loss (open circles) and of the 48 listeners with moderate-to-severe hearing loss (filled circles) that were obtained during Session 1. The +1 SD is shown for the former group, whereas the -1 SD is shown for the latter group.

Procedures

Each listener participated in two, 20-30-minute sessions. During each session the following were evaluated: (1) pure-tone thresholds, (2) word-recognition on 25-word NU No. 6 lists presented at two levels separated by 20-dB, and (3) two randomizations of either List 1 or List 2 of the 35-word WIN protocol (Wilson and Burks, 2005). The modified Hughson-Westlake procedure was used for the pure-tone testing (Carhart and Jerger, 1959). List 4 of the NU No. 6 (Department of Veterans Affairs, 2006) was used for word-recognition testing and was presented at 60- and 80-dB HL for Group 1 and 70- and 90-dB HL for Group 2. These two levels, which corresponded to the presentation levels of the words in the 0- and 20-dB S/N conditions of the WIN, helped ensure that audibility was not a major issue compounding the word-recognition performances in multitalker babble. Half the subjects in each group received the first half of List 4 first with the other half of the subjects receiving the second half of List 4 first. The lower presentation level always preceded the higher presentation level. The WIN materials used in this experiment involved the presentation of 5 words at each of 7 signal-to-noise ratios. The level of the multitalker babble was fixed at 80-dB SPL (Group 1) or 90-dB SPL (Group 2) and the level of the words was decremented in 4-dB steps from 24 to 0 dB S/N. Thus, Group 1 listened to the words presented at 104- to 80-dB SPL (84- to 60-dB HL) whereas Group 2 listened to the words presented at 114- to 90-dB SPL (94- to 70-dB HL). A previous study on listeners with mild-to-severe hearing loss demonstrated no differ-

ence between 50% points when the babble was fixed at 70-, 80-, or 90-dB SPL (Wilson, 2003). A stopping rule terminated the protocol when 5 words at one signal-to-noise ratio were missed. With this rule, many listeners only received the first five or six presentation levels. Again, the metric of interest with the WIN was the 50% point calculated with the Spearman-Kärber equation. In each group of 48 listeners, 24 listeners were assigned to List 1 and 24 were assigned to List 2. Within the groups of 24 listeners, 12 received Randomization A of the list followed by Randomization B, and vice versa. In this manner both randomizations of a list were given an equal number of times as Trial 1 and Trial 2 within a session. The list and order of the randomizations were the same in both sessions. With this design test-retest data were obtained both for intra- and inter-session conditions.

The materials were reproduced on a compact disc player (Sony, Model CDP-497) and routed through an audiometer (Grason-Stadler, Model 61) to a TDH-50P earphone encased in a Telephonics P/N 510C017-1 cushion. The non-test ear was covered with a dummy earphone.

Result and Discussion

Means and standard deviations for both Trial 1 and Trial 2 of Session 1 and Session 2 for both hearing loss groups are listed in Table 3. The mean 50% points on the WIN were 13.0- to 13.4-dB S/N (Group 1) and 15.3- to 15.8-dB S/N (Group 2) with standard deviations in the 3- to 4-dB range. To examine the effect of Session, Trial, and List, the data for the each hearing loss group were examined separately using the General Linear Model (GLM) repeated-measures analyses of variance (ANOVA). In each analysis there was one between-group variable (List) and two within-group variables (Session and Trial). The significance level for each of the ANOVAs and post-hoc analyses was $p < .05$. All statistical analyses were performed using SPSS for Windows Version 14.0.

For the analyses of both groups of listeners, none of the main effects or interactions was significant. These findings suggest the following for both hearing loss groups: (1) no significant performance differences on List 1 and List 2 of the WIN, (2) although performance improved slightly from Trial 1 to Trial 2, there was no significant intra-session practice effect, and (3) no significant inter-session performance differences were found. In addition to the lack of finding an effect of Session for either hearing loss group, the intra-class correlation for the test-retest data were 0.89 for the mild hearing loss group and 0.91 for the moderate hearing loss group. Both intra-class correlations suggest acceptable retest reliability (Nunally and Bernstein, 1994). An earlier report observed intra-session test-retest for the 70-

Table 3. The mean data (and standard deviations) from Experiment 2 for the listeners with mild-to-severe hearing loss and the listeners with moderate-to-severe hearing loss for the 50% points for the WIN, the two pure-tone averages, and the two speech in quiet presentation levels are shown for Session 1 and Session 2, which were separated by an average 40-day interval.

