Word Recognition Performance for Northwestern University Auditory Test No. 6 Word Lists in Quiet and in Competing Message

Daniel T. Stoppenbach*
Jeanne M. Craig*
Terry L. Wiley*
Richard H. Wilson'

Abstract

Word recognition norms were determined for the Northwestern University Auditory Test No. 6 (NU-6) released on the Department of Veterans Affairs (VA) Disc 1.1. Word recognition performance (in quiet and in competing message) was measured at 12 presentation levels for 24 young adults with normal hearing. Test–retest reliability also was evaluated. Word recognition scores for VA Disc 1.1 were generally higher than those reported for VA Disc 1.0. The differences in NU-6 scores across disc versions, however, were small and unlikely to affect clinical decisions based on word recognition tests. Score differences on test–retest for the VA Disc 1.1 version of the NU-6 with a competing message background also were small and unlikely to affect clinical outcomes. Overall, based on comparisons of scores for the disc versions of the NU-6, it appears that the two different recordings can be used interchangeably for clinical applications.

Key Words: Articulation functions, performance intensity functions, speech audiology, word recognition

Abbreviations: CD = compact disc, NU-6 = Northwestern University Auditory Test No. 6, VA = Department of Veterans Affairs, W/CML = word to competing message level

The use of recorded word lists for speech audiology has been advocated for more than 50 years. Watson and Tolan (1949, p. 442) advocated using recorded materials over live-voice audiology “...because a uniform standard can be employed.” Indeed, when live voice is used, test difficulty can change with the specific talker precluding test standardization (Kreul et al, 1968, 1969; Tillman and Olsen, 1973). When different talkers present the same word lists in a live-voice condition, each presentation becomes a different word recognition test. In spite of the need for test standardization, most practicing audiologists do not use recorded word lists for speech audiology (Wiley et al, 1995). The most recent survey of audiometric practices reported that 82 percent of respondents still use monitored live voice (Martin et al, 1998).

The reason(s) for audiologists not using recorded materials is unknown but may relate to matters of convenience and economy. In the past, word lists were recorded and played back using analog instrumentation. With analog tape and phonograph recordings, access to different tests was time consuming and the recordings deteriorated with use and time. Presently, however, several digital recordings are commercially available for clinical use (Wilson, 1997). Digital recordings, specifically compact discs (CDs), offer several benefits over analog recordings. Compact disc recordings allow for quick test access, are resistant to deterioration, and typically offer better fidelity than analog recordings (Wilson and Preece, 1990; Wilson, 1997). Digital recording methods provide excellent sig-
nal-to-noise levels, a wide frequency response, and near-perfect channel separation. If the same digital methods are used for each recording, then identical copies of the original speech materials can be reproduced CD to CD (Wilson, 1997). An advantage of producing identical recordings is that each reproduction is equivalent for testing and standardization purposes.

One of the more popular recorded word lists, the Northwestern University Auditory Test No. 6 (NU-6), was first developed by Tillman and Carhart in 1966 (Wilson and Preece, 1990; Wiley et al, 1995; Martin et al, 1998). Tillman and Carhart reported normative data for the NU-6 including articulation functions, test–retest reliability, and interlist equivalences for the word lists. Since that time, several analog and digital recordings of the NU-6 have been released for test purposes. Normative data have been published for some but not all of the recordings currently available (Wilson et al, 1976; Bess and Townsend, 1977; Beattie et al, 1977, 1978; Orchik et al, 1979; Causey et al, 1983; Frank and Craig, 1984; Maroonroge and Diefendorf, 1984; Wilson et al, 1990; Wilson and Oyler, 1997; Stuart et al, 1998).

