# Some Considerations in Evaluating Spoken Word Recognition by Normal-Hearing, Noise-Masked Normal-Hearing, and Cochlear Implant Listeners. I: The Effects of Response Format

**Mitchell S. Sommers**, **Karen Iler Kirk**, and **David B. Pisoni**
Department of Psychology (M.S.S.), Washington University, St. Louis, Missouri 63130;
Department of Otolaryngology-Head and Neck Surgery (K.I.K.), Indiana University School of
Medicine, Indianapolis, Indiana 46202; and Speech Research Laboratory (D.B.P.), Department of
Psychology, Indiana University, Bloomington, Indiana 47405

## Abstract

**Objective—**The purpose of the present studies was to assess the validity of using closed-set
response formats to measure two cognitive processes essential for recognizing spoken words—
perceptual normalization (the ability to accommodate acoustic-phonetic variability) and lexical
discrimination (the ability to isolate words in the mental lexicon). In addition, the experiments
were designed to examine the effects of response format on evaluation of these two abilities in
normal-hearing (NH), noise-masked normal-hearing (NMNH), and cochlear implant (CI) subject
populations.

**Design—**The speech recognition performance of NH, NMNH, and CI listeners was measured
using both open- and closed-set response formats under a number of experimental conditions. To
assess talker normalization abilities, identification scores for words produced by a single talker
were compared with recognition performance for items produced by multiple talkers. To examine
lexical discrimination, performance for words that are phonetically similar to many other words
(hard words) was compared with scores for items with few phonetically similar competitors (easy
words).

**Results—**Open-set word identification for all subjects was significantly poorer when stimuli
were produced in lists with multiple talkers compared with conditions in which all of the words
were spoken by a single talker. Open-set word recognition also was better for lexically easy
compared with lexically hard words. Closed-set tests, in contrast, failed to reveal the effects of
either talker variability or lexical difficulty even when the response alternatives provided were
systematically selected to maximize confusability with target items.

**Conclusions—**These findings suggest that, although closed-set tests may provide important
information for clinical assessment of speech perception, they may not adequately evaluate a
number of cognitive processes that are necessary for recognizing spoken words. The parallel
results obtained across all subject groups indicate that NH, NMNH, and CI listeners engage
similar perceptual operations to identify spoken words. Implications of these findings for the
design of new test batteries that can provide comprehensive evaluations of the individual
capacities needed for processing spoken language are discussed.

**REFERENCE NOTES**

The ability to recognize spoken words is critically dependent on the integration of a number of sensory, perceptual, and cognitive capacities (Klatt, 1989; Pisoni, 1985). Identifying and discriminating phonetic features, segmenting the speech waveform, compensating for talker differences, and accessing words from the mental lexicon are just some of the component operations necessary for transforming acoustic speech sounds into meaningful linguistic perceptions. Although a number of instruments currently are available for both clinical and experimental assessment of speech perception abilities (Dorman, 1993; Dowell, Brown, & Mecklenberg, 1990; Geers & Brenner, 1994; Rosen et al., 1985), relatively little systematic empirical research has been directed at establishing the specific perceptual and cognitive capacities that each of these tests measure. Consequently, little is known about either the utility or limitations of different assessment procedures for evaluating the individual abilities necessary to recognize spoken words.

For example, one of the most common methods of measuring speech discrimination is the closed-set test format. Closed-set tests such as the Modified Rhyme Test (MRT; House, Williams, Hecker, & Kryter, 1965) and portions of the Minimum Auditory Capabilities test (Owens, Kessler, & Schubert, 1981) require listeners to select one of several provided response alternatives that best matches a presented stimulus. This response format has been extremely useful for assessing certain perceptual capacities necessary to understand spoken language, especially in clinical populations with severely impaired auditory functions. Closed-set tests have been used to provide information about the discrimination of phonetic features, prosodic characteristics, and timing aspects of speech signals (Blamey, Dowell, Brown, Clark, & Seligman, 1987; Geers & Brenner, 1994; Tyler, Lowder, Otto, Preece, Gantz, & Mc-Cabe, 1984). In addition, closed-set response formats have been shown to be sensitive to word frequency (Elliott, Clifton, & Servi, 1983) and to differences in linguistic background (Garstecki & Wilkin, 1976).

