

Reliability of computer-automated hearing thresholds in cochlear-impaired listeners using ER-4B Canal PhoneTM Earphones

James A. Henry, PhD; Christopher L. Flick, BS; Alison Gilbert, MS; Roger M. Ellingson, MS;
Stephen A. Fausti, PhD

Department of Veterans Affairs (VA) Rehabilitation Research and Development National Center for Rehabilitative Auditory Research (NCRAR), VA Medical Center, Portland, OR; Department of Otolaryngology, Oregon Health and Science University, Portland, OR; VA Audiology Clinic, VA Medical Center, Portland, OR

Abstract—This paper describes the second phase of a study to determine test-retest reliability of hearing thresholds using a computer-automated technique with ER-4B Canal PhoneTM insert earphones. The first phase documented reliable hearing thresholds in 20 normal-hearing individuals. For this second phase, 20 individuals with cochlear hearing loss completed the same testing protocol as for phase one. During each of two sessions, hearing thresholds were obtained in one-third octave steps at 500 Hz to 16,000 Hz. The octave frequencies were immediately retested, followed by ear-tip reinsertion and further retesting at octave frequencies. Both groups showed overall good threshold reliability, with observable differences between groups. First, repeated testing resulted in improved hearing thresholds for the normal-hearing group, but not for the cochlear-loss group. Second, the normal-hearing group showed overall better response reliability, both within and between sessions, than the cochlear-loss group. These differences were small but consistent.

Key words: auditory threshold, hearing, reliability of results.

INTRODUCTION

In the development of a tinnitus measurement technique, we used computer automation to achieve the highest degree of interexaminer and intersession consistency in conducting the testing [1–5]. The overall goal of this

effort was to develop a standardized testing device that can be used efficiently in audiology clinics.

An important component of tinnitus testing is the precise measurement of hearing thresholds; thus, our computerized testing system has been programmed to perform such testing. The threshold testing protocol involved several unique features, however, that could potentially affect response reliability. First, the testing protocol was under full computer control and was directed by a testing algorithm that is designed to replicate human decision making. Second, thresholds were obtained to the nearest decibel, unlike conventional audiometric testing that uses 5 dB test increments. Third, the earphones used with the system were Etymotic Research

Abbreviations: ANOVA = analysis of variance, DLI = difference limen for intensity, ER = Etymotic Research, HL = hearing level, SD = standard deviation, SISI = Short Increment Sensitivity Index, SNHL = sensorineural hearing loss, SPL = sound pressure level.

This material is based upon work supported by the Department of Veterans Affairs Rehabilitation Research and Development (RR&D) Service (C891-RA and RCTR 597-0160).

All correspondence should be addressed to James A. Henry, PhD; National Center for Rehabilitative Auditory Research, VA Medical Center (NCRAR), P.O. Box 1034, Portland, OR 97207; 503-220-8262, ext. 57466; fax 503-402-2955; james.henry@med.va.gov.

(ER)-4B Canal Phone™ insert earphones that were designed for high-fidelity reproduction of music; i.e., they were not designed for audiometric testing. These earphones were deemed most suitable for an automated testing technique because of the need to use a single set of earphones for testing across a wide range of frequencies (500 Hz to 16,000 Hz). (Normally, audiometric testing at such high frequencies requires the use of special earphones.) For this testing system to be considered appropriate for clinical application, documentation of response reliability was essential. The unique features associated with this system necessitated a study to document the reliability of repeated threshold responses.

We previously reported a group of 20 normal-hearing individuals who were tested repeatedly, both within and between sessions, to determine the test-retest reliability of their threshold responses [6]. Response reliability was shown to be good and compared well to other studies that had evaluated hearing threshold reliability using conventional supra-aural earphones [7,8]. We also needed to determine if the same response reliability could be demonstrated with individuals who have hearing loss. This present study therefore was to replicate the previous study of normal-hearing individuals with a group of cochlear-impaired individuals.

METHOD

Subjects

Twenty individuals with cochlear hearing loss completed all testing. One ear was selected as the test ear for each subject, and only that ear was tested. Selection of the test ear was made randomly, based on alternating right and left ears for subsequent subjects (10 right ears and 10 left ears were tested). For the test ear, the subjects were required to have hearing thresholds exceeding 25 dB hearing level (HL) at two or more of the frequencies between 250 Hz and 8,000 Hz. All but two of the subjects did not have tinnitus. Subjects consisted of four females and 16 males (age range = 28 to 86 years; mean = 61.4; standard deviation [SD] = 14.8).

Instrumentation

The equipment and procedures used for this study have been described in detail [2] and were identical to the previous study using normal-hearing individuals [6] (also available on-line <http://www.vard.org/jour/01/38/5/>

[flick385.htm](http://www.vard.org/jour/01/38/5/flick385.htm)). Briefly, the system consisted of four major system components: (1) main controlling computer (Dell Dimension, 166 MHz Pentium central processing unit [CPU]) with a signal generator card (National Instruments, AT-DSP2200-128k) installed in an industry standard architecture (ISA) slot; (2) subject-response computer (Compaq Concerto 4/25); (3) signal processing module (custom built by Oregon Hearing Research Center, Oregon Health and Science University) used for signal mixing, attenuation, and earphone buffering; and (4) ER-4B Canal Phone insert earphones. An automated-calibration application was custom-programmed for the testing system.

