J Am Acad Audiol 17:350–366 (2006)

Evaluation of Equivalency in Two Recordings of Monosyllabic Words

Margaret W. Skinner*
Laura K. Holden*
Marios S. Fourakis†
John W. Hawks‡
Timothy Holden*
Jennifer Arcaroli§
Martyn Hyde**

Abstract

Thirty "new" lists of monosyllabic words were created at the University of Melbourne and recorded by Australian and American English speakers. These new lists and the ten original CNC lists (Peterson and Lehiste, 1962) were used during the feasibility study of the Nucleus Research Platform 8 Cochlear Implant System (Holden et al, 2004). Performance was similar across original and new lists for six implanted Australian subjects; for four implanted U.S. subjects, mean performance was 23 percentage points lower with the new than with the original lists. To evaluate differences between original and new lists for the American English recording, 22 CI recipients were administered all 40 CNC lists (30 new and 10 original lists). The overall mean word score for the new lists was significantly lower (22.3 percentage points) than for the original lists. Acoustic analysis revealed that decreased performance was most likely due to reduced amplitudes of certain initial and final consonants. The new CNC lists can be used as more difficult test material for clinical research.

Key Words: Acoustic analyses, clinical research, cochlear implant, consonants, monosyllabic words

Abbreviations: CI = cochlear implant; CNC = consonant-vowel nucleus-consonant; pps/ch = pulses per second per channel; RP8 = Research Platform 8

Sumario

Se crearon treinta "nuevas" listas de palabras monosilábicas en la Universidad de Melbourne, grabadas por hablantes de inglés australiano y americano. Se usaron estas nuevas listas y las diez listas CNC originales (Peterson y Lehiste,

*Washington University School of Medicine, St. Louis, Missouri; †University of Wisconsin-Madison, Madison, Wisconsin; ‡Kent State University, Kent, Ohio; §Cochlear Americas, Englewood, Colorado; **University of Toronto, Toronto, Ontario, Canada

Margaret W. Skinner, Department of Otolaryngology—Head and Neck Surgery, 660 South Euclid Avenue, Campus Box 8115, St. Louis, Missouri 63110; Phone: 314-362-7125; Fax: 314-362-7346; E-mail: skinnerm@ent.wustl.edu

Portions of this manuscript were presented as a poster at the VIII International Cochlear Implant Conference, Indianapolis, Indiana, May 10–13, 2004.

This research was supported by Grant R01 DC000581 from the National Institute on Deafness and Other Communication Disorders and Cochlear Americas.

1962) durante el estudio de factibilidad para el Sistema de Implante Coclear Nucleus Research Platform 8 (Holden y col., 2004). Los desempeños fueron similares con las listas originales y las nuevas para seis sujetos australianos implantados; para cuatro sujetos americanos implantados, el desempeño promedio fue 23 puntos porcentuales más bajo con las listas nuevas que con las originales. Para evaluar las diferencias entre las listas nuevas y las originales para la grabación en inglés americano, se le aplicaron todas las 40 listas CNC a 22 portadores de CI (30 listas nuevas y 10 originales). El puntaje global medio para las nuevas listas fue significativamente más bajo (22.3 puntos porcentuales) que para las listas originales: El análisis acústico reveló que el desempeño más pobre era posiblemente debido a amplitudes reducidas de ciertas consonantes iniciales y finales. Las nuevas listas CNC pueden ser utilizadas como un material de prueba más difícil para investigación clínica.

Palabras Clave: Análisis acústicos, investigación clínica, implante coclear, consonantes, palabras monosilábicas

Abreviaturas: CI = implante coclear; CNC = consonante-núcleo vocal-consonante; pps/ch = pulsos por segundo por canal; RP8 = Plataforma de Investigación 8

The CNC (consonant-vowel nucleus-consonant) monosyllabic word test (Peterson and Lehiste, 1962) has been widely accepted in North America for evaluating open-set speech recognition of adult cochlear implant (CI) recipients. The ten 50-word lists are part of the Minimum Speech Test Battery for Adult Cochlear Implant Users (Luxford et al, 2001). Each of these lists includes similar but not identical distributions of phonemes. When these lists were created in 1959 and revised in 1962, the phoneme distribution in each list was proportional to that found in American English CNC words (Thorndike and Lorge, 1944). The majority of words included were used frequently in written and spoken American English with only a few occurring with a frequency of less than five words per million words (word lists and frequency of each word's use is given in Peterson and Lehiste, 1962). Some infrequently used words in the CNC lists (e.g., salve, rout, mirth, and zeal) may be novel to CI recipients whose verbal skills and vocabularies are limited. Many other recipients may not understand some CNC words because of their impaired auditory processing capabilities. For both of these reasons, audiologists instruct CI recipients to respond by saying (or writing) what they hear even if they do not know some of the words.

Since 1995, many CI recipients have been evaluated with the CNC recorded lists recommended by Luxford et al, 2001 (originally recorded for Cochlear Corporation); relatively few have had scores that are below 10% or above 90%. Based on these results, this test has been of appropriate difficulty to monitor the performance of individuals and compare performance across recipients. With the advent of new cochlear implant systems and sound-processing strategies, recipients' performance is expected to improve with more scores reaching above 90%. When this so-called ceiling effect occurs, it is impossible to know how much better a recipient could perform. In this case, a recorded list of words that is more difficult to understand is required. Although performance can be degraded with noise, testing in quiet avoids the possibility of a particular recipient having a low score because he or she has more difficulty than other recipients understanding speech in noise.

Recently, 30 new lists of words were created at the University of Melbourne (McDermott, pers. comm., 2004) based on the same selection criteria as was used for the original ten CNC lists except that all 30 lists had identical distributions of phonemes. This distribution was the average of the occurrences of each phoneme across the ten original CNC lists. Words where the final phoneme is not explicitly pronounced in Australian English (e.g., "car") were excluded. The lists included seven words used in Australian English but not in American English. These words, their meanings, and the lists in which they occurred are as follows: "chook" (chicken; 20 and 28), "gorse" (type of bush; 23), "jape" (joke; 29), "kip" (nap; 26), "ruck" (set up in rugby; 10, 30), "toff" (snob; 16, 25), and "weir" (small dam or spillway; 10, 14). When the word "weir" is spoken, its sound is identical to the American English word "we're." For that reason, only six words were likely unknown to American English subjects. Recordings of these new CNC lists were made by Australian as well as American English talkers. For the American English recording, the new lists were spoken by the same male talker as for the original CNC word lists. These recordings of the original and new CNC lists were used during the feasibility study of the Nucleus Research Platform 8 (RP8) Cochlear Implant System in Australia and in the United States (Holden et al, 2004). Performance was very similar across original and new CNC lists for the six Australian subjects, but the four U.S. subjects' mean scores were 23 percentage points lower with the new lists than the original lists. Not all of the original and new lists were used for testing every subject in this feasibility study.

The purpose of this study was to determine whether the difference between CI recipients' recognition of the original and new CNC lists for the American English recordings was significant with a sufficiently large sample of subjects, and whether the intelligibility of the lists is equivalent among the original ten lists as well as among the 30 new lists. Performance was evaluated with all lists presented at 60 dB SPL to represent a normal conversational level. Warble-tone thresholds were obtained to determine whether subjects' speech processor programs made soft sounds audible.

