Department of Veterans Affairs

Journal of Rehabilitation Research and Development
Vol. 40, No. 4, July/August 2003
Pages 329–336

Learning effects associated with repeated word-recognition measures using sentence materials

Richard H. Wilson, PhD; Theodore S. Bell, PhD; John A. Koslowski, MS
James H. Quillen Department of Veterans Affairs Medical Center, Mountain Home, TN; Departments of Surgery and Communication Disorders, East Tennessee State University, Johnson City, TN; Department of Communication Disorders,California State University, Los Angeles, CA; VA Medical Center, Fort Howard, MD

Abstract—We investigated the learning effects of repeated presentation of sentence materials in an adaptive paradigm in five sessions over 5 to 10 days using 10 subjects in each of three age groups (<30 years, 40 to 60 years, and >65 years). Three target words, based on word-usage frequency and word confusability, were embedded within the seven to nine syllable sentences. Thresholds were obtained for the control lists in Sessions 1 to 5 and for the experimental lists in Sessions 1 and 5. The experimental lists were withdrawn in Sessions 2 to 4. The mean thresholds (1) for the three subject groups were significantly different, (2) for the experimental conditions and the control conditions were not significantly different, and (3) for Session 5 were significantly lower than in Session 1. The implication is that improved thresholds were the result of the subjects learning the test procedure (including the listening-response task, speaker familiarity, and test environment) and not from learning the test words/sentences.

Key words: adaptive procedure, auditory threshold, psychometric function, threshold reversal, word recognition.

INTRODUCTION

Current clinical procedures typically include recognition of isolated monosyllabic or spondaic words in a format that poorly represents communication in real-life situations. The validity of these tests depends on the assumption that the paradigms represent everyday speech. Although syllable and word tests may give insights into certain perceptual problems, syllable and word tests do not reasonably approximate everyday listening conditions. Several formats are employed that use sentence materials. They range from simple interrogative sentences, which the subject answers [1], to target-word formats in which the subject identifies target words within the sentence [2], to the SPIN format that controls the predictability of the target word [3], and to meaningful sentences of everyday language that have been employed with an adaptive procedure format that determines specific points (e.g., 50 percent) on the psychometric function [4].

Abbreviations: ANOVA = analysis of variance, CVC = consonant-vowel-consonant, HD = high density (frequency of usage word from a dense neighborhood), HL = hearing level, HS = high sparce (high frequency of usage word from a sparse neighborhood), LD = low density (low frequency of usage word from a dense neighborhood), LS = low sparce (low frequency of usage word from a sparse neighborhood), NU 6 = Northwestern University Auditory Test No. 6.

This material was based on work supported by a merit review from the Rehabilitation Research and Development Service, Department of Veterans Affairs. The senior author is a senior research career scientist supported by the Rehabilitation Research and Development Service.

Address all correspondence and requests for reprints to Richard H. Wilson, PhD; VA Medical Center, Audiology (126), Mountain Home, TN 37684; 423-979-3653; fax: 423-979-3403; richard.wilson2@med.va.gov.

The overall project involved the development and evaluation of sentence materials based on target words selected using two criteria: word-usage frequency and word confusability based on a single phoneme substitution metric [5]. The aspect of the project reported here examined the learning effects associated with repeated presentations of word-recognition materials with the use of sentence materials and an adaptive psychophysical procedure. Egan reported that with multiple trials of listening to words in an ambient noise [6], recognition performance improved from about 60 percent correct to about 80 percent correct over 5 successive practice days. Egan did not attribute the improved performance to any particular aspect of the listening-response task. The design of the current experiment permitted the examination of two questions. The first question addressed whether or not thresholds measured with sentence materials changed with repeated measures. The second question, which was based on a positive answer to the first question, addressed the source of possible learning effects, i.e., were the learning effects from repeated measures owing to (1) learning the words that compose the sentences; (2) learning the test environment and the listening-response task, including familiarity with the speaker; or (3) a combination of learning the materials and learning the listening task?