Procedures

Testing procedures for the cochlear-impaired subjects were identical to those previously described for the normal-hearing subjects [6]. Each subject attended two test sessions separated by 1 to 7 days. Testing time was approximately 1 to 1 1/4 hours for Session 1 and less than 1 hour for Session 2.

The full range of test frequencies for the automated testing protocol included 500, 620, 800, 1,000, 1,260, 1,600, 2,000, 2,520, 3,180, 4,000, 5,040, 6,360, 8,000, 10,080, 12,700, and 16,000 Hz (17 test frequencies separated by one-third octaves). Three stages of testing occurred during each session—Stage 1: hearing thresholds at all 17 frequencies, Stage 2: repeat hearing thresholds at the six octave frequencies (500, 1,000, 2,000, 4,000, 8,000, and 16,000 Hz) without removing the ear tip from the subject's ear canal, and Stage 3: repeat Stage 2 following removal and reinsertion of the ear tip.

At each test frequency, the initial presentation level was 60 dB sound pressure level (SPL). Step sizes for tone presentation were progressively reduced through a series of three bracketing protocols: (1) up 10 dB, down 20 dB; (2) up 5 dB, down 10 dB; and (3) up 1 dB, down 2 dB. During testing at a single frequency, the reversal rules were the same for each of the three bracketing series. The only differences between the series were the step sizes and the number of responses required during ascending tracks. For the start of each bracketing series, the tone presentations increased in level until a response was obtained. The response reversed the direction of stimulus output to descending steps. During the descending track, another reversal occurred when the first "no response" occurred following the presentation of a stimulus. For the first series, only one response was required during

ascending tracks, and thus, two reversals. For the second and third series, two responses were required during ascending tracks, with four reversals.

Pulsed pure tones of 400 ms duration and a 50 percent duty cycle were presented in segments of 2.4 s each. Thus five tones were presented per segment unless the subject responded during the segment, which terminated the stimulus presentation. Time intervals between segments were randomized between 1 s and 4 s following a response. (Intervals were fixed at 1 s when no response occurred.) The time required for each ascending or descending track varied according to subject response time. The average time to complete individual tracks was about 15 s.

The computer started the tone presentations for the second bracketing series at a level of 10 dB below the response level obtained during the first series. Two responses were obtained for the last two bracketing series and were averaged for each series. The level of the first tone presented during the third series was the average response level obtained during the second series, less 2 dB. The “thresholds” obtained at each frequency, and subsequently reported, are based on the averages of the two responses obtained during the final bracketing series.

RESULTS

Conventional Hearing Thresholds

The **Figure** shows the mean conventional hearing thresholds, in dB HL, for the test ears (using a Virtual Model 320 audiometer). Mean hearing thresholds are shown separately for the subjects in the present study and for the normal-hearing subjects from the previous study [6].

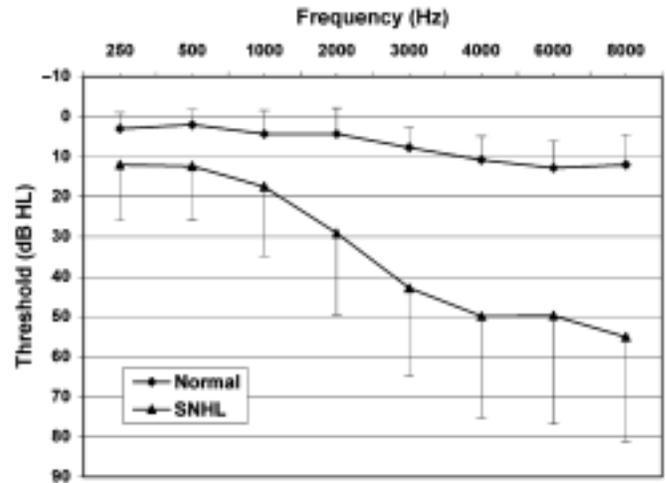


Figure.

Mean hearing thresholds in dB HL for test ears of subjects in normal-hearing and cochlear-loss groups.

Mean hearing thresholds were compared between the Virtual Model 320 audiometer and the automated system at the test frequencies that were common between systems: 500 Hz, 1,000 Hz, 2,000 Hz, 4,000 Hz, and 8,000 Hz. This comparison required use of the same decibel metric; thus, the dB HL thresholds obtained from the Model 320 were converted to dB SPL [9]. Note that, although the two earphones produced equal SPL in their respective calibration couplers, the sound pressure was not necessarily equal at the eardrum. It would thus be expected that differences in mean thresholds would be observed between the two systems even with the use of the same decibel metric. **Table 1** shows that these means differed by about 1 dB to 2 dB at 500 Hz, 1,000 Hz, and 2,000 Hz, and the differences increased to 4.1 dB at 4,000 Hz and to 7.3 dB at 8,000 Hz. To test for significance, we calculated t-tests. The multiple

Table 1.

Mean hearing thresholds, in dB SPL, obtained with two systems: (1) Virtual Model 320 audiometer with TDH-50P earphones and (2) automated system with ER-4B Canal Phone™ earphones.

