An Examination of Sources of Variability Across the Consonant-Nucleus-Consonant Test in Cochlear Implant Listeners

Julie Arenberg Bierer1, Eugene Spindler2, Steven M. Bierer1, and Richard Wright3

Abstract
The 10 consonant-nucleus-consonant (CNC) word lists are considered the gold standard in the testing of cochlear implant (CI) users. However, variance in scores across lists could degrade the sensitivity and reliability of them to identify deficits in speech perception. This study examined the relationship between variability in performance among lists and the lexical characteristics of the words. Data are from 28 adult CI users. Each subject was tested on all 10 CNC word lists. Data were analyzed in terms of lexical characteristics, lexical frequency, neighborhood density, bi-, and tri-phonemic probabilities. To determine whether individual performance variability across lists can be reduced, the standard set of 10 phonetically balanced 50-word lists was redistributed into a new set of lists using two sampling strategies: (a) balancing with respect to word lexical frequency or (b) selecting words with equal probability. The mean performance on the CNC lists varied from 53.1% to 62.4% correct. The average difference between the highest and lowest scores within individuals across the lists was 20.9% (from 12% to 28%). Lexical frequency and bi-phonemic probabilities were correlated with word recognition performance. The range of scores was not significantly reduced for all individuals when responses were simulated with 1,000 sets of redistributed lists, using both types of sampling methods. These results indicate that resampling of words does not affect the test–retest reliability and diagnostic value of the CNC word test.

Keywords
cochlear implants, speech perception, lexical frequency, monosyllabic words

Date received: 12 August 2015; revised: 17 March 2016; accepted: 2 April 2016

Introduction
The Consonant-Nucleus-Consonant (CNC) word lists, included in the Minimum Speech Test Battery for Adult Cochlear Implant Users (Luxford, 2001), are considered the gold standard in the testing and management of cochlear implant (CI) users. But despite the extensive use of CNC words, list-to-list variance in performance presents an issue in clinical and research applications because it diminishes the sensitivity of the test to assess speech recognition. This study examines the contributions of lexical characteristics of the CNC words to performance variability on the CNC word lists and evaluates the efficacy of lexically rebalanced word lists to reduce this variability.

The original CNC word lists (Lehiste & Peterson, 1959) were a set of 10 lists of 50 words each that were phonetically balanced (with roughly equal phonemic probabilities) and controlled for text-based lexical frequency across lists. They were an expansion and improvement on previously existing lists (Egan, 1948, 1957) and developed explicitly for use in hearing tests. Lehiste and Peterson (1959) collected 1,263 CNC words from Egan’s lists and The Teacher’s Word Book of 30,000 Words (Thorndike & Lorge, 1944) text corpus. They sought to improve phonemic balance by analyzing the frequency with which phonemes occurred and excluding all words occurring less than once per million. They subsequently published a revision in which lexical frequency was made to be more uniform across the lists by limiting
the occurrence of rare and common words (Peterson & Lehiste, 1962). These revised word lists are widely used clinically today. But while the lexical characteristics of the lists were reasonably balanced at the time they were compiled, there are two basic limitations to their application today. The first is that the lists were based on written text rather than on spoken language. The second is that contemporary spoken-frequency patterns have changed since the lists were developed.

Since the creation of the CNC word lists in 1962, extensive spoken language corpora, such as the Corpus of Contemporary American English (COCA; corpus.byu.edu/coca), and detailed lexical corpora, such as the Irvine Phonotactic Online Dictionary (IPhOD; Vaden, Halpin, & Hickok, 2009), have become available. The existence of these corpora makes possible a more real-world analysis of lexical and spoken lexical characteristics. Relating these characteristics to CI-user performance on the CNC lists may help reveal the causes of list-to-list variance in performance and inform the development of new lists with more uniform performance characteristics.