METHODS

Subjects

Three groups of 10 subjects each were studied. The three groups were included to determine if practice on the listening task had different effects on listeners of various age groups and degrees of hearing loss. The <30 years group consisted of young adults (mean = 23.7 years) with normal hearing <20 dB HL at the 250 Hz to 8,000 Hz octave frequencies [7]. The 40 to 60 years group consisted of adults (mean = 51.8 years) with mild-to-moderate sensorineural hearing loss, whereas the >65 years group consisted of adults (mean = 72.2 years) with mild-to-severe sensorineural hearing loss. The mean audiograms for the ears tested in each of the three groups are shown in **Figure 1**. The following standard deviations were observed for each subject group at the 250 Hz to 8,000 Hz octave intervals: 4, 4, 6, 7, 7, and 8 dB (<30 years group); 7, 8, 10, 12, 15, and 21 dB (40 to 60 years group); and 12, 11, 13, 20, 23, and 27 dB (>65 years).

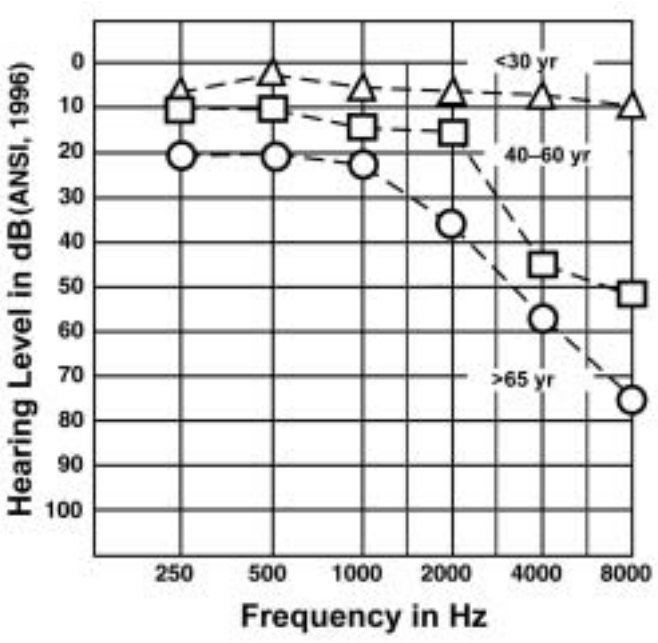

Figure 1.
Audiogram for three subject groups, Δ (<30 years), □ (40 to 60 years), and O (>65 years).

Materials

The sentences, each of which had seven to nine syllables and three target words, were homogeneous with respect to psychometric slope, range, and midpoint [5,8]. Development of the sentences is described elsewhere [5]. Briefly, the sentences were grouped into one of the following four categories based on the word category of the three target words: (1) HD (high density)—high frequency of usage from a dense neighborhood, (2) HS (high sparce)—high frequency of usage from a sparse neighborhood, (3) LD (low density)—low frequency of usage from a dense neighborhood, and (4) LS (low sparce)—low frequency of usage from a sparse neighborhood. The categorization of low versus high frequency of word usage was based on the entire set of familiar monosyllabic consonant-vowel-consonant (CVC) words in the pocket lexicon, which was described by a 7-point rating scale [9]. Sparse and dense refer to similarity neighborhoods in an assumed representation of the mental lexicon. A sparse word is relatively phonetically unique; i.e., it sounds like very few other words. A dense word is phonetically similar to many other words. Words from each lexical category were used three at time to form the sentences in a format similar to that used by Plomp and Mimpen or Bench and Bamford [4,10]. That is, the three

target words in each sentence were from the same word category. For this study, six lists of 25 sentences each were compiled. Because substantially more words met selection criteria in the HD and HS categories than in the LD and LS categories, the HD and HS categories had two lists each, whereas the LD and LS categories had one list each. Two randomizations of each list were available.