Frequency (Hz)	Mean Hearing Threshold (dB SPL)		p Value *
	TDH-50P Supra-aural Earphones	ER-4B Canal Phone™ Earphones	
500	25.8	24.5	0.1553
1,000	25.0	27.0	0.0329
2,000	40.1	41.2	0.4526
4,000	60.3	56.2	0.0874
8,000	68.0	60.7	<0.0001

*Results of paired t-tests—only the means at 8,000 Hz differed significantly after corrections for multiple tests using Bonferroni's method.

t-tests required Bonferroni's corrections that dictated significance levels that should be used to interpret the results: $p < 0.01$ to correspond with 0.05 level for a single t-test. Only the means at 8,000 Hz were significantly different ($p < 0.0001$).

We obtained all further data provided in this paper from the automated system using the ER-4B insert earphones. The data are presented in the same manner as shown previously for the normal-hearing subjects to allow a direct comparison between the two groups [6].

Between-Session Reliability

Within-Group Reliability

Table 2 shows the across-subjects mean thresholds, in dB SPL. A repeated measures of analysis of variances (ANOVAs) was calculated on the six means at each

octave frequency, and a t-test was calculated on the two means at each nonoctave frequency. None of these ANOVAs or t-tests revealed significant differences based on Bonferroni's correction for repeated statistical tests (significance levels of $p < 0.008$ to correspond with 0.05 level for a single ANOVA and $p < 0.005$ to correspond with 0.05 level for a single t-test).

Within-Subjects Reliability

Table 2 reveals good threshold reliability, both within and between sessions for the combined group of subjects. Of primary interest, however, was the reliability of responses for individual subjects. For each subject, differences were calculated between thresholds obtained at each session (Session 2 threshold minus Session 1 threshold). The means of these differences shown in **Table 2** reflect the actual differences, thus indicating the direction

Table 2.

Means of hearing thresholds, in dB SPL, obtained with automated system from 20 subjects. Between Stages 2 and 3 during each session, ear tips from insert earphones were removed and reinserted. Also shown are means of individual differences in hearing thresholds between Session 1 and Session 2.

Freq (Hz)	Session 1			Session 2			Difference Scores			
	Stage 1 (All Freqs)	Stage 2 (Octave Freqs)	Stage 3 (Octave Freqs)	Stage 1 (All Freqs)	Stage 2 (Octave Freqs)	Stage 3 (Octave Freqs)	Mean of Actual Diffs (dB)	SD of Diff Scores (dB)	Pearson r^*	Mean of Abs Values of Diffs
500	24.5	25.2	23.3	24.3	25.0	23.0	-0.2	11.5	0.75	6.3
620	25.8	—	—	22.6	—	—	-3.3	9.9	0.83	5.2
800	26.4	—	—	28.0	—	—	1.6	7.4	0.93	4.2
1,000	27.0	26.2	25.0	27.1	24.9	25.2	0.1	2.8	0.99	2.2
1,260	30.1	—	—	29.0	—	—	-1.2	5.5	0.96	3.1
1,580	33.6	—	—	34.1	—	—	0.5	4.0	0.98	2.9
2,000	41.2	41.0	40.8	41.3	40.9	40.6	0.1	2.7	0.99	1.9
2,520	45.8	—	—	47.0	—	—	1.5	4.6	0.98	2.5
3,180	52.7	—	—	54.1	—	—	1.4	4.7	0.98	2.5
4,000	56.2	57.1	56.4	56.3	55.3	55.8	0.1	2.5	0.99	1.8
5,040	56.1	—	—	55.3	—	—	-0.8	3.8	0.99	2.6
6,340 [†]	57.4	—	—	57.0	—	—	0.5	3.7	0.99	2.7
8,000 [†]	60.7	60.2	60.3	59.2	60.0	59.5	0.0	5.4	0.98	3.7
10,080 [†]	68.7	—	—	67.2	—	—	1.3	8.9	0.94	6.2
12,700 [†]	73.9	—	—	75.9	—	—	-0.9	3.4	0.99	2.0
16,000 [†]	93.3	93.8	91.2	91.5	79.6	95.2	-1.8	5.0	0.52	3.5
Average	—	—	—	—	—	—	-0.1	5.4	0.92	3.3

*All correlation coefficients significant at $p < 0.0001$, except $p = 0.3186$ at 16,000 Hz.

[†]Ns reduced to 19 at 6,340 Hz and 8,000 Hz, 10 at 10,080 Hz, 7 at 12,700 Hz, and 6 at 16,000 Hz.

of the responses between sessions. For these subjects, the means of the actual differences varied randomly across frequencies between positive and negative. Thus, in contrast to all the means being negative for the normal-hearing subjects [6], no trend was found for the threshold responses obtained at the second session to be less than those from the first session for the cochlear-impaired group (Wilcoxon, $p > 0.05$). At the test frequencies where there was an N of 20 (500 Hz to 5,040 Hz), 44 percent of the differences were positive, 34 percent of the differences were negative, and 22 percent of the responses were equal across sessions. The SDs of the differences are shown in **Table 2**, where it can be seen that they ranged from 2.5 dB to 11.5 dB, with an average SD of 5.4 dB. Pearson product-moment correlations were also evaluated for each frequency, and the Pearson r 's are shown in **Table 2**. Each of these r values was ≥ 0.747 , and all coefficients were significant at $p < 0.0001$, except at 16,000 Hz ($r = 0.520$; $p = 0.3186$).