Despite the importance of the CNC test material clinically, there have been only two peer-reviewed studies giving a detailed analysis of word composition. In 1970, Elkins evaluated the lexical and phonetic balance of the word lists in normal hearing listeners and confirmed that there was no significant difference when compared with the reference of a spoken corpus of transcribed telephone conversations from the 1930s. A more recent study by Skinner et al. (2006) showed that performance of CI listeners was quite variable from list to list, suggesting a limited sensitivity of the instrument in assessing speech perception.

As part of the Skinner study, performance of 22 CI listeners on a new set of monosyllabic word lists was compared with performance on the standard CNC set. They analyzed performance by list, word, consonant type, and consonant position in word. The results demonstrated that performance was quite variable across the CNC lists, ranging from 51.3% to 62.9% word correct with an overall mean of 56.7%. The newly developed word lists showed similar variability across lists, but performance was significantly poorer compared with the existing CNC lists.

While Skinner et al. (2006) investigated phonemic and word structure factors relating to cross-list variability, lexical frequency and other lexical characteristics have not yet been examined. These factors are known to contribute to intelligibility (e.g., Luce et al., 1990; Luce & Pisoni, 1998), so assessing their impact on CNC scores can potentially aid in the design of more reliable word lists. One solution proposed by Skinner et al. was to make list pairings to create equivalent mean performance across pairs. One obvious disadvantage of this approach is that it effectively halves the number of available lists while doubling the time to acquire test results. A second potential problem comes from the lack of subsequent published validation of the particular list pairings. An alternative pair of solutions are explored in this study, in which the existing pool of CNC words are redistributed into new lists based on (a) spoken-lexical frequency, resulting in a new set of lists with more equivalent lexical characteristics and (b) words are resampled randomly into sets of lists without regard for lexical frequency, thereby decreasing the likelihood of lexical bias in the lists. While these approaches sacrifice some of the phonemic balance across lists, they have the advantage of keeping 10 lists while diminishing variance in performance due to lexical characteristics.

Two additional studies have brought into question the value of phonemic balancing to the sensitivity or test–retest reliability of the CNC lists. Martin, Champlin, and Perez (2000) randomly generated lists from monosyllabic words drawn from a dictionary and compared performance of hearing-impaired listeners on their lists to the Northwestern University-6 (NU-6). They found no consistent benefit of phonemic balancing across lists. In a comparison of multiple word lists including the Central Institute for the Deaf (CID) W-22 and NU-6 speech material, Wilson, McArdele, and Roberts (2008) made similar conclusions. Moreover, strict adherence to principles of phonemic balancing can limit the number of word lists that can be generated from a selection of words.

Word lexical-frequency and phonological neighborhood density (phonological similarity of words) are two of the most studied lexical characteristics in the context of word recognition (e.g., Cluff & Luce, 1990). Both factors have been shown to affect the speed and accuracy of word recognition under a variety of conditions and with a variety of populations. For example, high-frequency words and low-density words are identified more accurately than their low-frequency and high-density counterparts for adults (e.g., Cluff & Luce, 1990), and children (Charles-Luce & Luce, 1990; Krull, Choi, Kirk, Prusick, & French, 2010; Paatsch, Blamey, Sarant, Martin, & Bow, 2004) with normal hearing in English-speaking populations. Similar effects have been reported for hearing-impaired adults (Kirk, Pisoni, & Miyamoto, 1997; Sommers, 1996) and children (Eisenberg, Martinex, Holowecky, & Pogorelsky, 2002; Kirk, Hay-McCutcheon, Sehgal, & Miyamoto, 2000; Kirk, Pisoni, & Osberger, 1995). The probability of phonemes co-occurring within words (biphonemic and triphonemic probabilities for CNC words) is related to neighborhood density and is therefore also examined in this study.

The current study examines the contributions of lexical properties to performance and explores two methods
to reduce variance across lists using the original CNC words: (a) lists are created using lexically balanced sampling and (b) lists are created using random sampling without replacement. The strongest lexical contributor to word recognition is lexical frequency. Neither of the sampling methods significantly reduced list-to-list variability compared with the original CNC lists.