For each subject, one list from both the HD and HS categories was designated as the "experimental" list with the other HD and HS lists assigned as "control" lists. The experimental lists were lists on which thresholds were established only in the first and fifth test sessions, i.e., the experimental materials were withdrawn from and not presented during Sessions 2, 3, and 4. The control lists were materials that were administered in each of the five sessions that served as practice items with multiple thresholds being established in each session. For a given subject, then, the control lists were used throughout the five sessions to provide practice on the listening and response task, whereas the experimental sentence lists were used only during Sessions 1 and 5. Because of the limited number of LD and LS sentences, the LD and LS lists were always used as experimental lists. Four experimental lists (HD, HS, LD, and LS) and two control lists (HD and HS) were selected for each listener.

Procedures

Each subject participated in a 1-hour listening session per day for 5 days during a 5- to 10-day interval. Eight thresholds for the sentence materials were established in each of the five sessions. During Sessions 1 and 5, thresholds for four experimental conditions (HD, HS, LD, and LS) were obtained, as were thresholds for four control conditions (two randomizations of HD and HS). During Sessions 2, 3, and 4, thresholds for eight control conditions were established (two conditions—HD and HS, by two randomizations, by two replications). Presentation orders during each session were randomized.

The sentences were reproduced from a digital audio tape (Sony, Model DTC-59ES), routed through an audiometer (Grason-Stadler, Model 10), and presented to the subjects through a TDH-50P earphone encased in a P/N 510C017-1 cushion. Left ears were used on the odd number subjects and right ears were used on the even number subjects. All testing was conducted in a sound booth (IAC, Model 1205). An adaptive procedure was used to establish thresholds for the sentence materials [11]. The task of the subject was to repeat the sentence, with scoring based on the three target words in each sentence. If two or three of the target words were recognized, then the level of the subsequent sentence was decreased 2 dB; if fewer than two of the target words were recognized, then the level of the subsequent sentence was increased 2 dB. The threshold track was terminated after 12 reversals with the first 3 reversals discarded and the last 9 reversals used to compute threshold, which was defined as the midpoint of the excursion. In addition to the threshold metric, the amplitudes of the excursions of the threshold tracks were evaluated.

Finally, a subset of six listeners from the 40 to 60 years group and six listeners from the >65 years group participated in a sixth session that replicated the eight conditions administered in Sessions 1 and 5. Session 6 was included to determine the extent of the carryover associated with the learning effects that were observed in the first five sessions. A 20- to 30-day hiatus separated Sessions 5 and 6.

RESULTS AND DISCUSSION

The relationships among the data for the four conditions were the same. Every subject in each of the two control conditions (HD and HS) and in each of the four experimental conditions (HD, HS, LD, and LS) had lower thresholds in Session 5 than in Session 1. Accordingly, an analysis of variance (ANOVA) indicated that thresholds obtained in Session 5 were significantly lower than the thresholds obtained in Session 1 [$F(1,27) = 685.0$, $p < 0.01$]. For this reason, only the data for the HD condition are presented to illustrate the various relations that were observed in the study. The mean word-recognition thresholds for the HD condition across the five sessions for the three subject groups are illustrated in **Figure 2** and are listed in **Table 1**, which includes the standard deviations (SDs) and the differences between the thresholds obtained in Session 1 and Session 5. Multiple trials of the various control conditions were administered during Sessions 1 and 5 (two trials) and Sessions 2, 3, and 4 (four trials). For each session, only the data from the first trial of each condition in that session are reported. The mean data in the lower, middle, and upper portions of the figure are from the <30 years group, the 40 to 60 years group, and the >65 years group, respectively. The experimental data (Sessions 1 and 5) are depicted with filled symbols connected with dashed lines; the open symbols with solid