We also determined the average magnitude of the differences between sessions by calculating the absolute value of each subject's between-session threshold difference at each frequency. The means of these absolute values are shown in the last column of **Table 2** and ranged from 1.8 dB to 6.3 dB. The average difference, ignoring the direction of the differences, was 3.3 dB (as compared to 2.5 dB for the normal-hearing subjects).

Confidence Intervals for Difference Scores

To show the range of individual between-sessions differences in hearing thresholds, **Table 3** displays the confidence intervals for the difference scores. A total of 279 between-sessions threshold differences were found, and difference scores were grouped according to the indicated confidence intervals ranging from ± 1 dB to ± 20 dB. Of the 279 differences, 233 (83.5 percent) were within ± 5 dB, 268 (96.1 percent) were within ± 10 dB, and 271 (97.1 percent) were within ± 15 dB.

We expanded the assessment of between-sessions confidence intervals to analyze confidence intervals at the individual test frequencies. **Table 4** shows that, in general, between-sessions responses were most reliable at frequencies between 1,000 Hz and 8,000 Hz. **Table 4** also displays the corresponding percentages obtained from the normal-hearing subjects for a direct comparison of reliability between the normal-hearing and cochlear-impaired groups [6]. At most test frequencies, between-sessions response reliability is seen to be slightly better for the normal-hearing subjects.

Within-Session Reliability

Table 5 shows the within-session mean threshold differences for the sensorineural hearing loss (SNHL) subjects. Thresholds were obtained three times during each session, and these three trials are referred to as Stage 1, Stage 2, and Stage 3, as in **Table 2**. There were three possibilities to calculate differences between responses

Table 3.
Confidence intervals for between-sessions differences in hearing thresholds.

Interval (dB) in Which Between-Sessions Threshold Differences Occurred		Cumulative Number of Differences*	Percent of Differences [†]
From (\geq)	To ($<$)		
-1	1	81	29.0 (29.0)
-2	2	137	49.1 (52.4)
-3	3	186	66.7 (71.6)
-4	4	220	78.9 (84.9)
-5	5	233	83.5 (91.5)
-10	10	268	96.1 (98.1)
-15	15	271	97.1 (99.4)
-20	20	279	100 (100)

*Total number of between-sessions threshold differences = 279.

[†]Percent of differences for normal-hearing subjects from previous study shown in parentheses [6].

Table 4.

Confidence intervals for between-sessions differences in hearing thresholds. Each value represents percentage of responses that occurred for intervals indicated. Corresponding responses from normal-hearing subjects are shown in parentheses [6].

Interval (dB)		Frequency (kHz)															
From (≥)	To (<)	0.5	0.62	0.8	1.0	1.26	1.58	2.0	2.52	3.18	4.0	5.04	6.34	8.0	10.08	12.7	16.0
-1	1	26 (30)	32 (25)	20 (15)	20 (45)	30 (10)	25 (30)	40 (55)	45 (35)	35 (40)	35 (30)	35 (21)	16 (42)	21 (25)	9 (25)	57 (20)	17 (30)
-2	2	37 (45)	53 (65)	40 (45)	60 (70)	55 (20)	40 (60)	70 (70)	65 (60)	55 (60)	50 (60)	55 (63)	37 (59)	42 (45)	9 (40)	57 (30)	33 (53)
-3	3	47 (70)	74 (75)	60 (80)	75 (85)	70 (50)	60 (90)	75 (85)	75 (90)	85 (80)	75 (75)	65 (79)	63 (68)	53 (50)	27 (60)	71 (50)	83 (65)
-4	4	63 (85)	79 (90)	60 (95)	90 (95)	85 (85)	80 (90)	90 (95)	80 (90)	90 (85)	95 (95)	75 (95)	79 (95)	68 (70)	45 (65)	86 (65)	83 (71)
-5	5	68 (90)	84 (100)	75 (100)	90 (100)	85 (100)	80 (95)	90 (100)	85 (95)	90 (85)	95 (95)	90 (100)	89 (95)	68 (85)	64 (80)	86 (75)	83 (76)
-10	10	95 (100)	89 —	95 —	100 —	95 —	100 (95)	100 —	95 (100)	95 (95)	100 (100)	95 —	100 (100)	95 (100)	82 (90)	100 (100)	83 (100)
-15	15	95 —	89 —	95 —	— —	95 —	— (100)	— —	95 —	95 (100)	— —	100 —	— —	95 —	91 (100)	— —	100 —
-20	20	100 —	100 —	100 —	— —	100 —	— —	— —	100 —	100 —	— —	— —	— —	100 —	100 —	— —	— —

Note: Each percentage is based on 19 or 20 responses, except at 10.08, 12.7, and 16 kHz, which had *N*s of 10, 7, and 6, respectively.

Table 5.

Means of actual values of individual differences in hearing thresholds. All means shown are for various combinations of within-session differences.