One potential disadvantage of closed-set measures, however, is that they may not adequately simulate the same cognitive demands that individuals confront in natural listening environments. Closed-set speech perception tests typically contain highly articulated tokens of words produced by a single talker and provide listeners with a restricted set of response alternatives. In contrast, real-world listening environments often contain poorly articulated stimuli spoken by multiple talkers with only minimal restrictions on potential response candidates. These differences between natural listening conditions and closed-set tests may limit the ability of closed-set formats to predict speech perception performance in everyday conversational situations. The purpose of the present studies, therefore, was to establish the validity of using closed-set tests to assess two cognitive abilities necessary for recognizing spoken words—normalizing for talker differences and isolating words in the mental lexicon. In addition, the experiments were designed to determine whether evaluation of these two capacities with closed-set formats differed in listeners with varying degrees of simulated or actual sensory impairment.

To achieve these goals, normal-hearing (NH), noise-masked normal-hearing (NMNH), and co-chlear implant (CI) listeners were tested using both open- and closed-set formats on their ability to accommodate changes in talker characteristics and to isolate words in long-term lexical memory. These two capacities were selected as the focus of our investigations because they have been shown to be critical components of the early stages of spoken language processing (Luce, 1986; Luce, Pisoni, & Goldinger, 1990; Mullennix, Pisoni, & Martin, 1989; Sommers, Nygaard, & Pisoni, 1994). Results indicating that open- and closed-set tests are equally effective in their ability to evaluate these two processes would provide empirical support for additional perceptual capacities that can be assessed with closed-set formats. In contrast, differential results using the two response formats would suggest that

closed-set measures of spoken word recognition may be limited in their ability to assess one or more perceptual capacities used in speech perception.

## THE IMPORTANCE OF TALKER NORMALIZATION AND LEXICAL DISCRIMINATION FOR SPOKEN WORD RECOGNITION

The ability to adapt rapidly to changes in talker characteristics is an essential aspect of processing spoken language because differences in the size and shape of vocal tracts result in a many-to-one mapping between acoustic speech signals and phonetic perceptions. Thus, for example, the same word produced by a man, a woman, and a child will have dramatically different acoustic properties due to differences in the physical characteristics of the talkers' vocal tracts (Peterson & Barney, 1952). Normal-hearing listeners, however, generally have little difficulty recognizing these distinct speech signals as phonetically equivalent (i.e., as instances of the same word). This ability to maintain perceptual constancy in the face of extensive acoustic-phonetic variability traditionally has been attributed to a stage of processing, referred to as perceptual normalization, during which listeners derive standardized phonetic representations that can then be matched to canonical forms stored in long-term memory (Johnson, 1990; Joos, 1948; Nearey, 1989; Pisoni, 1993).

The importance of talker normalization for spoken word recognition has now been well established (Mullennix & Pisoni, 1990; Mullennix et al., 1989; Sommers et al., 1994). For example, several investigators (Mullennix et al., 1989; Sommers et al., 1994) have examined the effects of requiring listeners to compensate or normalize for talker differences by comparing speech recognition performance for word lists produced by single and multiple talkers. The general finding from these studies is that multiple-talker contexts produce a 10 to 15% reduction in identification scores relative to the identical words spoken by a single talker. One hypothesis that has been proposed to account for this finding is that the greater demand for talker normalization in the multiple-talker contexts diverts limited processing resources from perceptual operations used in phonetic identification (Martin, Mullennix, Pisoni, & Summers, 1989; Mullennix et al., 1989; Sommers et al., 1994). That is, the mixed-talker condition requires more processing to maintain perceptual constancy, and, consequently, listeners have fewer cognitive resources available for identifying spoken words. If assessment instruments employing closed-set formats fail to engage all of the resources needed for speech perception under more natural (open-set) conditions, they may not be sensitive to the effects of talker variability or other factors that affect the acoustic-phonetic properties of speech signals.