Condition	Units	Session 1		Session 2	
Mild-to-severe hearing loss (n = 48)					
WIN Trial 1	dB S/N	13.4	(4.0)	13.2	(3.3)
WIN Trial 2	dB S/N	13.0	(3.5)	13.1	(3.6)
PTA (500, 1000, 2000 Hz)	dB HL	28.2	(9.1)	26.5	(9.6)
PTA (1000, 2000, 4000 Hz)	dB HL	44.3	(9.4)	43.5	(9.9)
Speech in quiet (60-dB HL)	% Correct	72.7	(18.3)	72.7	(20.1)
Speech in quiet (80-dB HL)	% Correct	81.4	(14.4)	82.4	(13.7)
Moderate-to-severe hearing loss (n = 48)					
WIN Trial 1	dB S/N	15.8	(3.2)	15.4	(3.0)
WIN Trial 2	dB S/N	15.5	(3.1)	15.3	(2.9)
PTA (500, 1000, 2000 Hz)	dB HL	45.8	(4.1)	43.2	(5.6)
PTA (1000, 2000, 4000 Hz)	dB HL	58.8	(5.8)	57.1	(6.4)
Speech in quiet (70-dB HL)	% Correct	50.9	(21.4)	49.7	(20.6)
Speech in quiet (90-dB HL)	% Correct	69.0	(17.0)	68.1	(14.3)

word WIN protocol also to be very good (Wilson et al, 2003). In that study, the 24 listeners with normal hearing (mean = 21.1 years) had mean 50% correct points of 4.1- and 4.0-dB S/N for Trials 1 and 2, respectively, whereas 24 listeners with sensorineural hearing loss (mean = 58.5 years) had mean 50% points of 9.4- and 9.1-dB S/N, respectively.

Critical difference values were calculated for the mean 50% points for Lists 1 and 2 as well as for the individual data collapsed across the two groups of listeners. To examine list equivalency the average standard deviation for Lists 1 and 2 among all 96 participants (3.6) was used to calculate the standard error of the mean (0.37). The 95% confidence interval around the across list mean average (14.5-dB S/N) for the 96 participants was 13.8- to 15.2-dB S/N. The mean 50% points for Lists 1 (14.3-dB S/N) and 2 (14.7-dB S/N) were within the 95% confidence interval suggesting equivalence for the two WIN lists. A critical difference value also was calculated for use with the individual data. A critical difference of 3.1 dB was calculated using the standard error of measurement (1.14) to find the standard error of the difference (1.60) and the 95% confidence interval for a true change in the 50% points. The critical difference value suggests that the score of an individual from one test administration to another must change by 3.1-dB S/N to be concluded with 95% confidence that a true change occurred. If List 1 and List 2 were both administered and averaged together, then the 95% critical difference value drops to 2.1 dB since the standard deviation used to calculate the critical difference is divided by the square root of 2 for the 2 lists.

The two-panel bivariate plot presented in Figure 7 gives a graphic analysis of the individual 50% points on the WIN established with the Spearman-Kärber equation for Group 1 (top panel) and Group 2 (bottom panel). In both panels the data are shown for Trial 1 (ordinate) and Trial 2 (abscissa) for both Session 1 (circles) and Session 2 (triangles). The diagonal line represents equal performance with the means represented by the large filled symbols. The numbers in parentheses indicate the percentage of 50% points that were above, on, or below the diagonal line. For both groups of listeners and for both sessions, the data in Figure 7, which demonstrate an even distribution of the individual datum points around the diagonal line with only an occasional outlier, indicate at the level of the individual subject that the two trials on the WIN materials produce the same result. The distributions of the performances at the 50% point are detailed further in Table 4 in 3.6-dB S/N steps, which are equivalent to 5 words in the 35-word paradigm. The distributions for both hearing loss groups are essentially normal with a proportional shift