Freq (Hz)	Session 1			Session 2		
	Stage 2 Minus Stage 1	Stage 3 Minus Stage 1	Stage 3 Minus Stage 2	Stage 2 Minus Stage 1	Stage 3 Minus Stage 1	Stage 3 Minus Stage 2
500	0.70	-1.15	-1.85	0.65	-1.35	-2.00
1,000	-0.85	-2.05	-1.20	-2.15	-1.90	0.25
2,000	-0.25	-0.40	-0.15	-0.40	-0.70	-0.30
4,000	0.85	0.20	-0.65	-0.95	-0.55	0.50
8,000*	0.5	-0.45	-1.05	-0.74	-1.32	-0.55
16,000*	0.50	-1.80	-1.60	3.00	3.67	0.667
Average	0.12	-0.82	-1.01	-0.51	-0.89	-0.36

**N* = 20 for all frequencies except 8000 Hz (*N* = 19) and 16,000 Hz (*N* = 6).

(Stage 1 versus Stage 2, Stage 1 versus Stage 3, and Stage 2 versus Stage 3). We calculated the difference values by subtracting earlier responses from later responses, and

Table 5 shows the means for each of these threshold differences. The mean differences are small, and no positive or negative trend was found (Wilcoxon, $p > 0.5$).

We examined the magnitude of the within-session threshold differences by determining the absolute value of each difference. Means of the absolute values are shown in **Table 6**. Ear-tip removal and reinsertion was completed between Stages 2 and 3 during each session, and threshold differences associated with ear-tip replacement are reflected by the means shown in the columns entitled “Stage 3 minus Stage 1” and “Stage 3 minus Stage 2.” The across-frequency averages, displayed in the bottom row, indicate that when ear tips were replaced, the average differences were within 1 dB of the mean differences when ear tips were not removed.

The potential effect of ear-tip replacement was evaluated by t-tests calculated at each frequency between the “Stage 2 minus Stage 1” means versus the “Stage 3 minus Stage 1” means. These t-tests were performed for the Session 1 and for the Session 2 pairs of means. Of the 12 t-tests that were conducted, only 2 resulted in $p < 0.05$. However, because of the number of t-tests that were performed, Bonferroni’s adjustment for repeated tests established a significance level of 0.004. At that level, significance did not occur at any of the test frequencies.

Confidence Intervals for Difference Scores

Table 7 provides confidence intervals for the within-subjects and within-sessions differences in hearing thresholds. Thresholds were repeated three times at the octave frequencies during each session. Therefore, three combinations of differences were available to be reported for the within-sessions repeated thresholds: (1) Stage 2 threshold minus Stage 1 threshold, (2) Stage 3 threshold minus Stage 1 threshold, and (3) Stage 3 threshold minus Stage 2

threshold. The first condition would reflect the ear tip being left in place, and the latter two conditions would reflect differences when the ear tips were removed and replaced between testing. We evaluated any effect of ear-tip replacement by comparing the percentages shown in **Table 7** for the three conditions in each session. Based on these data, reliability of responses was not reduced because of ear-tip replacement.

Table 7 also displays the corresponding data from the previous study with normal-hearing individuals [6]. Percentages for the normal-hearing subjects are shown directly below the corresponding percentages for the cochlear-impaired subjects in the present study. Although only slight differences were found between groups, this comparison reveals that the numbers are consistently better for the normal-hearing listeners.

DISCUSSION

Tinnitus is a subjective symptom, and worldwide efforts to quantify its “acoustic” aspects have not resulted in a widely accepted method for this purpose. We are responding to this need by developing computer-automated techniques that reliably measure tinnitus loudness and pitch and in a standardized format. The present study is part of a larger overall effort to develop computer-automated methodology to conduct clinical tinnitus assessment. Computer-automated testing has been shown to be effective in evaluating tinnitus loudness and pitch [2–5,10]. We are continuing to refine the automated system so as to provide a technique that can be used practically and efficiently in the

Table 6.

Means of absolute values of individual differences in hearing thresholds. All means shown are for various combinations of within-session differences.

Freq (Hz)*	Session 1			Session 2		
	Stage 2 Minus Stage 1	Stage 3 Minus Stage 1	Stage 3 Minus Stage 2	Stage 2 Minus Stage 1	Stage 3 Minus Stage 1	Stage 3 Minus Stage 2
500	3.10	2.45	3.45	5.25	5.05	2.70
1,000	1.65	2.65	2.10	2.75	2.40	1.95
2,000	0.95	2.00	1.65	1.60	3.20	2.50
4,000	1.45	1.60	2.25	2.45	2.35	2.00
8,000	1.84	3.25	2.53	2.00	4.26	2.95
16,000	1.83	2.60	2.40	3.67	4.33	2.67
Average	1.80	2.40	2.39	2.87	3.50	2.43

*N = 20 for all frequencies except 8,000 Hz (N = 19) and 16,000 Hz (N = 6).

Table 7.

Confidence intervals for within-sessions differences in hearing thresholds.

Interval (dB) in Which Within-Sessions Threshold Differences Occurred		Percent of Differences*					
		Session 1			Session 2		
From (\geq)	To ($<$)	Stage 2 Minus Stage 1	Stage 3 Minus Stage 1	Stage 3 Minus Stage 2	Stage 2 Minus Stage 1	Stage 3 Minus Stage 1	Stage 3 Minus Stage 2
-1	1	48.6 (51.3)	36.2 (40.2)	34.6 (49.6)	34.6 (33.3)	26.9 (29.1)	36.8 (30.8)
-2	2	66.7 (76.1)	60.0 (59.8)	67.3 (70.9)	66.3 (59.8)	52.9 (45.3)	57.5 (53.0)
-3	3	81.0 (88.9)	75.2 (77.8)	78.8 (82.9)	76.0 (80.3)	59.6 (69.2)	72.6 (69.2)
-4	4	92.4 (95.7)	88.6 (93.2)	88.5 (91.4)	81.7 (88.9)	75.0 (88.0)	83.0 (81.2)
-5	5	93.3 (97.4)	92.4 (94.9)	92.3 (94.9)	85.6 (94.0)	80.8 (92.3)	87.7 (88.9)
-10	10	98.1 (100)	97.1 (100)	96.2 (100)	95.2 (98.3)	94.2 (99.1)	96.2 (98.3)
-15	15	100	98.1	99.0	98.1 (99.1)	99.0 (99.1)	100 (99.1)
-20	20	—	100	100	100 (100)	100 (100)	— (100)
<i>N</i>	—	105	105	104	104	104	106