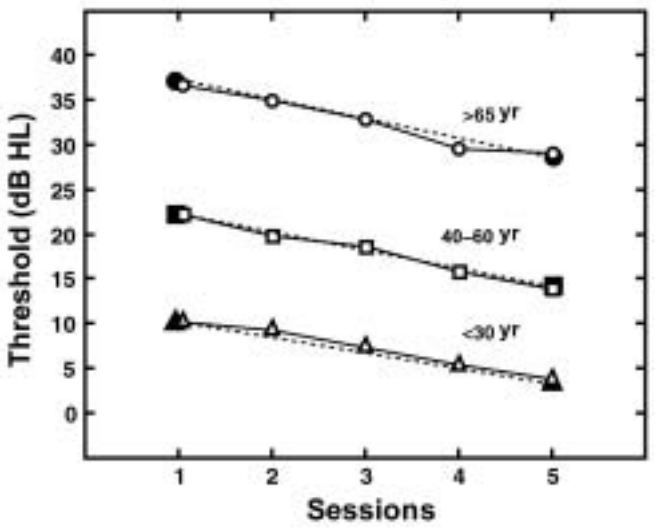

Figure 2.
Mean thresholds (dB HL) for HD experimental conditions (filled symbols and dashed lines) and for HD control conditions (open symbols) obtained in each of the sessions. In each panel, data for each of three age groups are shown as O (>65-years), □ (40–60 years), and Δ (<30 years).

lines represent the data for the control conditions used in each of the five sessions.

Three general results are apparent from the data in **Figure 2** and **Table 1**. First, the mean thresholds for the three groups of subjects are different. An ANOVA with repeated measures (two within [condition/session] and one between [subject group]) indicated that the threshold differences among the three subject groups were significant [$F(2, 27) = 21.2$, $p < 0.01$]. Post hoc Scheffé tests indicated that thresholds for each subject group were significantly different from the thresholds for the other subject groups ($p < 0.01$). This finding was expected as the subjects were selected from specific age-related categories that reflect different degrees of hearing loss. Second, as indicated earlier, the thresholds obtained in Session 5 were significantly lower than the thresholds obtained in Session 1. For each of the four experimental (HD, HS, LD, and LS) and two control conditions (HD and HS), the pair differences (Sessions 1 and 5) were significant ($p < 0.01$). Although the thresholds for the three groups of listeners are displaced from one another in **Figure 2**, the functions for the three groups across the five sessions are essentially parallel. This finding indicates that the improvements noted in recognition performances across the sessions were the same for all groups. Thus, age differences and degree of hearing loss did not influence the degree of improvement in performance noted between Sessions 1 and 5.

The thresholds for the HD condition displayed in **Figure 2** for the <30 years group were 10.5 dB HL for Session 1, decreasing to about 4 dB HL by Session 5. This 6 dB to 7 dB difference for the HD experimental conditions, which was significant, is highlighted in **Figure 3(a)** in which the thresholds from the HD, HS, LD, and LS conditions for the individual subjects from Session 1 (abscissa) are plotted against the threshold for the corresponding condition in Session 5 (ordinate). All the data points are to the right of the diagonal lines, indicating higher thresholds in Session 1 than in Session 5. With the 40 to 60 years group (**Figure 3(b)**) and the >65 years

Table 1.
Mean thresholds (dB HL) and SDs (in parentheses) for HD control and experimental conditions across five sessions for three groups of subjects.

Material/Condition	Session					Difference
	1	2	3	4	5	1 to 5
<30 Years Group						
Control	10.5 (1.4)	9.5 (0.7)	7.7 (0.9)	5.6 (0.8)	4.0 (0.5)	6.5
Experimental	10.5 (1.4)	—	—	—	3.7 (0.8)	6.8
40 to 60 Years Group						
Control	22.4 (4.5)	18.9 (4.4)	18.7 (4.0)	15.9 (3.7)	14.0 (2.8)	8.4
Experimental	22.5 (4.2)	—	—	—	14.4 (3.5)	8.1
>65 Years Group						
Control	36.8 (15.4)	35.0 (15.6)	33.0 (14.7)	29.6 (14.8)	29.1 (15.1)	7.7
Experimental	37.2 (15.2)	—	—	—	28.8 (14.3)	8.4