Specific to the present study, in 1991 the Department of Veterans Affairs (VA) released a CD recording of the NU-6 for which normative data are not available. The NU-6 was among several tests on a CD entitled Speech Recognition and Identification Materials, Disc 1.1. Disc 1.1 was an update of the recording Speech Recognition and Identification Materials, Disc 1.0, produced by the VA in 1989. Both CDs (Discs 1.1 and 1.0) include versions of the NU-6 in quiet and in a competing message. Each version was recorded using a female speaker for the test words, a male speaker for the competing message, and the Modified Bell Telephone Sentences (Fletcher and Steinberg, 1929) as the competing message. Although both recordings (Discs 1.0 and 1.1) used the same test materials, speakers, and analog master tape, the final recordings were digitally mastered and produced using different methods. On Disc 1.0, the NU-6 was produced using a 12-bit digitizer and a 5000-Hz filter (low pass) cut-off (115 dB/octave). On Disc 1.1, a 16-bit digitizer and an 8800-Hz filter cutoff (96 dB/octave) were used. Root-mean-square (RMS) measures of the noise floor on each recording indicated that the noise floor was 67.9 dB down on Disc 1.0 and 81.2 dB down on Disc 1.1 (Syntrillium Software Corporation, Cool Edit Pro™, 1977). In addition, the test words (relative to the calibration tone) were recorded 0.5 dB higher on Disc 1.1 than on Disc 1.0. Recording changes also were made in the alignment of the competing message. On Disc 1.0, the carrier phrases and the NU-6 words were recorded within the temporal envelope of the competing message but did not fall at a fixed time within the message. On Disc 1.1, each carrier phrase was recorded 200 msec after the onset of the competing sentence. Finally, the level of the competing sentences was recorded 1.6 dB lower on Disc 1.1 than on Disc 1.0.

In summary, the two CD versions of the NU-6 released by the VA were produced with differences in the digitization processes used, in the temporal alignment of test words, and in the level of the competing sentences. These differences, in turn, may have created spectral and timing effects unique to each recording that would affect listener performance. Wilson et al (1990) have published normative data for the Disc 1.0 recording of the NU-6; data are not available for the NU-6 on Disc 1.1. The purpose of this study, therefore, was to provide normative data for the Disc 1.1 version of the NU-6 and to determine if the two different recordings (Discs 1.0 and 1.1) can be used interchangeably. Specifically, listener performance for the Disc 1.1 recording was evaluated and compared to the normative data reported by Wilson et al (1990) for Disc 1.0. Test–retest data also were obtained for the Disc 1.1 recordings of the NU-6 test in a competing message.

**METHOD**

**Participants**

Seventeen women and seven men between 17 and 35 years of age (mean = 23 years) participated in the study. Each participant had pure-tone thresholds of −5 to 5 dB HL (ANSI, 1989) for the octave test frequencies 250 to 8000 Hz in the test ear. The ear with the better pure-tone average at 1000, 2000, and 4000 Hz was selected as the test ear. If the pure-tone average at 1000, 2000, and 4000 Hz was equivalent for both ears, the test ear was selected at random. All participants had normal otoscopic findings, a negative history of otic disorder, normal tympanometric measures of middle ear function (ASHA, 1990), and ipsilateral acoustic reflexes present at 1000 Hz (≤105 dB HL).

**Procedures**

Tympanograms and acoustic reflexes were obtained using an acoustic admittance screening instrument (Grason-Stadler, GSI 27A). Pure-tone air-conduction audiometry was performed...
using a diagnostic audiometer (Grason-Stadler, GSI 16). The NU-6 word lists (Disc 1.1 female speaker version) were presented monaurally in quiet and in a competing message background (Modified Bell Telephone Sentences) via a CD player (Sony, Model 497) routed through one channel of the audiometer and then to an earphone (TDH 50-P) mounted in a supra-aural cushion (Telephonics P/N 510C017-1). All testing was completed in a sound-treated room (Industrial Acoustics Company, 1200 Series) meeting ambient noise requirements (ANSI, 1991). Appropriate calibrations were done before data collection (ANSI, 1987, 1989).

Testing for each participant was completed in two 75-minute sessions. The first session included a brief case history, otoscopy, screening tympanometry, acoustic reflex measures, pure-tone audiometry, and word recognition testing. During the same initial session, participants listened monaurally to 12 50-word lists (NU-6) in the presence of a 40 dB HL competing sentence. The level of the competing message was held constant for increasing presentation levels of the test words (NU-6). The word lists (NU-6) were presented incrementally in 4-dB steps at 12 levels from 16 to 60 dB HL. An ascending presentation order was used to minimize learning effects. Each participant listened to one of four recorded word lists (1A, 2A, 3A, 4A) at each presentation level. The list presented at each level was chosen at random without replacement. After all four word lists were presented, all were replaced and the selection process was repeated. Each word list was presented an equal number of times to each participant (three) and an equal number of times at each presentation level (six).