**Methods**

**Subjects**

CNC performance data were analyzed from 28 adult CI subjects, including 6 tested at the University of Washington (UW; see Table 1) and 22 tested for a previous study at Washington University (Skinner et al., 2006). The UW subjects were added to increase the number of participants and to improve statistical power. All listeners became deaf after acquiring spoken language, except for one subject from the previous data set (W22). The listeners ranged in age from 22 to 79 years with a mean age of 56 years in the Skinner study, and from 50 to 80 with a mean age of 63 in the current set of subjects. Duration of severe-to-profound hearing loss prior to implantation was 0.66 to 43 years with a mean of 6 years in the Skinner study, and 1 to 50 with a mean of 24 years for the additional subjects. All of the subjects tested at the University of Washington wore the Advanced Bionics, HiRes90k device. The Human Subjects Review Boards at Washington University in St. Louis and the University of Washington in Seattle approved all procedures, and all subjects provided written informed consent.

**Speech Material and Equipment**

The 10 original CNC word lists were played through an external A/D device (SIIG USB SoundWave 7.1), amplified by a Crown Amplifier (D75) and presented at 60 dB SPL in the sound field inside a double-walled sound attenuating booth. The sound files, identical to those recommended for the MSTB for Adult Cochlear Implant Users (Luxford, 2001), were presented from a desktop PC using custom software (ListPlayer2 version 2.2.10.39, Advanced Bionics). The stimuli were calibrated to a 1 kHz tone with a sound level meter (Bruel and Kjaer, Hand-held Analyzer Type 2250 and ZC 0032 microphone) and presented through a loudspeaker (Bose 161) placed at ear level height, at 0 degrees azimuth and 1 meter from the subjects’ heads.

**Procedures**

Each of the 10 CNC word lists was presented once to each subject in random order. During word recognition testing, subjects used their everyday speech processor program (channel map and sensitivity and volume control settings). Subject responses were transcribed by the experimenter and scored for phonemic and whole word accuracy, and then reviewed for consistency. Transcribed responses collected by Skinner et al. (2006) were reviewed for consistency across scorers to insure that the rules for scoring provided in the Minimum Speech Test Battery were followed, as in correcting for plurals of singular tokens and insertions. The scoring also took into account the low back merger (e.g., “cot” vs. “caught”) in the scorers’ regional dialect (Western).

**Lexical Analyses**

Lexical frequency values were estimated from the *spoken corpus* subset of the COCA (Davies, 2008) and transformed using a \( \log_2 \) (binary logarithm) to approximate normal distributions for statistical analysis. The COCA is a 450 million word balanced corpus, and the 95 million word spoken subset is the largest known corpus of spoken American English. Values are in the form of corrected occurrence per million words in the corpus.

Lexical neighborhood density values were estimated from the IPhOD collection of English words (Vaden et al., 2009) using the estimation method employed in Vitevitch and Luce (1999). The values reported reflect the number of lexical neighbors of each word. Alternate methods for computing the likelihood of the phonemes in a particular token were explored with the bi-phonemic and tri-phonemic transitional probabilities, also estimated using IPhOD.

**Model-Simulated Lists**

To quantify the variability related to lexical frequency, the 500 CNC words were redistributed into new 50-word lists using the following procedure, carried out by a custom program in MATLAB (Mathworks). First, the words were sorted by lexical frequency then placed into
50 bins of 10 words, organized from least to most frequent in lexical frequency. Next, each word in a bin was randomly assigned to 1 of the 10 lists, with equal probability, using the MATLAB “randperm” function. This was repeated for all 50 bins, assuring that all lists had a comparable representation of lexical frequency. Finally, listener performance was recalculated for each new list based on the hit-or-miss identification of the individual words comprising that list. This simulation was repeated 1,000 times, creating 1,000 unique sets of 10 lists from the same 500 CNC words. Additionally, the randomized list reassignment was conducted without first ordering the words based on lexical frequency. This control simulation was also repeated 1,000 times.