METHOD

Subjects

Twenty-two adult CI recipients and five normal-hearing adults participated in the study. Demographic information for CI subjects is given in Table 1. Twenty-one of the CI subjects had postlinguistic and one had prelinguistic onset of profound hearing loss. For participation in the study, CI subjects were required to have scores of at least 20% (60 dB SPL presentation level) on the original CNC recordings at their most recent evaluation. Subjects ranged in age from 22 to 79 years with a mean of 56 years. Duration of severe-to-profound hearing loss prior to

Table 1. Biographical Information

Subject	Sex	Etiology	# Yrs. of Deafness	Age at Study	Length of CI Use (yrs.)
1	F	Unknown	.66	73	2.5
2	M	Unknown	43	43	1.5
3	F	Genetic	4	48	1
4	F	Unknown	2	78	2
5	F	Genetic	7	51	3
6	F	Unknown	2	54	1.5
7	F	Unknown	8	34	1
8	M	Genetic	2	69	4
9	F	Auto Immune	5	44	1.5
10	F	Otosclerosis	4	55	1
11	F	Genetic	6	53	3
12	M	Noise	9	75	4
13	M	Meningitis	2	60	1
14	M	Otosclerosis	11	77	16
15	M	Noise	2	79	4
16	F	Genetic	5	48	1
17	M	Genetic	1	51	7
18	F	Genetic	3	82	1.5
19	F	Otosclerosis	5	56	8
20	M	Ushers'	3	48	2
21	F	Genetic	6	22	3
22	F	Unknown	3	42	4

cochlear implantation ranged from .66 to 43 years with a mean duration of 6 years. Length of implant use ranged from 1 to 16 years with a mean length of 3 years. Sixteen subjects used the Nucleus 24 CI system, three used the Nucleus 22 CI system, and three used the Clarion II CI system. Table 2 provides specific details regarding each subject's cochlear implant system. The normal-hearing subjects ranged in age from 19–41 years with a mean age of 28 years and had no history of ear disease. Their pure-tone, air-conduction thresholds were ≤20 dB HL bilaterally from 250 through 8000 Hz.

Test Materials

Monosyllabic Word Tests

The ten lists of the original CNC Monosyllabic Word Test were spoken by a male talker of Midwestern American English and were digitally recorded in 1993 (compact disc version available in 1995 is identical to that recommended for the Minimum Speech Test Battery for Adult Cochlear Implant Users; Luxford et al, 2001). The recording was made in a studio recording booth with low ambient noise. A pop filter situated one inch in front of the microphone (ElectroVoice PL-20) served to distance the talker's mouth uniformly. The microphone output was fed to a preamplifier/signal processor (Symetrix 528 Voce Processor), although the processor portion of the circuitry was not activated. The signal was recorded on a digital audio (DAT) recorder (Tascam DA-30). Interconnections included: Canare cables and Neutrik gold-plated XLRs; all recording lines were balanced. Initial record levels were monitored visually by VU and maintained at approximately -2 dB. Words deemed too loud or too soft were re-recorded. The talker spoke so that every phoneme was clearly articulated; no carrier phrase was used. Target words were isolated from the original recording via a software editor (Sound Forge 5.0) and normalized by adjusting each token's overall amplitude to -24.6 dB RMS with peak amplitude values of ≤-2.2 dB. These recordings were made in the morning to minimize effects of fatigue on vocal effort. Three practice words were placed at the beginning of each list, and a single token of the word "ready" was edited in before each test word to alert the subject to listen. There were no repetitions of words across the ten lists.

The 30 new lists of CNC words were recorded in 2003 by the same talker and recording company with the same equipment, recording protocol, and wave file editing

Table 2. Ear of Implantation, Electrode Array, Depth of Insertion, Processor Type, Stimulation Rate (pulses-per-second per channel) and Speech Processing Strategy for Each of the Subjects

Subject	Ear	Array	Insertion Depth	Processor	Rate (pps/ch)	Strategy
1	R	Contour	Full	Sprint	1800	ACE
2	R	Contour	Full	Sprint	1800	ACE
3	R	CII	Full	BTE	2175	HiRes
4	R	Contour	Full	3G	1200	ACE
5	L	Straight	31 bands	Sprint	900	ACE
6	L	Contour	Full	3G	900	ACE
7	R	CII	Full	BTE	812.5	CIS
8	L	Straight	28 bands	Sprint	1800	ACE
9	L	Contour	Full	3G	1800	ACE
10	L	Contour	Full	3G	1800	ACE
11	R	Contour	Full	3G	900	ACE
12	R	Straight	23 bands	Sprint	1800	ACE
13	R	Contour	Full	3G	900	ACE
14	L	Nucleus 22	23 bands	Spectra	250	SPEAK
15	R	Straight	19 bands	Sprint	250	SPEAK
16	R	CII	Full	BTE	5156	HiRes
17	R	Nucleus 22	29 bands	Spectra	250	SPEAK
18	L	Contour	Full	3G	1800	ACE
19	L	Nucleus 22	16 bands	Spectra	250	SPEAK
20	L	Contour	Full	Sprint	250	SPEAK
21	L	Contour	Full	Esprit 24	250	SPEAK
22	L	Straight	30 bands	3G	1200	ACE

procedure. There were 572 new words included in the 30 lists (see Appendix A; six Australian English words not used in American English are shown in bold type. The word "weir" that sounds the same as the American English word "we're" is not shown in bold). Of the 500 words in the original ten lists, 355 (71%) were reused within the 30 new lists; 204 of these words were repeated twice; and 55 were repeated a third time. Within each of the three sets of ten lists (i.e., 1–10, 11–20, and 21–30), there were no repetitions of the original CNC words; however, there were repetitions across the three sets. The repeated words were uniquely spoken for each list. The word "ready" was placed before each test word. The male talker did not recall speaking the new lists differently from the original lists during recording; however, they were spoken ten years later.

Warble Tones

The warble tones (centered at .25, .5, .75, 1, 1.5, 2, 3, 4, and 6 kHz) were sinusoidal carriers modulated with a triangular function over the standard bandwidths recommended for use in the sound field by Walker et al (1984). The modulation rate was 10 Hz. A conversion from dB SPL to dB HL in the sound field was made using data obtained by Pascoe (1975) and Skinner (1988) at Central Institute for the Deaf (CID). For this conversion, the following values are subtracted from the dB SPL values: 15, 11, 9, 7, 5, 3, -2, -3, and 6 dB at .25, .5, .75, 1, 1.5, 2, 3, 4, and 6 kHz, respectively.

Equipment/Test Environment

The warble tones and recordings of the original and new CNC lists were presented to subjects in a double-walled sound attenuating booth (IAC; model 1204-A; 254 cm x 264 cm x 198 cm) through a loudspeaker placed at ear-level height at 0 degrees azimuth and 1.5 meters from the center of the subjects' heads in their absence. Test materials were presented via an IBM compatible, Pentium II computer that controlled a mixing and attenuation network (Tucker-Davis Technologies) to present sound through a power amplifier (Crown model D-150) and loudspeaker (Urei; model 809). Warble tones and words were stored as wave files on the hard disk. The sound pressure level (SPL) of the stimuli was measured with the microphone (Brüel and Kjaer, Model 4155) of the sound level meter (Brüel and Kjaer, Model 2230) at the position of the subject's head during testing. The overall SPL of the words was taken as the average of peaks on the slow, rms, linear scale. These measurements were made for the original and new lists. With software control, the overall presentation level of every list was the same, 60 dB SPL.

Procedures

This study followed a randomized-block (within subjects) design for which a different random ordering of the 40 lists (Lists 1–30: new lists; Lists 31–40: original lists) was assigned to each subject. Each list was presented only once, except that the last two lists (one new and one original list) assigned to each subject were presented as practice lists at the beginning of testing. These practice lists were used to familiarize subjects with the test materials and to minimize learning effects. Subjects used the same speech processor program or map and sensitivity and volume control settings for testing as they used in everyday life. At each of four test sessions spaced one week apart, ten CNC lists were presented. Warble-tone, soundfield threshold levels were obtained for each of the CI subjects during the first and third sessions to insure that all subjects had threshold levels at 30 dB HL or less across the frequency range and that no changes in hearing sensitivity occurred during the study.

For the past 20 years, speech recognition tests have been presented to cochlear implant recipients at a raised-to-loud level of 70 dB SPL. In this study, the words were presented at 60 dB SPL in quiet. A 60 dB SPL presentation level was chosen because it is more representative of conversational speech (Pearsons et al, 1976). In addition, a study by Firszt et al (2004) showed that a group of 78 CI users did not perform significantly poorer on CNC words (original recording) presented at 60 dB SPL (39%) than at 70 dB SPL (42%).