Once a standardized phonetic representation has been obtained through the normalization process, it must be matched to idealized representations stored in the mental lexicon. However, matching every representation derived from incoming speech signals to one of the tens of thousands of representations stored in long-term memory would place considerable, and almost certainly excessive, demands on the speech perception system. Speech researchers have therefore proposed several mechanisms that function to restrict the number of items within the mental lexicon that are compared with incoming speech waveforms (Luce et al., 1990; Marslen-Wilson, 1987).

One such proposal, the Neighborhood Activation Model of spoken word recognition (NAM; Luce et al., 1990), assumes that words in the mental lexicon are organized into similarity neighborhoods. A similarity neighborhood, according to the model, consists of a target word and all other words that can be created from that item by adding, deleting, or substituting a single phoneme. Thus, the neighborhood for the word "CAT" would include the words (referred to as neighbors) "COT," "KIT," "CAB," and "SCAT" (and all other words differing from CAT by a single phoneme). Luce et al. (1990) suggested that speech signals

activate only items within a single similarity neighborhood and that word recognition occurs by selecting among this restricted set of activated neighbors.

One prediction from the model that has received considerable empirical support (Kirk, Pisoni, & Os-berger, 1995; Luce, et. al, 1990; Sommers, 1996) is that the difficulty of isolating a target word from its neighbors will be determined by both the number of words within the neighborhood (neighborhood density) and the average frequency of those neighbors (neighborhood frequency) as determined by word frequency norms (Kucera & Francis, 1967). Specifically, the NAM predicts that words from high-density, high-frequency neighborhoods should be identified less accurately than words from low-density, low-frequency neighborhoods. That is, words with many similar-sounding, high-frequency neighbors (lexically hard words) will be more difficult to isolate than those that have only a few, low-frequency competitors (lexically easy words). Consistent with this prediction, a number of investigations have reported that both the speed and accuracy of processing spoken words is reduced for lexically hard compared with lexically easy items (Cluff & Luce, 1990; Luce et al., 1990; Sommers, 1996).

Taken together, the results of previous studies suggest that both talker variability and lexical difficulty can significantly affect spoken word recognition performance. However, these results have been obtained exclusively using open-set formats. Given the extensive use of closed-set measures in clinical assessments, it is essential to establish whether closed-set tests are also sensitive to the effects of talker variability and lexical difficulty. Differential effects of stimulus variability and lexical difficulty as a function of test format would indicate that the two types of assessment procedures are not equally capable of measuring the component operations needed for spoken word recognition. Furthermore, comparing the effects of response format in both hearing-impaired and normal-hearing subject populations will provide an indication of whether these groups engage similar processing mechanisms in recognizing spoken words.

## EXPERIMENT 1A

The purpose of Experiment 1A was to examine the effects of talker variability and lexical difficulty on perceptual identification in NH, NMNH, and CI subjects using both open- and closed-set response formats. The rationale for this approach is that it provides a methodology for examining the independent effects of hearing loss and cochlear implants on the perceptual operations used to recognize spoken words in open- and closed-set tests. Differences between the NH and the NMNH groups as a function of test format would suggest that reduced absolute sensitivity alters the mechanisms that listeners engage to recognize spoken words in open- and closed-set measures. Similarly, qualitative performance differences between the NMNH listeners and the CI listeners would suggest that, independent of hearing loss, processing limitations imposed by cochlear implants change the perceptual operations used for speech perception.

### Method

**Subjects**—Four groups of subjects were tested in Experiment 1A. Group 1 consisted of eight adult CI patients who were seen at Indiana University Medical Center as part of their regularly scheduled postimplant appointments. Seven of the CI participants used the Nucleus 22-channel implant, and one used the Clarion implant. Etiologies for the profound deafness in the CI patients were meningitis (2), Meniere's disease (3), and unknown (3). Mean age at implantation was approximately 40 yr, and mean length of device use was 3.6 yr.