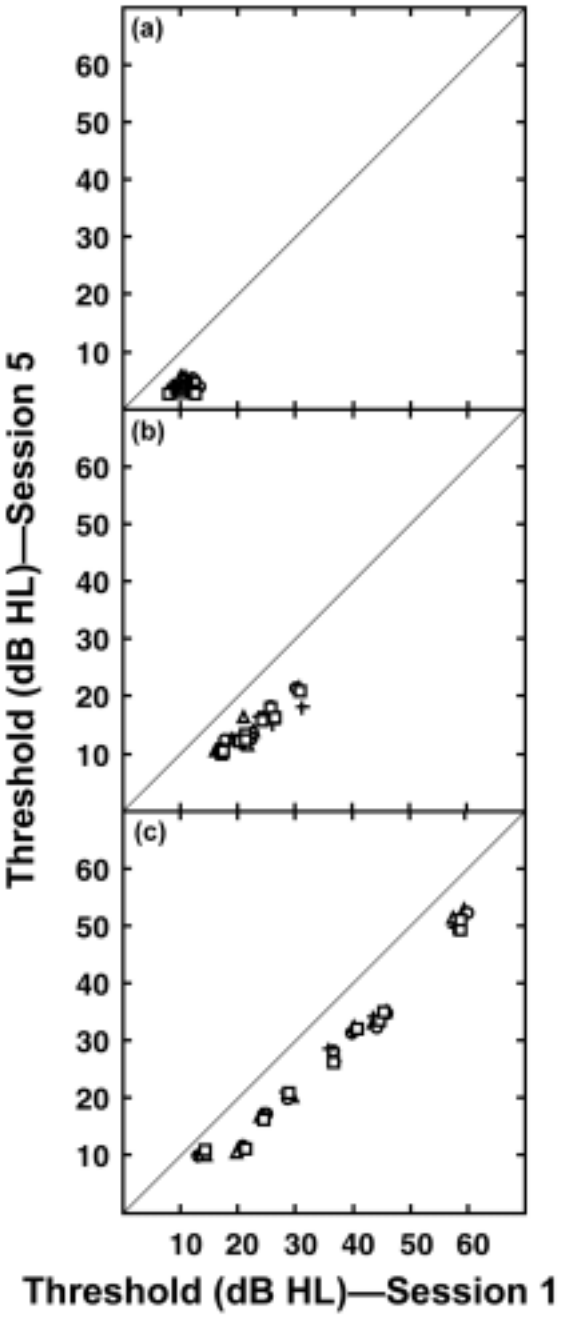

Figure 3.
A bivariate plot of individual thresholds (dB HL) for experimental conditions obtained in Session 1 (abscissa) and in Session 5 (ordinate). Diagonal lines represent equal performance. HD, HS, LD, and LS data are represented by □, O, Δ, and +, respectively. **(a)** <30 years, **(b)** 40 to 60 years, and **(c)** >65 years.

group (**Figure 3(c)**) a similar, but slightly larger, difference (7.7 dB to 8.4 dB) exists between the individual subject thresholds in Session 1 and Session 5. This second finding, which answers the first research question, indicates that for each of the three subject groups, the thresholds improved significantly between the first and fifth session. Egan and Dubno et al. observed similar improvements in word-recognition performance [6,12]. Egan noted a 20 percent (4 dB to 5 dB) improvement in the recognition of the PB-50 words presented in noise on 5 successive days. More recently, Dubno et al. in a similar study with the NU 6 materials presented in noise observed a 6 percent improvement in recognition performance over nine trials. Neither study, however, was designed to examine the learning components that contributed to the improved performances that were observed.

The distribution of the datum points for the <30 years group in **Figure 3(a)** is tightly grouped, reflecting homogeneity in recognition performance. In contrast to the <30 years group, the datum points for the other two subject groups progressively become less homogeneous as age and degree of hearing loss increase. This is an expected finding when the variability associated with the hearing sensitivity between groups and within groups reflected in the pure-tone audiometric findings is considered (see **Figure 1**).