Data Analysis

Rationale for Number of Subjects

The inclusion of 22 CI subjects was based on the following analysis. List equivalence can be defined by a consensus criterion that no two lists of 50 words each shall differ in expected score by three items or more (i.e., ≤6%). A contrast of any two lists is appropriately modeled by the difference in binomial variates. Such contrasts are least sensitive at the point of maximum variance, in the region of 50% correct. The sampling distribution of the difference is well-approximated by a normal variate with variance 2 x 50 x p(1-p) where p is the Bernoulli parameter for each list item; for maximum variance, p is 0.5. The conventional computation based on the standard normal deviate yields a minimum number of CI subjects of 22 for a two-sided alpha of 0.05 and a minimum power of 0.8 to detect a true difference in size of at least 6%.

Statistical Analyses

The first null hypothesis was that group mean scores (across CI recipients) would be equivalent for the 30 new CNC lists. The second null hypothesis was that group mean scores (across CI recipients) would be equivalent for the ten original CNC lists. The third null hypothesis was that the group mean scores (across CI subjects and lists) for the new and original CNC lists would be equivalent. These hypotheses were tested for word and phoneme scores separately.

The first two hypotheses were tested using the Friedman test (two-way analysis of variance [ANOVA] by ranks). The third hypothesis was tested with the Wilcoxon test using paired differences between subject means for original lists versus new lists. The data also were analyzed using parametric General Linear Model (GLM) Repeated Measures ANOVA with Geisser-Greenhouse adjustments, with similar results. The conventional arcsine (root) transformation of raw scores to promote variance homogeneity was applied only to the parametric analyses. Interactions and extreme values were explored by examining residuals from the parametric ANOVAs for the original and new sets of lists. In this way, potentially deviant lists could be identified. Differences between pairs of lists were examined using the Tukey HSD test based on the Studentized range.

Acoustic Analyses

As stated above, 355 of the 500 words in the original lists were used within the new lists, and many of these words were repeated twice and some three times in the new lists (henceforth called "shared words"). This repetition afforded the opportunity to examine the subjects' responses to different tokens of the same word. We chose three lists for analysis on which the mean scores across CI subjects were identical. These original lists (numbers 32, 35, and 36) contained 108 words that occurred 159 times in the new lists. The 108 words as spoken in the original and new lists were transcribed, and all subjects' responses to these words (in both sets of lists) were converted from the orthographic representation that the subjects used to a phonetic representation using standard American English as a guideline. Percent correct identification scores were computed for each shared word in the original and new lists. Words for which the identification score was at least 20% lower (n = 71) in the new compared to the original lists were selected for acoustic analysis. The analysis involved the examination of the waveform of each word, measurements of closure, burst, voice onset time, and RMS (root mean square) amplitude of consonant related intervals. Furthermore, confusion matrices were created for initial consonants, medial vowels, and final consonants.

Word duration measurements were made of 49 words in original lists 32, 35, and 36 for which CI subjects had scores that were ≥20% poorer for the same words that occurred 70 times in the new lists. These measures were obtained to determine whether there was a significant decrease in word duration that could have contributed to the poorer performance.

RESULTS

Warble-Tone, Soundfield Thresholds

Group mean warble-tone, soundfield thresholds from 250 to 6000 Hz for the CI subjects are shown in Figure 1. These

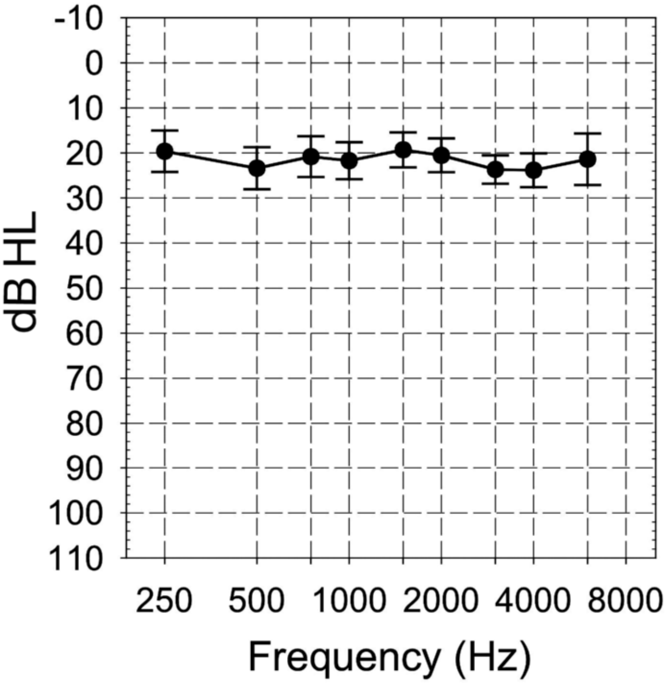

Figure 1. Mean warble-tone, soundfield thresholds (dB HL) from 250 through 6000 Hz across the CI subjects. Error bars are ±1 standard error of the mean.

thresholds range from 19.3 to 23.8 dB HL for individual frequencies; the group mean threshold across frequencies and CI subjects is 21.6 dB HL. According to Articulation Theory (e.g., Pavlovic et al, 1985), these thresholds suggest that audibility should not be a limiting factor in the recognition of these words presented at 60 dB SPL.

Individual CI Subjects' Performance across Word Lists

Each CI subject's mean word score across the 30 new lists as well as the 10 original lists is shown in Figure 2. The range of mean scores for the 30 new lists is 12.9 to 59.2% and for the original lists is 25.4 to 81.0%. No subject had lower than a 4% word score on any list. Scores on the original lists show that the performance criterion for subject selection of ≥20% was met, a wide range of performance was represented, and performance on the new lists was above chance for all subjects. Although the group mean difference in word score across subjects and lists was 22.3 percentage points between the new and original lists, there was

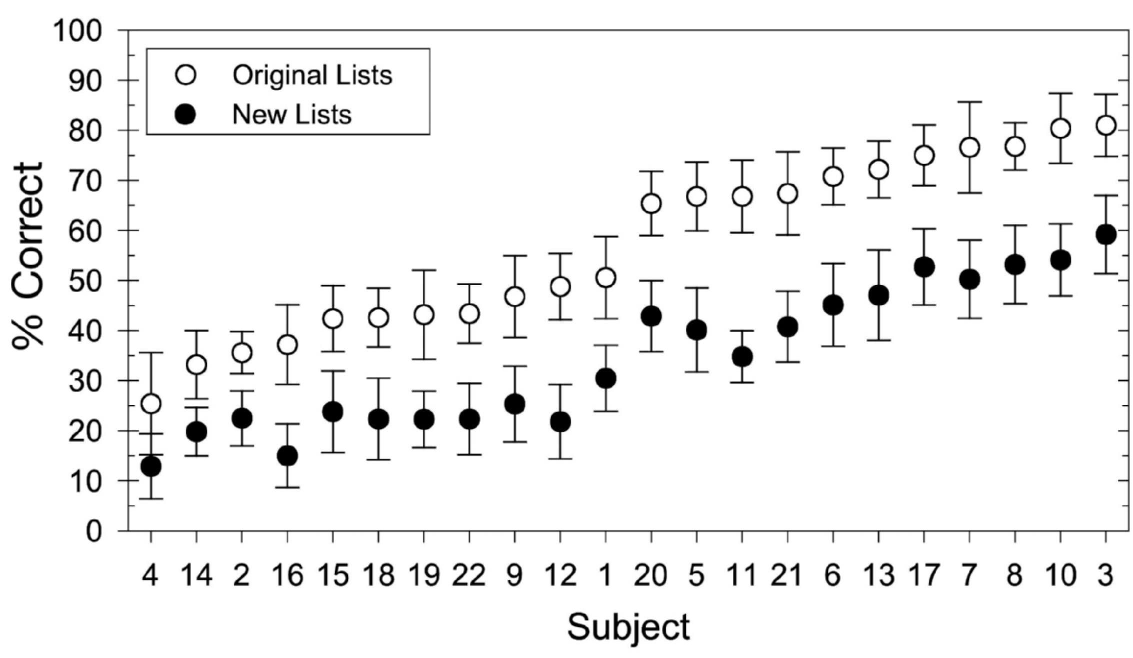

Figure 2. Individual subjects' word scores (percent correct) for the 30 new and 10 original lists for the 22 CI subjects. The filled and open circles represent the new and original list results, respectively. Error bars are ±1 standard error of the mean.

considerable individual variability in these score differences.