The remaining three groups of subjects all consisted of normal-hearing adult listeners tested under different signal-to-noise (S/N) ratios. Subjects in Groups 2 to 4 were recruited from the Washington University student population and surrounding community. All had pure tone air conduction thresholds of less than 20 dB HL for octave frequencies from 250 to 8000 Hz. Group 2 consisted of 11 listeners tested in quiet. Groups 3 and 4 were composed of 12 subjects each, and stimuli were presented at S/N ratios of +5 and −5 dB, respectively. Two different S/N ratios were used to examine how the extent of hearing loss affects the influence of response format on talker normalization and lexical access.

**Stimulus Materials—**Stimuli for the experiment were taken from a digital database containing 300 monosyllabic words from the MRT (House et al., 1965) recorded by 20 different talkers (10 male and 10 female). A total of 200 different words were selected from this database for use in Experiment 1A. Half of the items were produced by one of the 20 talkers and constituted the stimuli for the single-talker conditions. The remaining 100 items were used in the multiple-talker conditions and consisted of words produced by 10 different talkers (5 male and 5 female) with each talker contributing 10 items. Previous studies with stimuli from this database have reported that the intelligibility of the 20 talkers did not differ significantly (Martin et al., 1989; Mullennix et al., 1989).

In addition to dividing the 200 stimuli into single-and multiple-talker conditions, the words were further divided on the basis of lexical difficulty. Half of the words for the single- and multiple-talker conditions were lexically easy and half were lexically difficult, as defined by the NAM. The lexically easy words had an average neighborhood density of 11.3 (i.e., on average each word had approximately 11 neighbors) and an average neighborhood frequency of 43.4 (occurrences per million words). The corresponding values for the lexically hard words were a mean neighborhood density of 26 and a mean neighborhood frequency of 255.8. Thus, the easy words were selected from low-density, low-frequency neighborhoods, and the hard words were chosen from high-density, high-frequency neighborhoods. Each of the 10 talkers used in the multiple-talker conditions contributed five easy and five hard words.

**Procedure—**All subjects were tested using a repeated-measures design. Listeners received 100 items in both the open- and closed-set tests. Within each response format, two blocks of stimuli were presented. In one 50 item block, all of the words were produced by a single male talker. Half of these (25 items) were lexically easy and half were lexically difficult. In the other 50 item block for each format, the words were produced by all 10 talkers, and the voice presented on a given trial was selected randomly. Again, half of the mixed-talker words were lexically easy and half were lexically difficult. Order of presentation for the two blocks in each format was counterbalanced. Two versions of the test were constructed such that words presented in the open-set format on one form were presented in the closed-set format on the second form. Approximately half of the subjects in each group received each form of the test.

**Open-Set Tests—**For the CI listeners in Group 1, the digitized stimuli were converted to analog signals using a 12 bit D/A converter and a 10 kHz sampling rate. The signals were low-pass filtered at 4.5 kHz and recorded on audio tape. The stimuli were presented free field to CI patients sitting in a double-walled sound attenuating booth using a tape recorder (Nakamichi, CR-1A). Presentation level as measured at the output of the transducer was approximately 80 dB SPL. Subjects sat approximately 1.5 m from the transducer and wrote their responses on answer sheets provided by the experimenter. The interstimulus interval was 5 sec, which was generally sufficient for listeners to complete writing their responses. If the experimenter noticed a subject taking longer than normal to complete a response, the recorder was stopped until the listener completed the answer.

For the NH listeners (Groups 2 to 4), stimuli were converted to analog signals (12 bit D/A converter, 10 kHz sampling rate) and presented binaurally over matched and calibrated TDH-39 headphones. Levels for the speech signals were again approximately 80 dB SPL. The two simulated hearing loss groups were tested with speech signals presented in a background of white noise. The noise was generated using a noise generator (Grason Stadler, 90 IB) and was gated on and off coincident with presentation of the speech stimuli. The three groups of NH listeners all responded by typing their responses on a keyboard connected to a CRT terminal.

**Closed-Set Tests—**The procedures and equipment for testing listeners with closed-set response formats were identical to those used with the open-set tests except that all participants responded by circling one of six alternatives on a response sheet. The six response choices included the target item and five foils that differed from the target by a single phoneme. These were the same response alternatives used in the standard MRT.