The second research question addressed whether or not decreases in thresholds across sessions were owing to the subjects learning the test material, learning the test and listening environment, or a combination of both learning effects. From **Figure 2** and **Table 1**, comparison of the thresholds for the experimental and control conditions obtained in Session 1 indicates essentially no difference between the two sets of data. The same relation is observed for the data in Session 5. Consider the data from the 40 to 60 years group. The mean thresholds for the HD condition in Session 1 were 22.4 dB and 22.5 dB HL for the control and experimental conditions, respectively. As the session number increased, the mean threshold for the control condition systematically decreased from 18.9 to 18.7 to 15.9 dB HL in Sessions 2, 3, and 4, respectively. In Session 5, the mean threshold for the control condition was 14.0 dB HL that represented a threshold decrease of 8.4 dB from the 22.4 dB HL threshold obtained in Session 1. The threshold for the experimental condition in Session 5 was 14.4 dB HL. The decrease in threshold must be attributable to improvement that the subjects experienced with repeated practice in Sessions 2, 3, and 4 with the task. Because the materials for the control condition were presented in each of the five sessions, the control condition cannot possibly differentiate the effects associated

with learning the materials (i.e., the test words) from the effects associated with learning the test procedure and test environment (i.e., wearing earphones, sitting in a sound booth, listening to signals at low levels, repeating stimulus materials, becoming familiar with the various characteristics of the speaker, etc.). The data from the experimental conditions, however, provide insight that differentiates the two types of learning effects.

In comparison to the exposure received by the subjects to the materials in the control conditions, the practice received by the subjects to the materials in the experimental conditions was minimal. The thresholds for the control and experimental conditions were lowered by the same amount between Sessions 1 and 5 (**Table 1**, right column). The practice received on the control conditions improved the word-recognition performance by the subjects on the control materials. Likewise, the recognition performance of the subjects improved on the experimental materials on which no practice was received during Sessions 2, 3, and 4. The implication is that the subjects were not learning the test materials (words and/or sentences) but rather the subjects were learning the test procedure, test environment, etc. Although not presented, the HS, LD, and LS materials demonstrated these exact relations.

To this point, the focus has been on the control and experimental data from Sessions 1 and 5. Sessions 2, 3, and 4 were used as control conditions in which eight thresholds (four HD and four HS) were obtained from each subject in each session. These control conditions provided the subjects with practice on the listening task and materials. The mean thresholds for the HD condition (and SDs) for the four trials within each of Sessions 2, 3, and 4 are listed in **Table 2** for each of the subject groups. Two relations are of interest in the data. First, as was reported earlier, as the session number increased, the thresholds decreased, a relation that is true for each subject group. Second, the thresholds for a given condition changed little, if any, across trials of a given session. Collapsed across subject groups and sessions (2 to 4), the overall mean threshold for Trial 4 within a session (19.2 dB HL) was only 0.2 dB lower than the mean threshold for Trial 1 within a session (19.4 dB HL). Thus, the thresholds across trials within a session remain remarkably stable, which is in contrast to the systematic decrease in thresholds that was observed across sessions.

The threshold data for the HD condition from six of the 40- to 60-year-old listeners and six of the >65-year-old listeners who participated in Session 6 (20 to 30 days following Session 5) are depicted in **Figure 4**. For comparison, the mean data from the respective subsets of subjects obtained in Sessions 1 through 5 are shown. For both subject groups, recognition performance regressed slightly (2.9 dB to 3.3 dB) between Sessions 5 and 6, but all thresholds for the individual subjects were lower in Session 6 than in Session 1. These results indicate that over the 20- to 30-day interval, the listeners retained a portion of the listening sophistication that had been acquired during the initial five test sessions.