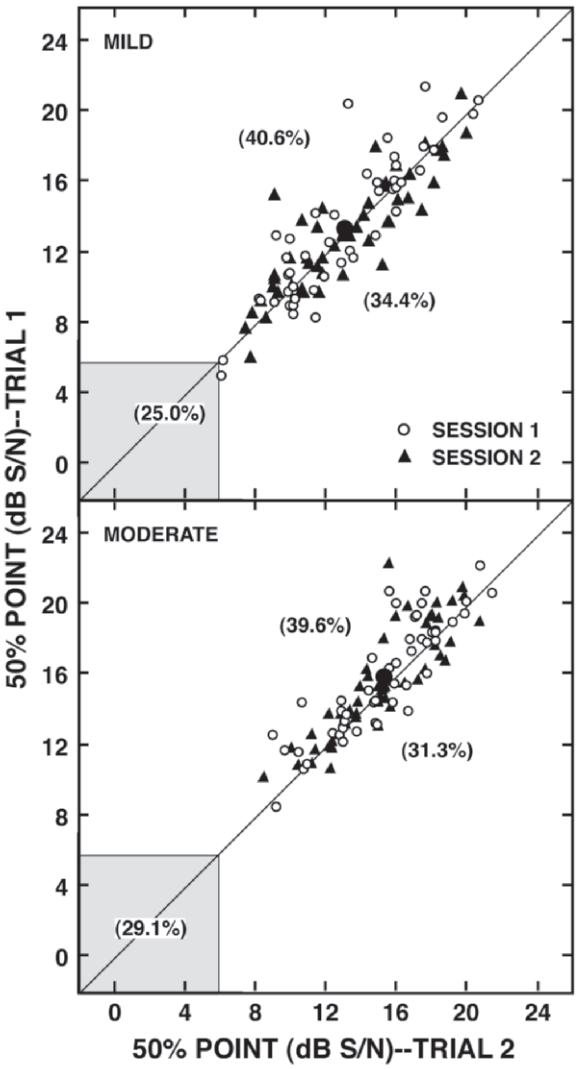

Figure 7. Bivariate plots of the 50% points calculated on the data for each participant with the Spearman-Kärber equation during Trial 1 (abscissa) and Trial 2 (ordinate) in Session 1 (open circles) and Session 2 (filled triangles). Data for the mild-to-severe group are shown in the top panel whereas data for the moderate-to-severe group are shown in the bottom panel. Again, the data were jittered using an algorithm that multiplied randomly each point by 0.975 to 1.025 in 0.005 increments. The numbers in parentheses indicate the percentage of the 50% points above, on, and below the diagonal line, which represents equal performance. The large, filled symbols represent the mean datum points and the shaded region defines performance by listeners with normal hearing (Wilson et al., 2003).

Table 4. The percentage distributions for the 50% points on the WIN are shown for both groups of 48 listeners with hearing loss in Experiment 2 calculated with the Spearman-Kärber equation for each of the two trials in each of the two sessions.

Range (dB S/N)	Session 1 Trial 1	Session 1 Trial 2	Session 2 Trial 1	Session 2 Trial 2
Mild-to-severe hearing loss (n = 48)				
≤6.1	4.2	4.2	2.1	0.0
6.5 - 10.1	22.9	29.2	20.8	25.0
10.5 - 14.1	35.4	29.2	39.6	37.5
14.5 - 18.1	25.0	31.3	33.3	29.2
18.5 - 22.1	12.5	6.3	4.2	8.3
22.5 - 26.1	0.0	0.0	0.0	0.0
Moderate-to-severe hearing loss (n = 48)				
≤6.1	0.0	0.0	0.0	0.0
6.5 - 10.1	2.1	6.3	2.1	4.2
10.5 - 14.1	37.5	27.1	39.6	35.4
14.5 - 18.1	33.3	52.1	33.3	41.7
18.5 - 22.1	27.1	14.6	25.0	18.8
22.5 - 26.1	0.0	0.0	0.0	0.0