## Results and Discussion

A mixed-design analysis of variance (ANOVA) was conducted on the identification scores for all four subject groups, with subject group and test version (version 1 or 2) as between-subjects factors and response format (open versus closed), lexical difficulty (easy versus hard), and stimulus variability (single versus multiple talker) as repeated measures. No significant main effects or interactions were observed for test version. Therefore, the remaining analyses were conducted with data from the two versions combined. Significant main effects were obtained for subject group ($F[3, 39] = 991.1$; $p < 0.001$) and response format ($F[1, 39] = 1358.5$; $p < 0.001$). As expected, performance was poorer for open- than for closed-set tests, and the differential hearing loss in the three hearing-impaired groups (actual loss in the case of the CI listeners and simulated losses in the case of the NH(+5) and NH(−5) listeners) produced systematic decrements in identification performance. Also consistent with previous findings (Luce et al., 1990; Mullennix et al., 1989; Sommers et al., 1994), identification scores were affected by both lexical difficulty ($F[1, 39] = 24.4$; $p < 0.001$) and stimulus variability ($F[1, 39] = 77.3$; $p < 0.001$); easy words were identified with greater accuracy than were hard words, and identification scores were higher for single-talker word lists than for multiple-talker word lists.

To determine whether open- and closed-set response formats were differentially sensitive to the effects of stimulus variability and lexical difficulty, several of the interactions obtained in the overall ANOVA were examined. First, a significant 2-way interaction between stimulus variability and response format was observed ($F[1, 39] = 30.2$; $p < 0.001$). Figure 1 displays identification scores for single and multiple talkers (collapsed across lexical difficulty) as a function of test format. Tukey HSD post hoc analyses indicated that all of the subject groups except the NH listeners tested in quiet[*] exhibited significantly poorer identification scores for multiple-talker lists in the open- but not in the closed-set response format (p < 0.05 for all open-set comparisons; $p > 0.2$ for all closed-set comparisons). It is important to note that this pattern of results was obtained despite significant differences in overall performance levels among the subject groups. Thus, although the NH(−5) and CI groups had significantly lower identification scores than did the NH(+5) listeners, they nevertheless exhibited effects of talker variability in open- but not in the closed-set tests. These results demonstrate that an important limitation on closed-set test formats is that they may not be sensitive to the effects of at least one source of acoustic—phonetic variability

---

[*]The absence of a significant difference in the open-set format for this group is most likely due to subjects approaching ceiling level performance.

present in many natural listening environments, namely the variability that results from changes in talker (vocal tract) characteristics.

A second important finding from this study was a reliable response format × lexical difficulty (F[1, 39] = 111.1, $p < 0.001$) interaction. Figure 2 displays identification scores for lexically easy and lexically hard words (collapsed across single- and multiple-talker conditions) as a function of test format. Tukey, HSD post hoc analyses indicated that, with the exception of the NH listeners tested in quiet (see Footnote), identification performance in open-set formats was significantly poorer for lexically hard compared with lexically easy words ($p < 0.05$ for all comparisons except the NH(Q) group). In contrast, examination of the data for the closed-set response formats revealed that none of the groups exhibited differences between easy and hard words when response alternatives were provided in the closed-set test (the effects of lexical difficulty for the NH(+5) group did approach significance ($p < 0.1$) in the closed-set format). Thus, the findings are similar to the results obtained with talker variability in that lexical difficulty influenced identification performance in open- but not in the closed-set tests.