Each CI subject's mean phoneme score across the 30 new lists as well as the ten original lists is shown in Figure 3. The range of mean scores across the new lists is 44.0 to 81.8% and across the original lists is 54.0 to 92.7%. The group mean difference in phoneme score across subjects and lists is 13.3 percentage points, a difference that is considerably smaller than that for words (22.3 percentage points). The reason for this difference is that subjects often responded to two out of three phonemes correctly in the words. For example, the subject with the lowest word score on any list (4%) noted above had a phoneme score of 48%.

Group Performance of CI and Normally Hearing Subjects for Word Lists

The mean word score for each list across CI subjects (circles) and across normally hearing subjects (diamonds) are shown in Figure 4. For CI subjects, mean list scores range from 26.9 to 42.9% with an overall mean of 34.4% for the new lists; scores range from 51.3 to 62.9% with an overall mean of 56.7% for the original lists. The large standard error ranges are expected given the considerable variability in CI subjects' performance shown in Figure 2. The total range of mean list scores is larger with the new lists (16.0%) than for the original lists (11.6%).

Statistical analyses revealed that all null hypotheses were rejected. That is, the group mean word scores for the 30 new lists were not equivalent ($p < 0.0001$; Friedman nonparametric ANOVA); group mean scores for the ten original lists were not equivalent ($p < 0.0001$; Friedman nonparametric ANOVA); and scores for the new lists were not equivalent to the original lists ($p < 0.0001$; Wilcoxon test).

Visual inspection of the rank-ordered mean scores across subjects for each of the new lists (Figure 5) suggests that List 6 (rank 1) and List 26 (rank 30) might be substantively different from the remaining body of lists. The differences between these lists and their

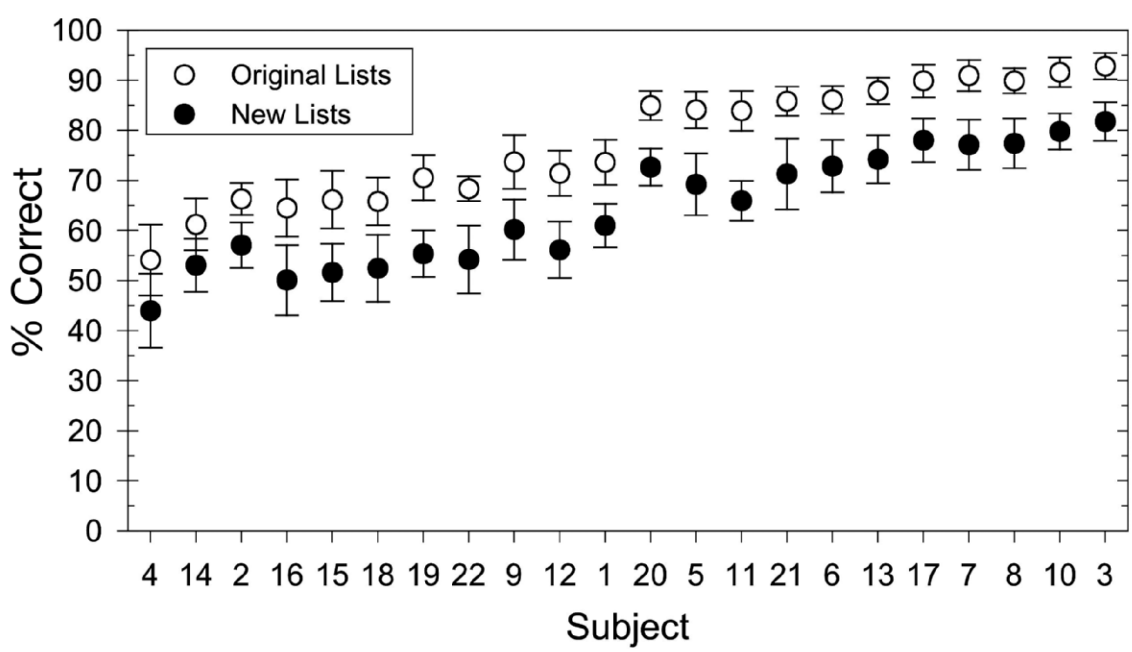

Figure 3. Individual subjects' phoneme scores (percent correct) for the 30 new and 10 original lists for the 22 CI subjects. The filled and open circles represent the new and original list results, respectively. Error bars are ±1 standard error of the mean.

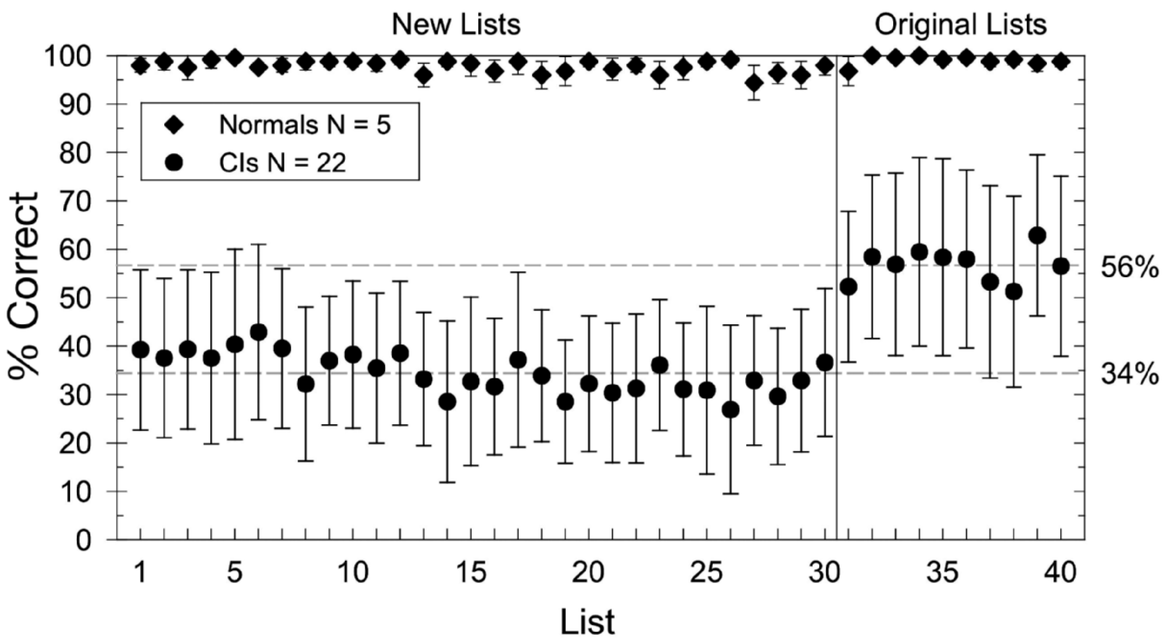

Figure 4. Group mean word score (percent correct) across CI subjects (circles) and normally hearing subjects (diamonds) for each of the new and original lists. Error bars are ±1 standard error of the mean. The upper and lower dashed gray lines represent the mean word score across original and new lists, respectively.