to the higher dB S/N values for the moderate hearing loss group. As seen in Table 4 a large number of listeners in both hearing loss groups had 50% points on the WIN >14-dB S/N, which translates into a signal-to-noise hearing loss of >8-dB (i.e., 14-dB S/N minus 6-dB S/N, which is the 90th percentile on the WIN for listeners with normal hearing). As mentioned previously, an 8-dB S/N hearing loss is a substantial impairment in communication (e.g., Plomp and Duquesnoy, 1982).

Because the psychometric functions for each of the four WIN trials (2 sessions by 2 trials) for the respective groups of listeners were essentially identical and to maintain graphic clarity, only the functions for Trial 1 of Session 1 are plotted in Figure 8. The data for Group 1 (mild-to-severe hearing loss) are shown as circles with the solid vertical lines representing ±1 standard deviation, whereas the data for Group 2 (moderate-to-severe hearing loss) are depicted with triangles with the dashed vertical lines illustrating ±1 standard deviation. The lines connecting the datum points are the best-fit, third-degree polynomials that are used to describe the data. The solid symbols (Q) are the mean percent correct recognition for the NU No. 6 words presented in quiet at the levels corresponding to the WIN words presented at the

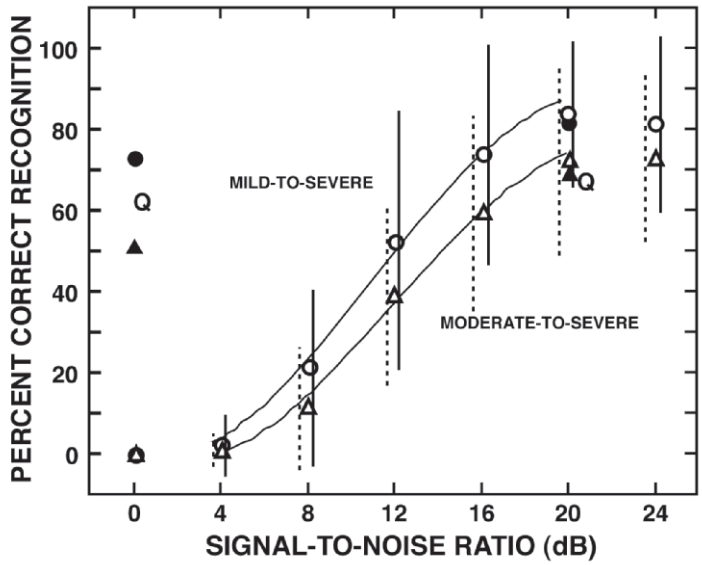

Figure 8. The psychometric functions for the WIN data obtained during Session 1 from the 48 listeners with mild-to-severe hearing loss (circles) and from the 48 listeners with moderate-to-severe hearing loss (triangles) in Experiment 2. The vertical lines indicate the standard deviations for the data from the listeners with mild-to-severe hearing loss (solid) and from the listeners with moderate-to-severe hearing loss (dashed). The lines through the datum points are the best-fit, third-degree polynomials used to describe the data. The filled symbols depict the data obtained with the NU No. 6 lists presented in quiet (Q) at the presentation levels that corresponded to 0- and 20-dB S/N.

two respective SNRs. The data for the quiet conditions in Figure 8 indicate again that recognition performance at the lower presentation level decreased slightly from the performance observed at the higher presentation level. The conclusion from this relation is that the decrease in performance observed on the WIN at the less favorable SNRs was not the result of decreased audibility but rather was the result of the increasing interference created by the multitalker babble.

The mean data and standard deviations for NU No. 6 presented in quiet at the two presentation levels for the two groups of listeners are listed in Table 5. As expected, (1) for both groups recognition performance increased with increased presentation level, and (2) better performance was obtained by Group 1 with the milder hearing loss then by Group 2. The mean differences for corresponding conditions between sessions for each subject group were ≤1%. The standard deviations for corresponding conditions also were similar. These relations between corresponding conditions in the two sessions indicate there is no difference between performances on the NU No. 6 materials at corresponding conditions for the two groups of listeners. At the level of the individual listener, the majority of listeners had differences between corresponding conditions in the two sessions that ranged from 0 to 3 words.