Finally, in addition to the threshold measure obtained from each set of data, the amplitudes of the last nine excursions of the threshold tracks were evaluated. The mean track excursion sizes (in decibel) are depicted in **Figure 5** for the various HD listening conditions, listening sessions, and subject groups. A general characteristic of both the control data (open symbols with solid lines) and the experimental data (filled symbols with dashed lines) is that the excursion size decreased as the number of sessions increased. For both the <30 years group and the 40 to 60 years group, the excursion size systematically decreased ~0.5 dB from 2.8 dB in Session 1 to 2.3 dB in Session 5. With the >65 years group, the decrease in excursion size was reduced to about 0.2 dB, from 2.7 dB to 2.5 dB. Although these differences are small, the findings indicate that as practice on the listening task increased, the subjects became more proficient at the listening-response task. By Session 5, threshold was “bracketed” typically by little more than one-step size (2 dB). Because the sizes of the track excursions of the control and experimental conditions changed by similar amounts between Sessions 1 and 5, one can conclude, as with the threshold measures, that the equivalent changes in performance on the control and experimental conditions were owing to practice or exposure on the test procedure and test environment and not to learning the target stimulus words and sentences.

CONCLUSIONS

For a word-recognition task using sentence materials, young subjects with normal hearing and older subjects with hearing loss demonstrated improved thresholds with an adaptive psychophysical procedure over test sessions on 5 days. Comparison of the data from the control and experimental conditions indicated that the improvements

Table 2.
Mean thresholds (dB HL) and SDs (in parentheses) for the HD control conditions across four trials in Sessions 2, 3, and 4.

Group/Session	Trial			
	1	2	3	4
<30 Years Group				
Session 2	9.5 (0.7)	9.7 (1.1)	9.2 (0.8)	9.3 (0.9)
Session 3	7.7 (0.9)	7.2 (1.3)	6.8 (1.1)	6.9 (1.1)
Session 4	5.6 (0.8)	5.6 (0.9)	5.2 (0.6)	5.3 (0.8)
40 to 60 Years Group				
Session 2	19.9 (4.2)	20.2 (4.1)	20.3 (4.3)	20.1 (4.1)
Session 3	18.7 (4.0)	17.9 (3.5)	18.0 (3.3)	18.7 (4.3)
Session 4	15.9 (3.7)	16.5 (4.2)	16.5 (3.5)	16.4 (4.0)
>65 Years Group				
Session 2	35.0 (15.6)	34.8 (15.2)	34.3 (15.1)	34.8 (15.1)
Session 3	33.0 (14.7)	32.6 (14.9)	32.4 (15.1)	31.9 (14.9)
Session 4	29.6 (14.8)	30.0 (15.1)	29.3 (14.8)	29.1 (14.6)

in performance were owing to the listeners learning (i.e., becoming more sophisticated), the listening-response task, and the listening environment. The improvement was not attributable to the listeners learning the test words and/or sentences. Based on the data across the test sessions, the test-retest characteristic of the materials and procedures would be poorer; however, based on the data across trials within a session, the test-retest characteristic would be good. The current findings provide one type of evidence that subjects can improve word-recognition performance with practice on the listening task. Because the adaptive psychophysical procedure and sentence materials involved in the current study are different from the traditional clinical method used to access word-recognition

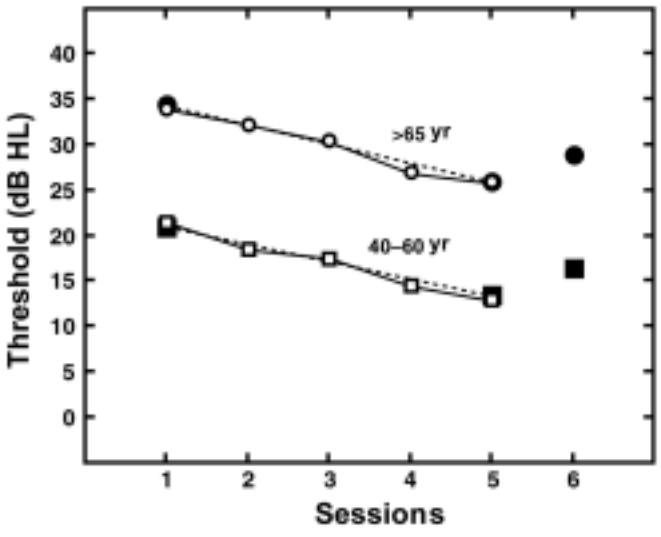

Figure 4.
Mean thresholds (dB HL) for HD experimental conditions (filled symbols and dashed lines) and for HD control conditions (open symbols) obtained in six sessions from a subset of subjects from 40 to 60 years and >65-years age groups. A 20- to 30-day interval occurred between Sessions 5 and 6.