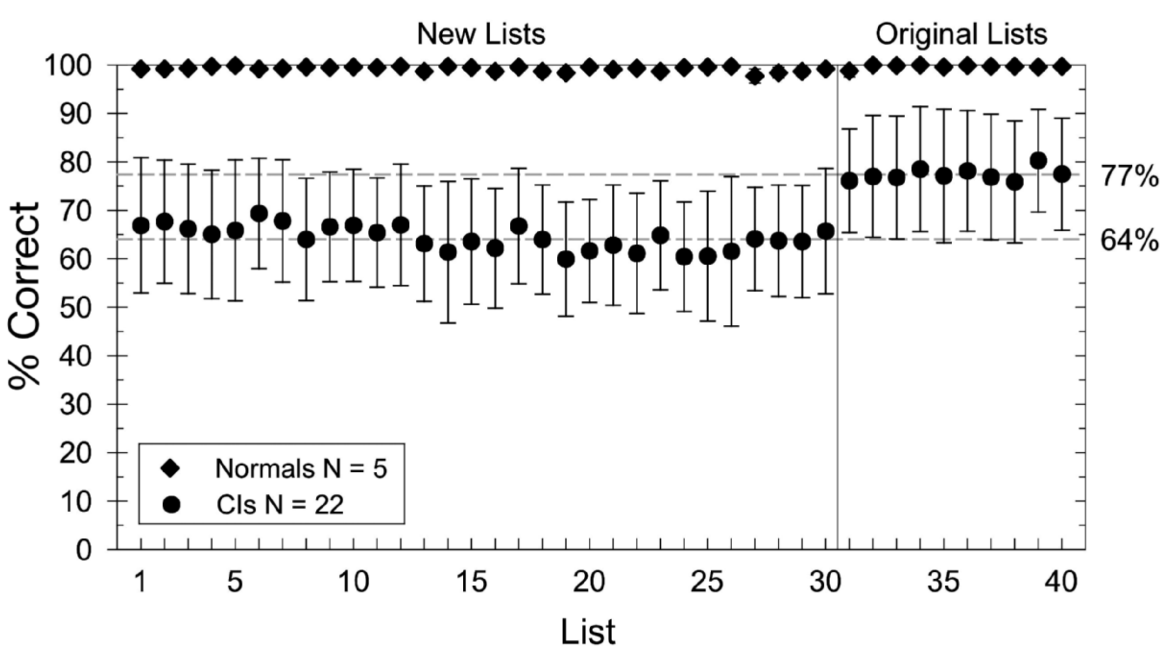

Figure 5. Group mean phoneme score across CI subjects for the new and original lists. The scores are rank ordered from lowest to highest score. Error bars are ±1 standard error of the mean. The upper and lower dashed gray lines represent the mean phoneme score across original and new lists, respectively.

nearest neighbors are larger than other differences between adjacent-ranked lists but did not achieve statistical significance on the conservative post hoc Tukey HSD test (with a computed 95% critical Studentized range difference of 5%). Nevertheless, the possibility that Lists 6 and 26 are genuinely deviant from the main body of lists should be considered carefully in the light of a variance component. Of the old lists, visual inspection of Figure 5 shows that List 39 (rank 40) is possibly deviant as well. While the extreme-ranked new lists are clearly, significantly different from each other on the basis of the global ANOVA, judgments about the "deviance" of individual lists from this data set are speculative and hypothesis generating. Confirmatory studies would be required to establish deviance of the suspect lists using a priori contrasts with nominal type I error rates.

The group-mean scores for the normal-hearing subjects on the new lists ranged from 94.4 to 99.6% with an overall mean of 97.7%. The scores on the original lists ranged from 96.0 to 100% with an overall mean of 99%. Although only five normal-hearing subjects were tested, the results showed near perfect performance for the original as well as the new lists. For this reason, additional normal-hearing subjects were not evaluated and further statistical analyses were not performed.

The mean phoneme score for each list across CI subjects (circles) and normal-hearing subjects (diamonds) are shown in Figure 6. For CI subjects, list scores range from 57.2 to 69.5% with an overall mean of 64.1% for the new lists; scores range from 75.9 to 80.3% with an overall mean of 77.4% for the original lists. As for the words, mean phoneme scores for new Lists 6 and 26 as well as for original List 39 were found to be possibly deviant.

Statistical analyses revealed that the following null hypotheses were rejected. The group mean phoneme scores for the new lists, for the new lists with Lists 6 and 26 removed, and for the new lists compared to the original lists were not equivalent at the 0.0001 level. In addition, the group-mean scores for the original lists were not equivalent at the 0.012 level. However, when List 39 was removed, the group mean scores across the remaining nine original lists were equivalent (i.e., the scores were no longer significantly different; p = 0.20). For normal-hearing subjects, group mean phoneme scores ranged from 98.2 to

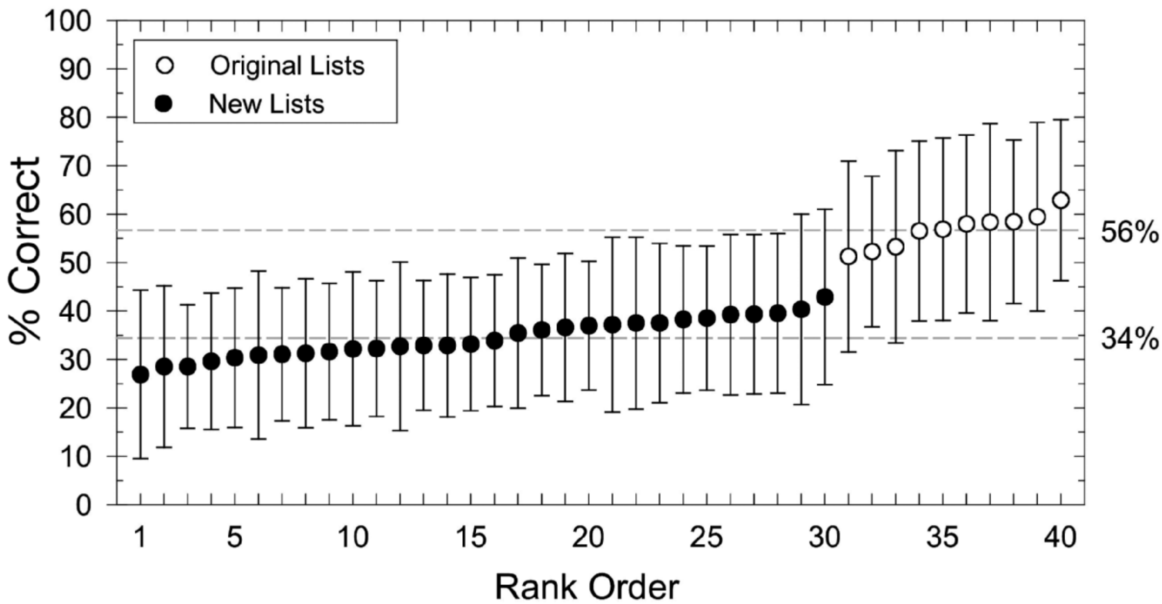

Figure 6. Group mean word score (percent correct) across CI subjects (circles) and normally hearing subjects (diamonds) for each of the new and original lists. The filled and open symbols represent the new and original list results, respectively. Error bars are ±1 standard error of the mean. The upper and lower dashed gray lines represent the mean word score across original and new lists, respectively.

100% across new and original lists with an overall mean of 99%. Statistical analyses of these scores were not performed because the scores were nearly perfect.

Comparison of Performance on Australian and American English Words

The six Australian English words in the new lists were unknown to the subjects in this study. The mean scores of the normal-hearing and CI subjects respectively on these words are as follows: "chook" (60%; 14%); "gorse" (40%; 0%); "jape" (100%; 0%); "kip" (100%; 0%); "ruck" (100%; 9%); and "toff" (100%; 0%). The normal-hearing subjects identified four of the six words with 100% accuracy; for the other two words, all but one of the errors were for one phoneme in each word. For the CI subjects, most of the errors also were for one phoneme in each word. Whereas the majority of these "unknown" words were identified accurately by the normal-hearing subjects, the CI subjects identified most of them incorrectly. For both groups of subjects, errors were mainly on one of the three phonemes in each word. It is important to consider these results in the comparison with American English words in the new lists with which both normal-hearing and CI subjects had difficulty. For example, normal-hearing and CI subjects had mean scores, respectively, on the following words: "bane" (40%; 0%); "fin" (20%; 0%); and "sheaf" (40%; 0%). That is, low performance occurred for both Australian and American English words.