SUMMARY AND CONCLUSIONS

In Experiment 1, the recognition performances of 315 listeners with sensorineural hearing loss on the NU No. 6 in quiet and on the 70-word version of the WIN test were evaluated in two sessions separated by 12 months, which provided a measure of the retest stability of the materials. Performances on NU No. 6 in quiet at 60-dB HL were 69% and 66% correct for the two sessions, whereas at 80-dB HL both mean performances were 84% correct. The mean 50% points on the WIN were 12.5- and 12.8-dB S/N for the two sessions with a critical difference of 3.5 dB and an intra-class correlation coefficient of 0.88.

In Experiment 2, the recognition performances of 48 listeners with mild-to-severe hearing loss (Group 1) and 48 listeners with moderate-to-severe hearing loss (Group 2) on the NU No. 6 in quiet and on the 2, 35-word versions of the WIN were evaluated in two sessions separated by 1 to 3 months. Test, retest measures of the WIN were made in each session, which provided retest reliability data. Essentially identical performances in the two sessions were obtained with the NU No. 6 materials presented in quiet. The mean 50% points on the WIN during the two sessions were 13.0- and 13.4-dB S/N (Group 1) and 15.3- and 15.8-dB S/N (Group 2) with a critical difference of 3.1 dB calculated for the groups combined and intra-class correlations of 0.89 and 0.91 for Group 1 and Group 2, respectively.

The results from both the 70- and 35-word versions of the WIN indicate that for listeners with various degrees of sensorineural hearing loss the WIN provides both a stable and reliable measure of word-recognition performance in background noise. As was mentioned earlier, it is difficult to predict from either pure-tone thresholds or word-recognition performance in quiet the ability of a listener to understand speech in background noise. If you want to know the ability of a listener to understand speech in noise, then that ability must be measured.

Table 5. The mean percent correct recognition and standard deviations obtained during the two sessions are shown for the two presentation levels of each group of listeners.

	Session 1		Session 2	
	60-dB HL	80-dB HL	60-dB HL	80-dB HL
Mild-to-severe hearing loss (n = 48)				
Mean	72.7	81.4	72.7	82.4
SD	18.3	14.4	20.1	13.7
	Session 1		Session 2	
	70-dB HL	90-dB HL	70-dB HL	90-dB HL
Moderate-to-severe hearing loss (n = 48)				
Mean	50.9	69.0	49.7	68.1
SD	21.4	17.0	20.6	14.3

Acknowledgments. We would like to thank Amanda Pillion, Kelly Koder, Sherri Smith, and Joseph Mikolic for their assistance with the data. We would also like to acknowledge the contributions of Paige Harden, Judith Reese and Maureen Wargo.

REFERENCES

Abrams HB, Doyle P. (2000) Functioning, Disability, and Quality of Life in the Hearing Impaired. Merit Review. VA Rehabilitation Research and Development.

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

Beattie RC. (1989) Word recognition functions for the CID W-22 in multitalker noise for normally hearing and hearing-impaired subjects. *J Speech Hear Dis* 54:20–32.

Carhart R, Jerger JF. (1959) Preferred method for clinical determination of pure-tone thresholds. *J Speech Hear Dis* 24:330–345.

Carhart R, Tillman TW. (1970) Interaction of competing speech signals with hearing losses. *Arch Otolaryng* 91:273–279.

Cox RM, Alexander GC, Gilmore C. (1987) Development of the Connected Speech Test (CST). *Ear Hear* 8:119S–125S.

Demorest ME, Erdman SA. (1988) Retest stability of the Communication Profile for the Hearing Impaired. *Ear Hear* 9:237–242.

Department of Veterans Affairs. (1998) *Speech Recognition and Identification Materials*. Disc 2.0. Mountain Home, TN: VA Medical Center.