During the second session, word recognition performance was measured for the NU-6 word lists in quiet at presentation levels from 0 to 44 dB HL in 4-dB increments. The word lists in quiet were chosen and presented using the same selection process used for the word lists presented in a competing message. Again, each word list was presented an equal number of times to each participant and an equal number of times at each presentation level. Two additional word lists were randomly selected and presented to evaluate test–retest reliability. These word lists were presented at 32 dB HL and 44 dB HL with a 40 dB HL competing message. Presentation levels were again arranged in an ascending order to minimize learning effects. The lowest levels were presented first and the highest levels last. For all tests, participants wrote their responses to each word.

**RESULTS**

**Word Recognition in Quiet**

Mean word recognition scores in quiet for the Disc 1.1 version of the NU-6 test are shown in Figure 1 (triangles). The data previously reported by Wilson et al (1990) for the Disc 1.0 recordings also are shown (squares). Summary information for paired t-tests, mean recognition scores for all presentation levels (in quiet), and standard deviations are provided in Table 1. Although the articulation functions obtained for both the Disc 1.1 and 1.0 versions of the NU-6 exhibit similar shapes (see Fig. 1), the mean recognition performance was better for the Disc 1.1 version of the NU-6 at all presentation levels (see Table 1). These differences were significant at 12 dB HL ($t_{(23)} = 0.0429$, $p < .05$), 16 dB HL ($t_{(23)} = 0.0161$, $p < .05$), 20 dB HL ($t_{(23)} = 0.0115$, $p < .05$), and 24 dB HL ($t_{(23)} = 0.0082$, $p < .05$). Each articulation curve was fit with a third-degree polynomial function. The presentation level at which equivalent performance would occur for both data sets was then calculated at each 10 percent increment from 20 percent to 80 percent correct. Based on these calculations, equal word recognition performance would have occurred at levels from 2.2 to 2.9 dB lower for the Disc 1.1 version of the NU-6 than for the Disc 1.0 version of the test.

![Figure 1](image-url)

Figure 1 Articulation functions for the NU-6 word lists, in quiet, on VA Disc 1.1 (triangles) and those reported by Wilson et al (1990) for the quiet version on VA Disc 1.0 (squares). The lines connecting the data points are best-fit, third-degree polynomials (Disc 1.1: $y = 0.5290 + 1.1116x + 0.2148x^2 - 0.0048x^3$; Disc 1.0: $y = -0.8989 - 0.3962x + 0.2858x^2 - 0.0056x^3$).
Table 1  Mean Percent Correct Recognition and Standard Deviations by Presentation Level for VA Discs 1.0 and 1.1 Recordings of the NU-6 Word Lists Presented in Quiet

<table>
<thead>
<tr>
<th>Presentation Level (dB HL)</th>
<th>Disc 1.0 Mean (SD)</th>
<th>Disc 1.1 Mean (SD)</th>
<th>Mean Difference t</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.0 (0)</td>
<td>2.6 (9.0)</td>
<td>2.6 1.4</td>
</tr>
<tr>
<td>4</td>
<td>1.5 (2.1)</td>
<td>7.3 (18.0)</td>
<td>5.8 1.6</td>
</tr>
<tr>
<td>8</td>
<td>9.3 (7.2)</td>
<td>16.2 (15.0)</td>
<td>6.9 2.0</td>
</tr>
<tr>
<td>12</td>
<td>25.0 (12.1)</td>
<td>34.8 (18.7)</td>
<td>9.8 2.1</td>
</tr>
<tr>
<td>16</td>
<td>44.0 (15.6)</td>
<td>57.8 (20.9)</td>
<td>13.8 2.6</td>
</tr>
<tr>
<td>20</td>
<td>63.8 (13.5)</td>
<td>74.3 (12.9)</td>
<td>10.5 2.8</td>
</tr>
<tr>
<td>24</td>
<td>76.0 (10.6)</td>
<td>84.3 (9.3)</td>
<td>8.3 2.9</td>
</tr>
<tr>
<td>28</td>
<td>86.8 (5.6)</td>
<td>90.3 (6.7)</td>
<td>3.5 2.0</td>
</tr>
<tr>
<td>32</td>
<td>92.7 (3.8)</td>
<td>93.4 (5.5)</td>
<td>0.7 0.5</td>
</tr>
<tr>
<td>36</td>
<td>94.1 (4.3)</td>
<td>96.3 (3.7)</td>
<td>2.2 1.9</td>
</tr>
<tr>
<td>40</td>
<td>97.9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>44</td>
<td>97.5</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Summary information for paired t-tests and mean differences (in percent) also is provided. All t-tests had 23 degrees of freedom. t values marked with an asterisk (*) are significant (p < .05). Data were not reported by Wilson et al (1990) for VA Disc 1.0 at 40 and 44 dB HL.