**Results**

Speech perception scores on the original CNC lists, averaged across listeners, ranged from 53.1% (List 8) to 62.4% (List 9) with a mean score and SD of 57.7 ± 3.0%. Figure 1 shows the percent words correct as a function of list number, sorted from lowest to highest average score. The data in Figure 1(a) are from Skinner et al. (2006), and the data in Figure 1(b) include the additional six UW subjects tested in the present study. Note that the list order changed slightly such that the adjacent lists 3 and 10, and 2 and 4 are reversed. A mixed-model ANOVA shows a significant effect of list (Greenhouse-Geisser, $F(6.25, 162.37) = 2.73, p = 0.014$).

Although the range of scores across the 10 lists when averaged across listeners is only approximately nine percentage points, the range within individual subjects varies from 12% to 28%, with an average of 20.85%. This performance variability is evident in Figure 2(a), which plots the list scores of every subject in order from lowest to highest average score. Interestingly, both broad and narrow ranges are evident at both ends of the graph. Subject 5, for instance, exhibits a much wider dispersion of list scores compared to subject 18, even though both listeners had comparably low mean performance. The overall variability in ranges across subjects is more clearly seen in Figure 2(b), in which mean score is subtracted for each subject and then plotted according to the average range of within-subject scores. As expected from a visual inspection of Figure 2(a) and (b), the average subject performance was not correlated with variability in performance (Pearson correlation; $r = 0.02, p = .92$).

Figure 3 examines the relationship of two lexical factors, lexical frequency and phonological neighborhood density, with recognition of individual CNC words, averaged across the 28 subjects. The distribution of COCA values for lexical frequency is skewed by outliers, with a rate of occurrence that varies across lists; therefore, the frequencies were transformed on a log2 $n$ (binary logarithm) scale to yield an approximately normal distribution. A comparison of these values with CNC word
recognition, shown as a scatter plot in Figure 3(a), indicates that the more frequently a word occurs in spoken American English, the more likely a listener was to accurately identify the word. Lexical neighborhood density, however, did not show a relationship with word accuracy, as observed in Figure 3(b). A multiple linear regression analysis of word accuracy with lexical frequency, neighborhood density, bi-, and tri-phonemic probabilities was performed. The analysis revealed a significant correlation between word recognition and the four factors, with a $R^2$ of 0.037, $F(4) = 4.738$, $p = 0.001$, with lexical frequency as the strongest predictor, $t(4) = 3.279$, $p = 0.001$. The next strongest predictor was bi-phonemic probability, $t(4) = 2.114$, $p = 0.035$. There was no statistically reliable correlation in the complete model for neighborhood density, $t(4) = 1.174$, $p = 0.241$ or tri-phonemic probability, $t(4) = 0.703$, $p = 0.482$.

An examination of the relationship between lexical frequency and performance by individual lists is shown in Figure 4. Note that only three lists (2, 6, and 9) show significantly higher recognition scores for words that occur more frequently in contemporary spoken American English. Therefore, these three lists may have a disproportionate influence on the overall effect indicated by the multiple linear regression analysis.

If variability in lexical frequency contributed to the wide ranges in list scores observed for most listeners (Figure 2), then balancing the lists should reduce those ranges. This hypothesis was tested by generating new CNC word lists with a more equitable distribution of lexical frequencies than the original, as described in the Methods section. Because testing every subject on the balanced lists would be impractical, performance was simulated by re-tallying scores obtained with the original lists according to the occurrence of each word in the new lists. The simulations were performed 1,000 times in order to evaluate how different word arrangements can impact the range of list scores. (Note that, because word-by-word performance is unchanged, mean performance...
for each subject is a constant and identical to that of the original lists.)