*Percent of differences for normal-hearing subjects from previous study shown in parentheses [6].

clinical setting to perform a comprehensive tinnitus assessment battery [11]. An essential component of the automated system is the capability to obtain hearing thresholds at each test frequency with 1 dB precision.

The present study is a follow-up to our previous study for which we evaluated the test-retest reliability of hearing thresholds in normal-hearing listeners using our computer-automated testing system [6]. The computer-automated system included a number of design features that might have affected response reliability, including (1) a custom algorithm for obtaining thresholds entirely by computer control [2], (2) threshold testing in 1 dB increments, (3) use of the ER-4B Canal Phone insert ear-

phones, and (4) removal and replacement of the ear tips from the insert-style earphones.

The previous study showed that the reliability of hearing thresholds obtained with this system was well within a clinically acceptable range [6]. For these normal-hearing individuals, the means of the absolute values of the between-sessions differences ranged from about 1 dB to 3 dB, with an average across frequencies of 2.5 dB. The between-sessions differences were slightly higher for the cochlear-impaired subjects in the present study, with an average difference across frequencies of 3.3 dB. Thus, for all test frequencies combined, the average between-sessions threshold difference for the cochlear-impaired subjects was 0.8 dB higher than for the normal-hearing

subjects. For the cochlear-impaired subjects, this degree of reliability would still be well within a clinically acceptable range.

Confidence intervals were calculated for the cochlear-impaired subjects, both for the between-sessions differences in thresholds (**Tables 3 and 4**) and for the within-sessions differences (**Table 7**). The confidence-interval data show the percentages of differences in thresholds that occurred within various intervals, from ± 1 dB up to ± 20 dB. A side-by-side comparison between the normal-hearing and cochlear-impaired subjects indicates that normal-hearing subjects provided consistently better response reliability but that the cochlear-impaired subjects provided reliability that was as good or better than other published reports [7,8,12]. Our data indicate that 83.5 percent and 96.1 percent of the between-sessions threshold differences for the cochlear-impaired subjects were within ± 5 dB and ± 10 dB, respectively.

In the previous study, the normal-hearing subjects revealed a significant trend for an improvement in hearing thresholds as a function of repeated testing, both within and between sessions [6]. Although statistically significant, the mean differences were small—amounting to average differences across frequencies of less than -1 dB. For the present group of cochlear-impaired subjects, no trend in the direction of responses was apparent—either within or between sessions—even though all testing conditions were identical between groups.

These findings may offer support for a learning or practice effect among normal-hearing listeners but not for cochlear-impaired listeners. A review of the literature reveals that this effect has not been confirmed. The preponderance of studies, however, supports the occurrence of a learning or practice effect for auditory thresholds. High and Glorig obtained repeated hearing thresholds [13] and found that the threshold for the first tone presented (1,000 Hz) was always higher (poorer) than the threshold for the next frequency and that repeated testing at 1,000 Hz generally resulted in an improvement in the threshold at that frequency. Studies by Burns and Hinchcliffe and Robinson and Whittle have both shown improvements in hearing thresholds of 1 dB to 2 dB over time that they ascribed to a learning effect [14,15]. Corso and Cohen showed improved thresholds both within and between sessions [16]. Hickling showed that response reliability improved both with listening practice and as

the interval between successive tests was reduced [17,18].

Other studies have shown no improvement in thresholds as a result of repeated testing. Erlandsson et al. showed no tendency toward between-sessions improvement of the summed hearing thresholds from 2,000 Hz to 8,000 Hz [19]. Brown reported that practice did not appear to have an effect on the measurements [20]. Additional studies reported only chance threshold fluctuations across repeated tests [21–23].

The present studies with our automated system suggest that a practice or learning effect does exist with repeated testing for auditory thresholds, although only for normal-hearing individuals. These results agree with the one study that systematically tested for this effect [24], presumably with normal-hearing listeners. There is, however, the unexplained lack of improved thresholds for the cochlear-impaired subjects who were tested in an identical fashion as for the normal-hearing subjects. A possible explanation for this discrepancy may be related to the difference limen for intensity (DLI).