Table 2  Mean Percent Correct Recognition and Standard Deviations by Word/Competing Message Level (W/CML) for VA Discs 1.0 and 1.1 Recordings of the NU-6 Word Lists Presented in a Competing Sentence

<table>
<thead>
<tr>
<th>W/CML Level in dB</th>
<th>Disc 1.0 Mean (SD)</th>
<th>Disc 1.1 Mean (SD)</th>
<th>Mean Difference t</th>
</tr>
</thead>
<tbody>
<tr>
<td>-24</td>
<td>—</td>
<td>1.2 (2.3)</td>
<td>—</td>
</tr>
<tr>
<td>-20</td>
<td>0.8 (2.1)</td>
<td>4.3 (4.8)</td>
<td>3.5 3.2*</td>
</tr>
<tr>
<td>-16</td>
<td>6.3 (4.7)</td>
<td>11.6 (8.1)</td>
<td>5.3 2.8*</td>
</tr>
<tr>
<td>-12</td>
<td>15.3 (6.6)</td>
<td>28.3 (12.0)</td>
<td>13.0 4.6*</td>
</tr>
<tr>
<td>-8</td>
<td>32.1 (8.2)</td>
<td>45.7 (15.6)</td>
<td>13.6 3.8*</td>
</tr>
<tr>
<td>-4</td>
<td>49.8 (8.7)</td>
<td>60.6 (13.3)</td>
<td>10.8 3.2*</td>
</tr>
<tr>
<td>0</td>
<td>63.2 (7.8)</td>
<td>76.8 (9.4)</td>
<td>13.6 5.5*</td>
</tr>
<tr>
<td>4</td>
<td>74.3 (4.6)</td>
<td>85.8 (6.6)</td>
<td>11.5 7.0*</td>
</tr>
<tr>
<td>8</td>
<td>84.3 (5.0)</td>
<td>89.9 (5.6)</td>
<td>5.6 3.7*</td>
</tr>
<tr>
<td>12</td>
<td>89.3 (3.8)</td>
<td>93.3 (4.1)</td>
<td>4.0 3.6*</td>
</tr>
<tr>
<td>16</td>
<td>92.9 (3.7)</td>
<td>95.4 (3.8)</td>
<td>2.5 2.3*</td>
</tr>
<tr>
<td>20</td>
<td>97.3</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Summary information for paired t-tests and mean differences (in percent) also is provided. All t-tests had 23 degrees of freedom. t values marked with an asterisk (*) are significant (p < .05). Data were not reported by Wilson et al (1990) at the -24 dB and 20 dB W/CML.

Word Recognition in Competing Message

Mean word recognition scores (Disc 1.1) in competing message (40 dB HL) as a function of presentation level are shown in Figure 2.

![Figure 2](image)

Figure 2  Articulation functions for the NU-6 word lists in a competing message on VA Disc 1.1 (triangles) and those reported by Wilson et al (1990) for VA Disc 1.0 (squares). The competing message was 40 dB HL for all presentation levels tested. The lines connecting the data points are best-fit, third-degree polynomials (Disc 1.1: y = 71.8769 + 3.1841x - 0.0558x^2 - 0.0025x^3; Disc 1.0: y = 61.7770 + 3.5549x - 0.0492x^2 - 0.0036x^3).

Test–Retest Reliability

Test–retest reliability was evaluated by comparing mean percent correct scores obtained for two randomly selected lists repeated at two W/CMLs. Pearson correlation coefficients were 0.9 and 0.8 for repeated tests at the -8 dB and the 4-dB W/CMLs, respectively. At the -8 dB W/CML, mean word recognition scores improved from 45.7 percent on the first trial to 56.5 percent on the second trial. At the 4-dB W/CML, the mean score improved from 85.8 percent to 90.3 percent across trials. These improvements in mean scores were significant at the -8 dB...
W/CML (t [23] = 7.1, p < .05) and at the 4-dB W/CML (t [23] = 4.6, p < .05).