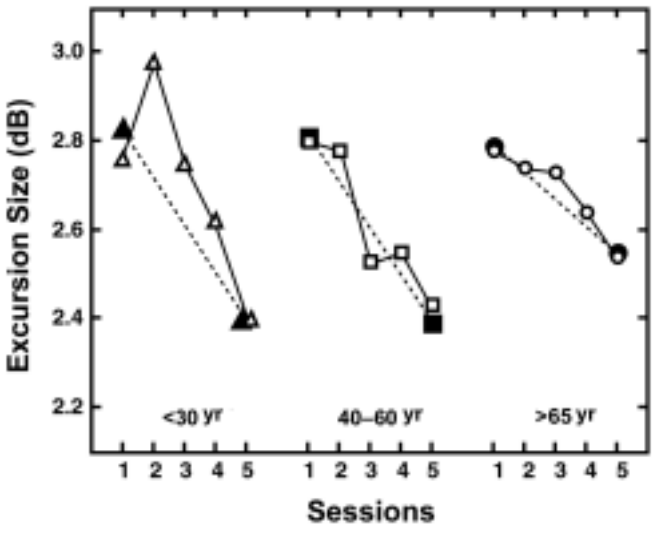

Figure 5.
Mean excursion amplitudes (dB) for HD experimental conditions (filled symbols and dashed lines) and for HD control conditions (open symbols) obtained in each of sessions. In each panel, data for each of three age groups are shown as circles (>65 years), squares (40 to 60 years), and triangles (<30 years).

abilities using monosyllabic words presented at a fixed level, one must use caution in applying the findings from the former to the latter.

REFERENCES

1. Fletcher H. Speech and Hearing. (1st ed.) New York: Van Nostrand; 1929.
2. Silverman SR, Hirsh IJ. Problems related to the use of speech in clinical audiometry. Ann Otol Rhinol Laryngol 1955;64:1234–44.
3. Kalikow DN, Stevens KM, Elliot LL. Development of a test of speech intelligibility in noise using sentence materials with controlled word predictability. J Acoust Soc Am 1977;61:1337–51.
4. Plomp R, Mimpen AM. Improving the reliability of testing the speech reception threshold for sentences. Audiology 1979;18:43–52.
5. Bell TS, Wilson RH. Sentence recognition materials based on frequency of word usage and lexical confusability. J Am Acad Audiol 2001;12:524–32.
6. Egan JP. Articulation testing methods. Laryngoscope 1948; 58:955–91.
7. American National Standards Institute (ANSI). Specification for audiometers (ANSI S3.6-1996). New York: ANSI; 1996.
8. Bell TS, Wilson RH. A new clinical word recognition test using sentence materials. Association for Research in Otolaryngology; 1994 February; St. Petersburg, Florida.
9. Nusbaum HC, Pisoni DB, Davis CK. Sizing up the Hoosier mental lexicon: measuring the familiarity of 20,000 words. Research on speech perception; Progress Report No. 10. Bloomington (IN): Indiana University Press; 1984.
10. Bench J, Bamford J. Speech-hearing tests and the spoken language of hearing-impaired children. London: Academic Press; 1979.
11. Levitt H. Transformed up-down methods in psychoacoustics. J Acoust Soc Am 1971;49:467–77.
12. Dubno JR, Horwitz AR, Ahlstrom JB. Speech recognition in noise at higher than normal levels: Decreases in scores and increases in masking. International Hearing Aid Research Conference; 2000; Lake Tahoe, California.

Submitted for publication August 30, 2002. Accepted in revised form April 1, 2003.