Department of Veterans Affairs. (2006) *Speech Recognition and Identification Materials*. Disc 4.0. Mountain Home, TN: VA Medical Center.

Finney DJ. (1952) *Statistical Method in Biological Assay*. London: C. Griffen.

Holma T, Laitakari K, Sorri M, Winblad I. (1997) New speech-in-noise test in different types of hearing impairment. *Acta Otolaryng Suppl* 529:71–73.

Killion MC. (2002) New thinking on hearing in noise: a generalized Articulation Index. *Sem Hear* 23:57–75.

Killion MC, Niquette PA, Gudmundsen GI, Revit LJ, Banerjee S. (2004) Development of a quick speech-in-noise test for measuring signal-to-noise ratio loss in normal-hearing and hearing-impaired listeners. *J Acoust Soc Am* 116:2395–2405.

McArdle RA, Wilson RH, Burks CA. (2005) Speech recognition in multitalker babble using digits, words, and sentences. *J Am Acad Audiol* 16:726–739.

Nilsson M, Soli SD, Sullivan JA. (1994) Development of the hearing in noise test for the measurement of speech reception thresholds in quiet and in noise. *J Acoust Soc Am* 95:1085–1099.

Nunnally JC, Bernstein IH. (1994) *Psychometric Theory*. 3rd edition. New York: McGraw-Hill.

Plomp R. (1978) Auditory handicap of hearing impairment and the limited benefit of hearing aids. *J Acoust Soc Am* 63:533–549.

Plomp R, Duquesnoy AJ. (1982) A model for the speech-reception threshold in noise without and with a hearing aid. *Scand Audiol* 15:95–111.

Sperry JL, Wiley TL, Chial MR. (1997) Word recognition performance in various background competitors. *J Am Acad Audiol* 8:71–80.

Strom KE. (2006) The HR 2006 dispenser survey. *Hear Rev* 13:16–39.

Thornton AR, Raffin MJM. (1978) Speech-discrimination scores modeled as a binomial variable. *J Speech Hear Dis* 21:507–518.

Tillman TW, Carhart R. (1966) *An Expanded Test for Speech Discrimination Utilizing CNC Monosyllabic Words. Northwestern University Auditory Test No. 6*. USAF School of Aerospace Medicine Technical Report. Brooks Air Force Base, TX: USAF School of Aerospace Medicine.

Wilson RH. (2003) Development of a speech in multitalker babble paradigm to assess word-recognition performance. *J Am Acad Audiol* 14:453–470.

Wilson RH, Abrams HB, Pillion AL. (2003) A word-recognition task in multitalker babble using a descending presentation mode from 24-dB S/N to 0-dB S/N. *J Rehabil Res Dev* 40:321–328.

Wilson RH, Burks CA. (2005) The Use of 35 words to evaluate hearing loss in terms of signal-to-noise ratio: a clinic protocol. *J Rehabil Res Dev* 42:839–852. http://www.rehab.research.va.gov/jour/05/42/6/pdf/wilson.pdf.

Wilson RH, Carnell C, Trussell A. (2007a) The Words-in-Noise (WIN) test with multitalker babble and speech-spectrum noise maskers. *J Am Acad Audiol* 18:522–529.

Wilson RH, McArdle R, Smith SL. (2007b) An evaluation of the BKB-SIN, HINT, QuickSIN, and WIN materials on listeners with normal hearing and listeners with hearing loss. *J Speech Lang Hear Res* 50:844–856.

Wilson RH, Morgan DE, Dirks DD. (1973) A proposed SRT procedure and its statistical precedent. *J Speech Hear Disord* 38:184–191.

Wilson RH, Weakley DG. (2004) The use of digit triplets to evaluate word-recognition abilities in multitalker babble. *Semin Hear* 25:93–111.

Wilson RH, Weakley DG. (2005) The 500-Hz masking-level difference and word recognition in multitalker babble for 40 to 89 year old listeners with symmetrical sensorineural hearing loss. *J Am Acad Audiol* 16:367–382.