Figure 5 plots the performance ranges of the original lists (red) and the distribution of those ranges for the resorted simulated lists based on lexical frequency (blue) and random selection (green). For 15 of the 28 subjects, the score range is larger for the original lists than the median of the simulated distributions; an additional 8 subjects showed no change, and only 5 subjects exhibited an increase in mean range. Despite the reduction in performance range for the majority of subjects the range of scores with the original lists falls within 75% confidence limits of the distributions of simulated data suggesting the original lists were not significantly worse than the new lists. Additionally, a paired \( t \)-test of ranges obtained with the median balanced lists with those obtained from the original lists did not show a significant change at the 5% confidence level, \( t(27) = -1.973, p = 0.059 \). Similarly, ranges obtained with the control simulation were not significantly different from those of the original set, \( t(27) = -1.875, p = 0.072 \), or from the balanced simulation, \( t(27) = -0.441, p = 0.663 \). Nevertheless, it is worth noting the relationship between mean range on the original lists and the change in range after reordering: subjects with the highest original ranges (i.e., those to the right of Subject 9 in Figure 5) tended to show a slight benefit from word redistribution, whether the new lists were lexically balanced or not. This interaction between range reduction and range average was a likely factor in the marginal statistical outcomes described earlier.

**Discussion**

Performance variability on the CNC word lists poses difficulties for monitoring patient progress as well as for reliability of the lists as a sensitive research tool. In an effort to understand factors related to the performance variability, this study investigated lexical characteristics of the lists in relation to listener performance. Two lexical factors, lexical frequency and bi-phonemic probability, were shown to correlate with performance on the CNC word lists. Tri-phonemic probability failed to reach significance making the bi-phonemic effect less clearly interpretable. Therefore, lexical frequency was chosen as the main lexical effect to concentrate on because it is both a stronger and more widely used lexical factor.

The list-pairing approach proposed by Skinner et al. (2006) halves the number of lists available for testing and experimentation, thereby exacerbating the well-known problem of reusing static lists across testing sessions. Two alternative solutions were explored in this study using different sampling strategies of the CNC
words: (a) based on lexical frequency and (b) based on random sampling with equal probability. Results indicate that a slight reduction in performance variability among lists can be achieved using either lexical or random sampling. If the results of the simulations extend to clinical settings, then the variability reduction could benefit measurements of individual-listener performance.

While this study’s method of generating lists discards phonemic balancing across lists, it has two distinct advantages: the first is that it may slightly reduce performance variance across lists; the second is that it can potentially make a much larger number of lists available for testing and research, thereby reducing the need to recycle static lists from test to test. Further experiments are needed to determine optimal word tests for clinical use and reducing list-to-list variability.

Acknowledgements
We would like to thank the cochlear implant listeners for their time and commitment. We would also like to thank Laura Holden, Jill Firszt, and the late Margaret Skinner for sharing their published CNC word data with us.

Declaration of Conflicting Interests
The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding
The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by NIH DC012142 (JAB).

References
Charles-Luce, J., & Luce, P. A. (1990). Similarity neighbourhoods of words in young children’s lexicons. *Journal of Child Language, 17*(1), 205–215.
Cluff, M. S., & Luce, P. A. (1990). Similarity neighborhoods of spoken two-syllable words: Retroactive effects on multiple activation. *Journal of Experimental Psychology: Human Perception and Performance, 16*(3), 551–563.
Davies, M. (2008). The corpus of contemporary American English: 450 million words, 1990–present. Retrieved from http://corpus.byu.edu/coca/.
Egan, J. (1948). Articulation testing methods. *Laryngoscope, 58*(9), 955–991.
Egan, J. (1957). Remarks on rare PB words. *Journal of the Acoustical Society of America, 29*, 751.
Eisenberg, L. S., Martinek, A. S., Holoweycky, S. R., & Pogorelsky, S. (2002). Recognition of lexically controlled words and sentences by children with normal hearing and children with cochlear implants. *Ear and Hearing, 23*(5), 450–462.
Elkins, E. (1970). Analyses of the phonetic composition and word familiarity attributes of CNC intelligibility word lists. *Journal of Speech and Hearing Disorders, 35*(2), 156–160.
Kirk, K. I., Hay-McCutcheon, M., Sehgal, S. T., & Miyamoto, R. T. (2000). Speech perception in children with cochlear implants: Effects of lexical difficulty talker variability and word length. *Annals of Otology, Rhinology, and Laryngology Suppl*, 185, 79–81.