Each list in which an Australian English word occurred included only one of these words. For these lists (i.e., 10, 16, 20, 23, 25, 26, 28, 29, and 30), the group mean scores (see Figure 4 and Appendix B) for the CI subjects were close to the mean across lists (34.4%) except for List 26. For this list, the group mean score for the normal-hearing and CI subjects was 100% and 26.9%, respectively. These results suggest that the inclusion of the nine tokens of the six Australian English words among a total of 1500 word tokens included in the 30 new lists did not significantly impact the overall performance of the CI subjects.

Comparison of Performance on Shared Words

Overall

The shared words were identified by the CI subjects with 58.8% accuracy when presented from the original recording and with 41.2% accuracy when presented from the new recording. The difference of 17.6 percentage points is consistent with the overall mean difference of 22.3 percentage points found between the new and original lists. Next, confusion matrices were examined to determine whether there was any specific class of sounds that was affected more than others, starting with broad classes such as initial and final consonants or vowels as well as more specific subclasses (i.e., stops and fricatives). Overall, initial consonants showed the highest decrement, with 73.5% correct identification in the original lists versus 56.1% identification in the new lists. Final consonants also showed a decrement with 75.3% identification in the original lists versus 63.3% in the new lists. Vowels also were affected but not as much as the consonants. Vowels in the original lists were identified with 84.4% accuracy versus 78.2% in the new lists. Because consonants were most affected, initial and final consonants were divided into the following classes: stops [p, b, t, d, k, g]; weak fricatives [f, v, θ, ð, h]; strong fricatives [s, ʃ, z, ʒ]; affricates [tʃ, dʒ]; nasals [m, n] and [ŋ] in final position only; and [w, y, r, l] in initial position and [r, l] in final position.

Initial Consonants

Table 3 shows the percent correct identification scores for the different consonant classes in initial position for the new and original lists. The largest decrements in score between the original and new lists were for stop (30 percentage points), affricate (19 percentage points), and weak fricative (16 percentage points) classes. Percent correct identification scores for each stop is shown in Table 3. The largest decrement in score between the original and new lists was for the alveolar stop [t] (40 percentage points) followed by [d] (36 percentage points), [k] (34 percentage points), and [g] (29 percentage points).

Table 3. Mean Percent Correct Identification Scores for the Different Consonant Classes in Initial Word Position for Both the Original and New CNC Lists

	Consonant Class					
List	Stops	Fricatives (Weak)	Fricatives (Strong)	Affricates	Nasals	w, r, y, l
Original Lists	73	50	90	92	74	75
New Lists	43	34	86	73	64	67
	Stops					
List	p	b	t	d	k	g
Original Lists	35	71	92	80	79	83
New Lists	13	53	52	44	45	54
	Weak Fricatives and Affricates					
List	f	v	θ	h	tʃ	dʒ
Original Lists	56	26	5	66	88	93
New Lists	43	16	7	47	76	68

Examination of the confusion matrices showed an increase of [h] error responses to words starting with stops, in both absolute numbers (original lists: 68; new lists: 277) as well as in percentage points (original: 8.1; new: 19.7). In addition, there was an increase in the number of "no response" errors to word initial stops (original: 2.5%; new: 8.9%). Although weak or absent bursts are very common in running speech, the CNC words were spoken in citation form following one-second silence between "ready" and the stimulus word. Correct identification of a stop consonant in the initial position is critically dependent on the presence of the burst, regardless of the place of articulation. When no burst is heard, then the presence of any aspiration noise may induce the perception of a weak fricative like [h]. This explanation is consistent with the large increase in error responses of [h] with the new lists. Acoustic analyses showed two probable reasons why the bursts were not correctly identified in the new lists. First, the bursts were on average 9 dB lower in amplitude, and second, some words had no bursts compared to the same words in original lists. For example, the words "dock" and "care" had strong initial bursts in the original lists and no discernible bursts in the new lists. These words were associated with a large decrement in score from the original to the new lists. The mean score for "dock" decreased from 82% to 5% and for "care" from 65% to 0% for the original and new lists, respectively.

Table 3 also shows the percent correct identification scores for each of the initial weak fricatives and affricates that occurred among the shared words. The voiced affricate [dʒ] and the voiceless glottal fricative [h] had the largest decreases in correct identification scores (25 and 19 percentage points, respectively). The voiced affricates were most often confused as stops, with 43 [t] and [d] responses, accounting for more than two-thirds of the error responses. This result suggests that subjects were not hearing the frication part of the affricate. The acoustic analysis did not show any consistent pattern that could be identified as a cause for this increase. The decrease in the correct identification of [h] resulted from an increase in "no responses" from 1.5% in the original lists to 11.3% in the new lists, as well as an increase in errant stop responses from 22.7% in the original lists to 29.5% in the new lists. Acoustic analysis revealed two factors that might account for this decrement. One was a slight average decrease of 3.2 dB in overall amplitude that was coupled with an average 29 msec decrease in duration of this phoneme (mean duration in old lists = 83 msec; in new lists = 54 msec).

Final Consonants

Table 4 shows the percent correct identification scores for each subclass of consonants in final position for the shared words. The greatest decrease in performance occurred for final nasals (24 percentage points) in the new lists compared to the

Table 4. Mean Percent Correct Identification Scores for the Different Consonant Classes in Final Word Position for Both the Original and New CNC Lists

	Consonant Classes					
List	Stops	Fricatives (Weak)	Fricatives (Strong)	Affricates	Nasals	r, l
Original Lists	80	37	90	86	68	88
New Lists	67	34	86	65	44	84
	Stops					
List	p	b	t	d	k	g
Original Lists	67	75	87	83	81	71
New Lists	58	42	70	68	71	69

original lists, followed by affricates (21 percentage points) and stops (13 percentage points). Excluding cases where the nasality feature was correctly identified but place of articulation was not, there appear to be two acoustic factors contributing to the decrement in correct identification of final nasals in the new lists. That is, these nasals were about 64 msec shorter (25%) and 7.6 dB weaker than in the original lists. This may account for the increase of errant stop responses from 8% to 11% and [v] responses from 2.4% to 6%. In addition, place of articulation errors increased from 10% in the original lists to 15.5% in the new lists. The decrease in affricate identification was due to an increase of [t] and [d] error responses which occurred only 6% of the time with the original lists but 18% of the time with the new lists. These error responses were not supported by any consistent acoustic differences between shared words in the two sets of recordings.

Table 4 shows the percent correct identification scores for each stop separately. It can be seen that the voiced bilabial stop [b] was most affected; that is, it was identified with 75% and 42% accuracy in the original and new lists, respectively. For example, acoustic analysis of the word "robe" showed that the final [b] in this word in the original recording exhibited very strong voicing during closure and a very strong burst that was 18 dB stronger than the burst for the final [b] in the same word in the new recording. As noted for initial stops, the decrease in score was accompanied by much lower burst amplitude in the new versus the original lists.

Word Duration Measurements

The mean duration of 49 of the shared words in original Lists 32, 35, and 36 was 568 msec (SD = 74.0 msec). There were 70 occurrences of these words in the new lists; their mean duration was 521 msec (SD = 77.5 msec). Thus, the words in the new lists were 8.3% shorter than those in the original lists. A pairwise t-test showed that there was a significant difference in mean duration ($t [69] = 5.646$, $p < .01$). However, regression analysis attempting to predict the score difference from the durational difference yielded a nonsignificant standard beta coefficient. Therefore, it is unlikely that the shortening of these words in the new lists is responsible for the 22.3 percentage point decrease in overall performance with the new lists.

Summary

The decrease in performance for words occurring in both the new and original lists is likely due to the reduced amplitudes for certain classes of consonants for which amplitude can be an important perceptual cue. Most affected were stops, affricates, and weak fricatives in initial position and nasals, affricates, and stops in final position.