DISCUSSION

As indicated in Tables 1 and 2, mean word recognition scores at equal presentation levels were higher for the NU-6 released on Disc 1.1 than for the Disc 1.0 version. Although the precise reasons for the improvement in performance for the Disc 1.1 recording are unclear, several differences in the recording procedures for the two recordings offer a viable explanation. First, the NU-6 recorded on Disc 1.1 was digitized using 16 bits, which divides the amplitude range into 65,536 values ($2^{16}$). Disc 1.0 was digitized using 12 bits, which divides the amplitude range into 4096 possible values ($2^{12}$). Materials digitized in 16 bits, then, are a more accurate representation of the analog waveform than are materials digitized with 12 bits. The audio trade literature relates 16 bits comparable to CD-quality recordings and 12 bits comparable to high-quality cassette recordings (Cool Edit Pro™, 1997). The dynamic range for 16 bits is 96 dB, whereas the dynamic range for 12 bits is 72 dB. The measured noise floor for the NU-6 on Disc 1.1 was lower (81.2 dB down) than the noise floor for the NU-6 on Disc 1.0 (67.9 dB down). Second, the test words relative to the calibration tone were recorded 0.5 dB higher on Disc 1.1 than on Disc 1.0, which could partially account for the higher scores on Disc 1.1 in quiet. Third, the level of the competing sentences was 1.6 dB lower on Disc 1.1 than on Disc 1.0. This difference in the level of the competing message also could account for a portion of the improvement in scores for Disc 1.1 in a competing message. A higher score, for example, might be expected for the recording (Disc 1.1) with the slightly lower competing message. Fourth, the temporal alignment of the NU-6 materials within the envelope of the competing message was different for Disc 1.0 and Disc 1.1. With Disc 1.1, the onset of the carrier phrase was always 200 msec after the onset of the competing sentence. On Disc 1.0, however, the onset of the carrier phrase was at some random interval after the onset of the competing sentence. The random nature of the relation between the onset of the word and the competing sentence in the Disc 1.0 materials is a source of uncertainty for the listener with respect to when the words occur within the envelope of the competing sentences. This listener uncertainty could potentially have resulted in portions of the reduced performance for the competing message tests on Disc 1.0 relative to that for Disc 1.1.

Test–retest reliability for the Disc 1.1 version of the NU-6 in a competing message was good. Although word recognition scores were significantly higher from session to session, the differences were small relative to the expected variability for middle-range scores (Thornton and Raffin, 1978). The slight improvement from test to test could be attributed to learning effects or practice with a difficult listening task.

Finally, although several of the observed score differences across disc versions were statistically significant, absolute differences were small, particularly when compared to the score variability predicted by the binomial model. The largest difference (14%), for example, occurred for middle-range scores where variability is greatest (Thornton and Raffin, 1978). Calculations of equivalent word recognition performance by presentation level indicated that comparable scores could be obtained across the two disc versions with only a 2.2 to 4.2 dB adjustment in presentation level. Differences this small are unlikely to affect clinical decisions related to word recognition performance.

CONCLUSION

Current digital recording methods allow for the production of multiple generations of identical audio recordings. Even slight changes in the recording or digital mastering process, however, can alter the eventual test recording and can affect listener performance. Carhart (1965, p. 254) stressed the importance of the issue over 30 years ago, stating, “There might be as much difference between one recording and another as between two live-voice talkers.” The results of the current study exemplify this point. Differences in the recording methods used for Discs 1.0 and 1.1 were associated with differences in word recognition performance. These results reinforce the need for standardization of all versions of recorded speech recognition tests.

Clinically, the differences in scores between test versions (Discs 1.0 and 1.1) and across test sessions (Disc 1.1) are probably of minor consequence with regard to diagnostic decisions. The observed differences were small relative to known characteristics in binomial score variability. These small differences in word recognition scores suggest that the different versions of the NU-6 on Discs 1.0 and 1.1 could be used
with little or no difference in clinical outcomes. An important part of the standardization process for any test, including tests of speech recognition, is the collection of data from the target population(s) for whom the test is intended (Conrad, 1951; American Psychological Association, 1974; Bilger, 1984). Although the present study was limited to data analyses for normal-hearing listeners, the data are useful for comparison purposes in epidemiologic studies (Wiley et al., 1998) and in studies using listeners with varying degrees and configuration of hearing loss. In this regard, future research directed toward obtaining additional data from listeners with varying degrees and configuration of hearing loss are necessary and should prove useful for future clinical applications using NU-6 with these target populations.

REFERENCES


REFERENCES