It is well known that individuals with cochlear impairment have reduced DLIs relative to normal-hearing listeners [25,26]. The DLI phenomenon led to the development of the Short Increment Sensitivity Index (SISI) Test that requires patients to identify 1 dB shifts in intensity as evidence of cochlear pathology [27]. Test increments of 1 dB were later validated as optimal for use with the SISI [28]. Therefore, a possibility exists that, in the present study, the lack of improved auditory thresholds for the SNHL subjects was due to their reduced DLIs. That is, subjects with cochlear pathology may possess a greater inherent ability to identify the first audible tone 1 dB above threshold. Such ability could preclude any practice effect that might be associated with a lesser ability to discriminate between tones that vary in intensity by such a small amount. This line of reasoning, however, would also suggest that reliability of threshold responses should be better for listeners with cochlear impairment, which was not demonstrated by our findings. The between-sessions threshold differences were greater (by an average of 0.8 dB) for the cochlear-impaired group relative to the normal-hearing group. Because the difference was so small, however, it may reflect normal variability and not an actual difference between groups in response reliability. This question can only be resolved by further investigation.

In both the previous and present studies [6], hearing threshold data were obtained from all subjects with the ER-4B insert earphones and the TDH-50P supra-aural earphones. Although not the specific purpose of these studies, these data allowed a direct comparison of results at octave frequencies between the two types of earphones (see **Table 1** in both the previous and present papers). Some comments are necessary regarding technical aspects of these comparisons. There is a recommended technique for the transfer of reference equivalent threshold values from a standard reference earphone to an earphone of a different type. This procedure is spelled out in the current American National Standard Specification for Audiometers (ANSI) (S3.6-1996). This procedure is based upon probe-tube measurements that Corliss and Burkhard initially developed [29]. The procedure used here was an initial behavioral comparison of the ER-4B insert earphone and the TDH-50 supra-aural earphone. The findings from this pilot work are encouraging, and our next step is to use the probe-tube technique to establish reference equivalent threshold values for the ER-4B transducer. The difference in HL between the ER-4B and the TDH-50 earphones decreases systematically at the higher two frequencies (4,000 Hz and 8,000 Hz). This is true for both normal-hearing individuals and for patients with hearing loss in the present study [6]. Only probe-tube measurements will help us determine if this effect is real or an artifact of using, with the ER-4B earphone, reference equivalent threshold SPLs that have been established only for the ER-3A insert earphones.

Some studies have been conducted to describe normal variability of auditory thresholds for identifying true threshold shifts that would indicate noise damage or ototoxicity [30–32]. In general, these studies have observed that the majority of random threshold shifts fall within ± 5 dB. Dobie and Simpson et al. both reported SDs of the difference scores [30,32], which were comparable between their two studies.

The present study used 1 dB increments and showed that repeated thresholds differed, on average, by 3.5 dB or less (**Tables 2** and **6**). This average held true across the entire range of test frequencies for both within- and between-sessions differences. The SDs of difference scores averaged 2 dB to 3 dB better than the studies by Dobie and by Simpson et al. [30,32]. This could be explained by the 1 dB steps that were used in the final stage of testing for the present experiment, as well as the shorter time intervals involved. The good reliability shown

both in the previous study with normal-hearing listeners and in the present study with cochlear-impaired listeners suggests that this technique might offer an improvement to techniques that are used for serial monitoring.

In both the previous and the present study, the between-sessions differences in repeated thresholds never met the criteria that would have indicated ototoxicity according to the national guidelines for ototoxicity monitoring [33]. Therefore, this automated technique potentially may reduce false positive responses that would meet criteria for ototoxicity (or noise damage).

The ideal ototoxicity monitoring protocol would obtain daily measures of threshold sensitivity to determine, within one day, when ototoxic effects occur. Such information could only be obtained with an automated self-testing device that could be left in the hospital room or taken home by the patient. This technology could be accomplished using testing procedures similar to those described in this report. The device would also require the capability of downloading data via telephone lines to a central database for analysis. Development of such a device would provide data never before observed in patients receiving ototoxic drugs, that is, the day-to-day patterns of shifts in threshold sensitivity that occur during treatment.

CONCLUSION

The present study completes the evaluation of our automated system for its capability to obtain reliable measures of auditory sensitivity. In both groups of normal-hearing and cochlear-impaired listeners, good reliability of threshold responses was observed. The normal-hearing group had better reliability, but the differences observed between groups were small and would not be of any consequence clinically. This automated technique could thus be adapted for use as a standard hearing-test device. We plan to establish reference equivalent threshold values for the ER-4B insert earphones by conducting probe-tube measurements.

The automated technique continues to reveal reliable responses for use as a tinnitus-assessment instrument. In addition, the technique may have further potential for any application that requires serial monitoring of auditory thresholds. The automated system, with its use of insert earphones, could even be incorporated into a portable monitoring device that could be used for daily self-testing in any reasonably quiet environment.

ACKNOWLEDGMENT

We would like to thank Craig Dennis and David Lilly, PhD, for their helpful assistance in the preparation of this manuscript.