DISCUSSION

The decrease in group mean score (22.3 percentage points) with the new versus the original CNC lists for the CI subjects in this study agrees closely with the 23 percentage point mean decrease in score by U.S. recipients of the Nucleus RP8 CI System (Holden et al, 2004). Acoustic analysis coupled with analysis of confusion matrices of words occurring in both original and new lists suggests that the major reason for this decrement in score for the CI recipients was a decrease in amplitude of stop bursts, affricates, weak fricatives, and nasals in the new lists. That is, these phonemes were incorrectly identified more often on the new lists because their cues were either absent or weak. This difference is seemingly not related to the recording equipment, recording procedure, or editing software because all were identical for the original and new lists. In addition, it cannot be argued that the original and new lists were spoken at speaking rates that differed significantly for the following reasons. First, two studies of the effect of clear versus conversational speaking rates on speech recognition (Picheny et al, 1985 and 1986) revealed that words were twice as long (100 versus 200 words per minute) for the clear rate. Second, analyses in the present study showed that the duration of shared words in the original lists were only about 8% longer than those in the new lists. This difference in rate is much smaller than that found by Picheny et al. It is improbable that word duration was an important reason for the decrease in performance with the new lists. The same talker was used for recording both the original and new lists. When the new lists were recorded ten years after the original lists, it appears that he spoke some consonant classes (listed above) with substantially reduced amplitude when recording the new lists. Finally, the normal-hearing subjects obtained scores that were well above 90% on original and new lists, indicating that the decrease in phoneme amplitude did not affect their performance nearly as much as it did for the CI subjects. If the new lists are to be used with CI recipients, one must take into account the phonemes with reduced amplitude and their effect on performance. As described above, the inclusion of six Australian English words unknown to the American English speaking subjects did not appear to have a substantive negative impact on overall performance on the new lists.

While the initial intent of creating the 30 new lists was to generate lists of equivalent difficulty to the original ten lists, the result is that the new lists are more difficult than the original. However, these new lists may be useful as a more difficult speech recognition test for CI recipients. The CI subjects in this study were chosen to demonstrate a wide range of speech recognition performance. Individual mean word scores across the ten original CNC lists ranged from 25% to 81%. However, half of the subjects in this study had mean word scores (across original lists) of greater than 65% and mean phoneme scores of greater than 80%. As technology continues to advance, speech recognition will likely continue to improve as well. Thus, a more challenging monosyllabic word test is needed to assess CI recipients' ability to understand speech.

Furthermore, in clinical research, it is important to minimize sources of test-score variability as well as to minimize learning effects. To accomplish this, a number of equally intelligible lists are needed. Statistical analysis revealed that the ten original CNC lists were not equally intelligible nor were the 30 new lists. However, by averaging mean scores for the pair of lists with the lowest and highest scores and working toward the middle lists pair-wise (e.g., Lists 26 and 6, Lists 14 and 5, Lists 19 and 7, etc. [see Appendix B]) will provide two-list (100 word) combinations that should yield very homogeneous scores. Using this method gives 15 pairs of new lists with scores ranging from 33.5% to 34.9% and five pairs of the original lists with scores ranging from 55.9% to 57.4% (see Appendixes B and C). This strategy of pairing observed extreme-ranked lists is likely to produce more homogeneous combined lists, because the differences between the extreme-ranked lists are highly significant statistically and are not attributable to chance. However, the strategy capitalizes on chance as expressed by random errors in this particular dataset. The homogeneity of the specified combined lists would be established more definitely by confirmatory studies.

CONCLUSION

For the CI subjects, the overall mean word score (across subjects and lists) for the 30 new CNC lists created at the University of Melbourne and recorded in the United States is significantly lower (22.3 percentage points) than the same score for the ten original CNC lists. Five normal-hearing subjects had near perfect scores for both the original and new lists. Statistical analysis revealed that the new lists are not equally intelligible nor are the original lists; however, pairing extreme-ranked lists gives 15 (100-word) pairs of new lists and five (100-word) pairs of original lists that yield very homogeneous scores (range: <3% for both sets), which can be used for clinical assessment as well as for research.

Acoustic analysis of the same words occurring in both the new and original lists revealed that the difference in performance between the two sets of lists is most likely due to reduced amplitudes of certain initial and final consonants in the new lists despite the use of identical overall presentation level for original and new lists.

Acknowledgments. This research was supported by Grant RO1 DC000581 from the National Institute on Deafness and Other Communication Disorders. Cochlear Americas funded the recording of original and new word lists, evaluation of normal-hearing subjects' performance, and generation of phonetic transcriptions of CI subjects' responses to "shared words." Appreciation is expressed to the 27 subjects who graciously gave their time and effort to participate in this study; to Jon Shallop who was the talker for the original and new CNC recordings; to Kennet Plantell and Bjarne Blume of AVCOM Media Productions, Inc. who made the original and new recordings; and to Brenda Gotter, Christine Brenner, and Sallie Vanderhoof for assistance in data collection and entry. This research was approved by the Human Studies Committee at Washington University School of Medicine.

REFERENCES

Firszt JB, Holden LK, Skinner MW, Tobey EA, Peterson A, Gaggl W, Runge-Samuelson C, Wackym, PA. (2004) Recognition of speech presented at soft to loud levels by adult cochlear implant recipients of three cochlear implant systems. *Ear Hear* 25(4):375–387.

Holden LK, Plant K, Skinner MW, Arcaroli J, Whitford L, Nel E, Cowan R. (2004) Evaluation of new coding strategies in the Nucleus research platform 8 system. Poster presentation at the VIII International Cochlear Implant Conference, Indianapolis.

Luxford W, Ad Hoc Subcommittee. (2001) Minimum speech test battery for postlinguistically deafened adult cochlear implant patients. *Otolaryngol Head Neck Surg* 124:125–126.

Pascoe DP. (1975) Frequency responses of hearing aids and their effects on the speech perception of hearing-impaired subjects. *Ann Otol Rhinol Laryngol* 84 (Suppl. 23):1–40.

Pavlovic CV, Studebaker GA, Sherbecoe RL. (1985) An articulation index based procedure for predicting the speech recognition performance of hearing-impaired individuals. *J Acoust Soc Am* 80:50–57.

Pearsons KS, Bennett RL, Fidell S. (1976) Speech levels in various environments. Bolt, Beranek and Newman Report No. 321. Canoga Park, CA: Bolt, Beranek and Newman.

Peterson GE, Lehiste I. (1962) Revised CNC lists for auditory tests. *J Speech Hear Disord* 27:62–70.

Picheny MA, Durlach NI, Braida LD. (1985) Speaking clearly for the hard of hearing I: Intelligibility differences between clear and conversational speech. *J Speech Hear Res* 28:96–103.

Picheny MA, Durlach NI, Braida LD. (1986) Speaking clearly for the hard of hearing. II: Acoustic characteristics of clear and conversational speech. *J Speech Hear Res* 29:434–446.

Skinner MW. (1988) *Hearing Aid Evaluation*. Englewood Cliffs, NJ: Prentice Hall.

Thorndike EL, Lorge I. (1944) *The Teacher's Word Book of 30,000 Words*. New York: Bureau of Publications.

Walker G, Dillon H, Byrne D. (1984) Sound-field audiometry: recommended stimuli and procedures. *Ear Hear* 5:13–21.