Kirk, K. I., Pisoni, D. B., & Osberger, M. J. (1995). Lexical effects on spoken word recognition by pediatric cochlear implant users. *Ear and Hearing*, 16(5), 470–481.

Kirk, K. I., Pisoni, D. B., & Miyamoto, R. T. (1997). Effects of stimulus variability on speech perception in listeners with hearing impairment. *Journal of Speech, Language, and Hearing Research*, 40(6), 1395–1405.

Krull, V., Choi, S., Kirk, K. I., Prusick, L., & French, B. (2010). Lexical effects on spoken-word recognition in children with normal hearing. *Ear and Hearing*, 31(1), 102–114.

Lehiste, I., & Peterson, G. E. (1959). Linguistic considerations in the study of speech intelligibility. *Journal of the Acoustical Society of America*, 31(3), 280–286.

Luce, P. A., Pisoni, D. B., & Goldinger, S. D. (1990). Similarity neighborhoods of spoken words. In G. T. M. Altmann, (Ed.), *Cognitive models of speech processing: Psycholinguistic and computational perspectives*. Cambridge, MA: MIT.

Luce, P. A., & Pisoni, D. B. (1998). Recognizing spoken words: The neighborhood activation model. *Ear and Hearing*, 19(1), 1–36.

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

Martin, F., Champlin, C., & Perez, D. (2000). The question of phonetic balance in word recognition testing. *Journal of the American Academy of Audiology*, 11(9), 489–493.

Paatsch, L. E., Blamey, P. J., Sarant, J. Z., Martin, L. F., & Bow, C. P. (2004). Separating contributions of hearing, lexical knowledge and speech production to speech-perception scores in children with hearing impairments. *Journal of Speech, Language and Hearing Research*, 47(4), 738–750.

Peterson, G. E., & Lehiste, I. (1962). Revised CNC lists for auditory tests. *Journal of Speech and Hearing Disorders*, 27(1), 62–70.

Skinner, M. W., Holden, L. K., Fourakis, M. S., Hawks, J. W., Holden, T., Arcaroli, J., & Hyde, M. (2006). Evaluation of equivalency in two recordings of monosyllabic words. *Journal of the American Academy of Audiology*, 17, 350–366.

Sommers, M. S. (1996). The structural organization of the mental lexicon and its contribution to age-related declines in spoken-word recognition. *Psychology and Aging*, 11(2), 333–341.

Thorndike, E., & Lorge, I. (1944). *The teacher’s word book of 30,000 words*. New York, NY: Bureau of Publications Teachers College, Columbia University.

Vaden, K. I., Halpin, H. R., & Hickok, G. S. (2009). Irvine phonotactic online dictionary, Version 2.0. [Data file]. Retrieved from http://www.iphod.com/.

Vitevitch, M. S., & Luce, P. A. (1999). Probabilistic phonotactics and neighborhood activation in spoken word recognition. *Journal of Memory and Language*, 40, 374–408.

Wilson, R., McArdle, R., & Roberts, H. (2008). A comparison of recognition performances in speech-spectrum noise by listeners with normal hearing on PB-50, CID W-22, NU-6, W-1 spondaic words, and monosyllabic digits spoken by the same speaker. *Journal of the American Academy of Audiology*, 19, 496–506.