REFERENCES

- Henry JA, Meikle MB. Pulsed versus continuous tones for evaluating the loudness of tinnitus. *J Am Acad Audiol* 1999;10:261–72.
- Henry JA, Flick CL, Gilbert AM, Ellingson RM, Fausti SA. Reliability of tinnitus loudness matches under procedural variation. *J Am Acad Audiol* 1999;10:502–20.
- Henry JA, Flick CL, Gilbert AM, Ellingson RM, Silaski GC, Fausti SA. Fully-automated system for tinnitus loudness and pitch matching. In: Hazell J, editor. *Proceedings of the Sixth International Tinnitus Seminar*. Cambridge, UK: The Tinnitus and Hyperacusis Centre; 1999. p. 520–21.
- Henry JA, Fausti SA, Mitchell CR, Flick CL, Helt WJ, Ellingson RM. Computer-automated clinical technique for tinnitus quantification. *Am J Audiol* 2000;9:36–49.
- Henry JA, Flick CL, Gilbert AM, Ellingson RM, Fausti SA. Comparison of two computer-automated procedures for tinnitus pitch-matching. *J Rehab Res Devel* 2001;38:557–66.
- Henry JA, Flick CL, Gilbert A, Ellingson RM, Fausti SA. Reliability of hearing thresholds: Computer-automated testing with ER-4B Canal Phone Earphones. *J Rehab Res Devel* 2001;38:567–81.
- Studebaker GA. Intertest variability and the air-bone gap. *J Speech Hear Disord* 1967;32:82–86.
- Frank T. High-frequency hearing thresholds in young adults using a commercially available audiometer. *Ear Hear* 1990;11:450–54.
- ANSI. American National Standard: Specification for Audiometers. American National Standards Institute 1996; ANSI S3.6; 1996.
- Henry JA, Fausti SA, Mitchell CR, Flick CL, Helt WJ. An automated technique for tinnitus evaluation. In: Reich GE, Vernon JA, editors. *Proceedings of the 5th International Tinnitus Seminar 1995*. Portland: American Tinnitus Association; 1996. p. 325–26.
- Westphal BE, Ellingson RM, Schechter MA, Fausti SA, Henry JA. Development of a new clinical system for automated tinnitus evaluation. In: Patuzzi R, editor. *VII International Tinnitus Seminar Proceedings*. Perth, Western Australia: Physiology Dept., Univ. of Western Australia; 2002. p. 18.
- Larson VD, Cooper WA, Talbott RE, Schwartz DM, Ahlstrom C, DeChicchis AR. Reference threshold sound-pressure levels for the TDH-50 and ER-3A Earphones. *J Acoust Soc Am* 1988;84:46–51.
- High WS, Glorig A. The reliability of industrial audiometry. *J Audit Res* 1962;2:56–65.
- Burns W, Hinchcliffe R. Comparison of the auditory threshold as measured by individual pure tone and by Bekesy audiometry. *J Acoust Soc Am* 1957;29:1274.
- Robinson DW, Whittle LS. A comparison of self-recording and manual audiometry: some systematic effects shown by unpractised subjects. *J Sound Vibrat* 1973;26:41–62.
- Corso JF, Cohen A. Methodological aspects of auditory threshold measurements. *J Experim Psychol* 1958;55:8–12.
- Hickling S. The validity and reliability of pure tone clinical audiometry. *New Zealand Med J* 1964;63:379–82.
- Hickling S. Studies on the reliability of auditory threshold values. *J Audit Res* 1966;6:39–46.
- Erlandsson B, Hakanson H, Ivarsson A, Nilsson P. The reliability of Bekesy sweep audiometry recording and effects of the earphone position. *Acta Otolaryngol (Stockh)* 1980;366:99–112.
- Brown REC. Experimental studies on the reliability of audiometry. *J Laryngol Otol* 1948;42:487–524.
- Harris JD, Myers CK. Experiments on the fluctuation of auditory acuity. *J Gen Psychol* 1954;50:87–109.
- Herman G. Variability of the absolute auditory threshold—a psychophysical study. *J Acoust Soc Am* 1953;25:822.
- Munson WA, Wiener FM. Sound measurements for psychophysical tests. *J Acoust Soc Am* 1950;22:382–86.
- Zwislocki J, Maire F, Feldman AS, Rubin H. On the effect of practice and motivation on the threshold of audibility. *J Acoust Soc Am* 1958;30:254–62.
- Luscher E, Zwislocki J. A simple method for indirect monaural determination of the recruitment phenomenon (difference limen in intensity in different types of deafness). *Acta Otolaryngol (Stockh)* 1949;78(Suppl):156–68.
- Brunt MA. Tests of cochlear function. In: Katz J, editor. *Handbook of clinical audiology*. Baltimore: Williams & Wilkins; 1994. p. 165–80.
- Jerger J, Shedd J, Harford E. On the detection of extremely small changes in sound intensity. *Arch Otolaryngol* 1959;69:200–11.
- Sanders JW, Simpson ME. The effect of increment size on short increment sensitivity index scores. *J Speech Hear Res* 1966;9:297–304.
- Corliss ELR, Burkhard MD. A probe tube method for transfer of threshold standards between audiometric earphones. *J Acoust Soc Am* 1953;55:990–1003.
- Dobie RA. Reliability and validity of industrial audiometry: implications for hearing conservation program design. *Laryngoscope* 1983;93:906–27.

31. Brummett RE, Morrison RB. The incidence of aminoglycoside antibiotic-induced hearing loss. *Arch Otolaryngol Head Neck Surg* 1990;116:406–10.
32. Simpson TH, Schwan SA, Rintelmann WF. Audiometric test criteria in the detection of cisplatin ototoxicity. *J Am Acad Audiol* 1992;3:176–85.
33. ASHA (American Speech-Language-Hearing Association). Guidelines for the audiologic management of individuals receiving cochleotoxic drug therapy. *Asha* 1994;36 Suppl 12:11–19.

Submitted for publication May 21, 2002. Accepted in revised form December 20, 2002.