Appendix A. Words (n = 572) in New Lists That Were Not in the Original CNC Lists

BADGE	COIL	FAWN	HIS	LID	NIT	REEL	SIEVE	VERB
BAIT	COME	FEAR	HIVE	LIED	NOD	RHYME	SIGHT	VERGE
BAN	CON	FED	HOARD	LIES	NODE	RICH	SIGN	VERSE
BANE	COOK	FEED	HOG	LIKE	NOOK	RIFE	SILL	VET
BANG	COOP	FEIGN	HONE	LIME	NOON	RIGHT	SIP	VICE
BARE	COP	FELL	HOOK	LINE	NORM	RILE	SIRE	VIED
BARGE	COPE	FETCH	HOOT	LIT	NOT	RIM	SITE	VILE
BARK	CORD	FIB	HORN	LOAM	NOTCH	RIP	SOAK	VOGUE
BARN	CORK	FIGHT	HORSE	LOBE	NUB	RISE	SOIL	VOWED
BASE	CORN	FILE	HOSE	LODGE	NUN	ROAM	SOOT	WAD
BAT	COULD	FIN	HUB	LOIN	ONE	ROARED	SORT	WADE
BATCH	COURSE	FIZZ	HUFF	LOLL	PAIN	ROCK	SOUP	WAGE
BEAD	COURT	FOIL	HUG	LOOM	PAIR	ROD	SUIT	WAIF
BEEF	COVE	FORCE	HUM	LORD	PAL	RODE	SUM	WAIT
BEER	COWS	FORM	HUNG	LOUSE	PARCH	ROGUE	SUP	WALK
BEES	CUD	FOURTH	HUTCH	LOUT	PARK	ROOK	SURF	WALL
BIDE	CUFF	FOYER	HYPE	LOWER	PART	ROVE	SURGE	WANE
BILE	CULL	FUN	JAB	LOYAL	PAT	ROYAL	SWORD	WARD
BILL	CURB	FURL	JACK	LUG	PAWN	RUB	TAB	WARM
BIRD	CURL	FURS	JAG	LURCH	PAWS	**RUCK**	TACK	WARP
BOARD	CURSE	GAFF	**JAPE**	LURE	PEACE	RUDE	TAG	WART
BOG	CURT	GALL	JEEP	LURK	PEAS	RUG	TAIL	WAVE
BOON	DAD	GAME	JEER	LUSH	PECK	RULE	TAME	WAYS
BOOTH	DALE	GAPE	JEWEL	MACE	PEEK	RUNG	TAN	WEAN
BORN	DAME	GATE	JIG	MAD	PEEP	RUSE	TAP	WEAR
BOSS	DARE	GAUGE	JIVE	MADE	PEER	RUT	TART	WED
BOUT	DARK	GAVE	JOG	MAIL	PEG	SAGE	TAUGHT	WEDGE
BOWEL	DARN	GEL	JOWL	MAIM	PEN	SALE	TEACH	WEIR
BUCK	DASH	GIG	JOYS	MARE	PERT	SANG	TEAK	WHARF
BUDGE	DAUB	GILL	JUNE	MARK	PET	SASH	TEAR	WHERE
BUFF	DEAF	GIRD	JUT	MARSH	PHASE	SAT	TEETH	WHIFF
BULL	DEBT	GIRTH	JUTE	MASH	PIES	SAUCE	TEN	WHIRL
BUS	DEED	GNAT	KEEL	MASS	PIKE	SAWN	TERSE	WHIZZ
BUYS	DEEM	GNOME	KILL	MAT	PIN	SAYS	THEME	WHO'D
CAD	DEER	GOAD	KIN	MAUVE	PIP	SEAL	THICK	WICK
CAKE	DEIGN	GOAT	**KIP**	MAZE	PIPE	SEEM	THIEF	WIDE
CANE	DELL	GOD	KISS	MEAL	PIT	SET	THIGHS	WINE
CAP	DIAL	GOOF	KIT	MEN	POKE	SEWED	THONG	WIPE
CARVE	DICE	GOON	KNACK	MERE	PORCH	SEWER	THORN	WISE
CASE	DIED	**GORSE**	KNAVE	MESH	PORK	SEWN	THOUGHT	WOKE
CHAFE	DILL	GOUT	KNEEL	MIGHT	PORT	SHALE	THUD	WOOED
CHAFF	DINE	GUILE	KNOWN	MILE	POUCH	SHAME	THUG	WOOL
CHAP	DIRT	GUISE	LACE	MIME	POURED	SHAPE	TIDE	WORN
CHARM	DOES	GUM	LAD	MISS	POUT	SHARE	TIED	WORSE
CHASE	DOLL	GUSH	LAID	MITT	PUP	SHARK	TIER	WORTH
CHEAT	DOME	GUT	LAIR	MOAN	PURRS	SHAVE	TIFF	WRAP
CHESS	DONE	GYM	LAME	MOAT	PUS	SHEAF	TIGHT	WREATH
CHEWED	DOT	HACK	LANE	MOOSE	PUSH	SHEAR	TILE	WREN
CHIDE	DOTE	HAD	LASS	MOOT	PUT	SHEATH	**TOFF**	WRETCH
CHIME	DOWEL	HAG	LATCH	MOPE	PUTT	SHED	TON	WROTE
CHIP	DUB	HANG	LAWS	MOTH	RACE	SHEET	TONG	YACHT
CHIRP	DUCK	HARK	LAYER	MOWER	RACK	SHELL	TONGUE	YAK
CHOKE	DUD	HARM	LAZE	MUCK	RAKE	SHIN	TORCH	YAP
CHOOK	DUG	HAT	LEAF	MULL	RAM	SHIRE	TORN	YARD
CHOP	DUKE	HATCH	LEASH	MUM	RAN	SHIRK	TOUGH	YARN
CHOSE	DUTCH	HAUL	LEDGE	MURK	RANG	SHOES	TOUR	YAWN
CHUCK	DYKE	HAWK	LEECH	MUSH	RAPE	SHOOK	TOYS	YEAR
CHUG	FAD	HEAR	LEG	NAB	RARE	SHORES	TUB	YELL
CHURCH	FAIR	HEARSE	LESS	NAN	RASH	SHORT	TUCK	YEN
CHURN	FAME	HEARTH	LET	NAPE	RATE	SHOVE	TUG	YET
COACH	FANG	HEATH	LEWD	NAUGHT	RAVE	SHOWER	TURF	
COD	FARM	HEIGHT	LIAR	NERVE	REAM	SHOWN	TYPE	
CODE	FAT	HELL	LICE	NIB	REEF	SHUN	USE	
COG	FATE	HERB	LICK	NIL	REEK	SICK	VEER	

Note: Six Australian English words not used in American English are shown in bold type.

Appendix B. New Lists Rank Ordered from Lowest to Highest Score in the First Column

New Lists

List Number	Mean Score	Paired Lists	Mean Score
26	26.9	26, 6	34.9
14	28.5	14, 5	34.5
19	28.5	19, 7	34.0
28	29.6	28, 3	34.5
21	30.4	21, 1	34.8
25	30.9	25, 12	34.7
24	31.1	24, 4	34.3
22	31.3	22, 2	34.4
16	31.6	16, 17	34.4
8	32.2	8, 9	34.6
20	32.3	20, 30	34.4
15	32.7	15, 10	34.6
27	32.9	27, 23	34.5
29	32.9	29, 11	34.2
13	33.2	13, 18	33.5
18	33.9		
11	35.5		
23	36.1		
10	36.5		
30	36.6		
9	37.0		
17	37.2		
2	37.5		
4	37.5		
12	38.5		
1	39.3		
3	39.4		
7	39.5		
5	40.4		
6	42.9		

Note: Mean word scores across subjects for each of the new lists are shown in the second column. Lists with extreme scores are paired (third column), and their mean scores are shown in the fourth column.

Appendix C. Original Lists Rank Ordered (first column) from Lowest to Highest Mean Word Score across Subjects (second column)

Original Lists

List Number	Mean Score	Paired Lists	Mean Score
38	51.3	38, 39	57.1
31	52.3	31, 34	55.9
37	53.3	37, 32	55.9
40	56.5	40, 35	57.4
33	56.9	33, 36	57.4
36	58.0		
35	58.4		
32	58.5		
34	59.5		
39	62.9		

Note: Lists are paired by extreme scores (third column), and their mean score is shown in the fourth column. Original Lists 31–40 are identical to Lists 1–10 on the Minimum Speech Test Battery for Postlinguistically Deafened Adult Cochlear Implant Users.