One unanticipated finding revealed in the ANOVA was a significant subject group × format (F[3, 39] = 100.8; $p < 0.001$) interaction; the effects of changing from open- to closed-set formats differed significantly across the four subject groups. Figure 3 displays scores for all groups except the NH(Q) participants (performance for this group was already close to ceiling in the open-set test) as a function of test format. To further explore differences between open- and closed-set recognition performance, ratios of scores for the two test formats were computed for all listeners (except those in the NH(Q) group). Post hoc analyses computed on these ratios indicated that the CI patients exhibited the greatest benefit of changing from open- to closed-set measures ($p < 0.05$). One explanation for this finding is that the impoverished acoustic signal that CI patients receive from their devices may force them to rely on deriving broad phonetic categories (Ship-man & Zue, Reference Note 2) rather than obtaining detailed phonetic information in recognizing spoken words. Such a strategy will be most beneficial in closed-set tests where the response alternatives already are given. Koch, Carrell, Tremblay, and Kraus (Reference Note 1) recently have provided evidence to suggest that the ability to group speech sounds into broad phonetic classes is highly predictive of CI patients' ability to understand everyday speech. Thus, closed-set tests may be particularly beneficial to CI patients because the response alternatives provided allow them to map their broad phonetic classifications onto individual spoken words.

Examination of the remaining 2-way interactions revealed a significant subject group X variability (F[3, 39] = 3.8; $p < 0.05$) effect; the reduction in identification scores resulting from increased stimulus variability was significantly larger for the two simulated loss groups than for either the CI or NH(Q) listeners (as noted, the small effects for the NH(Q) group probably are due to listeners approaching ceiling level performance). In addition, a reliable 3-way subject group × format × variability interaction was obtained (F[3, 39] = 3.6; $p < 0.05$), indicating that the differential effects of stimulus variability across subject groups were limited to the open-set format. None of the remaining effects were statistically reliable.

Although the present study failed to find effects of lexical difficulty and talker variability in closed-set formats, this result may have been due, in part, to the specific response alternatives used in the MRT. If the five foils used with each item in the closed-set tests

---

[2]Shipman, D. W., & Zue, V. W. (1982). *Properties of large lexicons: Implications for advanced isolated word recognition systems.* Paper presented at the IEEE Conference on Acoustics of Speech and Signal Processing, Paris.
[1]Koch, D. B., Carrell, T. D., Tremblay, K, & Kraus, N. (1996). *Perception of synthetic syllables by cochlear-implant users: Relation to other measures of speech perception.* Poster presented at the mid-winter meeting of the Association for Research in Otolaryngology, St. Petersburg Beach, FL.

were not sufficiently confusable with the target word, the absence of significant lexical difficulty and stimulus variability effects may have been due to the relative ease of discriminating target words from response alternatives. Under conditions in which the target is easily distinguished from the foils, the effects of variables such as lexical difficulty and stimulus variability may be obscured. Therefore, Experiment IB was designed to examine the effects of lexical difficulty and talker variability in closed-set tests containing response alternatives that were systematically selected to be most confusable with the target items. Results similar to those of Experiment 1A would suggest that the failure to find effects of stimulus variability and lexical difficulty was not due to the nature of the response foils but to changes in task demands with closed-set test formats.

## EXPERIMENT IB

### Method

**Subjects—**Implant patients were not available for testing in Experiment IB. Therefore, only the three NH subject groups (NH(Q), NH(+5), and NH(−5)) were examined. Each group consisted of 15 listeners with pure tone air conduction thresholds less than 20 dB HL for octave frequencies from 250 to 8000 Hz. All participants were native speakers of English and reported no history of hearing loss or other auditory dysfunction at the time of testing.

**Selection of Response Alternatives—**As noted, the primary goal of Experiment IB was to determine whether the effects of lexical difficulty and stimulus variability could be obtained in a closed-set format when response alternatives were systematically selected to maximize confusability with the target word. To determine the five most confusable foils for each item of the MRT, phoneme confusion matrices derived by Xuce (1986) were examined. Luce (1986) measured consonant and vowel confusions for each position (initial consonant, medial vowel, final consonant) in consonant-vowel-consonant stimuli at a number of S/N ratios. To determine the relative confusability of a specific response alternatives, the probability of misidentifying each of the individual phonemes in the target with the corresponding phoneme in the alternative was determined. The product of these individual probabilities was then multiplied by a log transform of word frequency to obtain a measure of the overall confusability of the alternative.

For example, to calculate the probability of confusing the target word CAT with the response alternative KIT, the separate probabilities of /k/|/k/ (i.e., the probability of saying /k/ when /k/ was presented as the initial phoneme), /i/|/æ/ (the probability of saying /i/ when /æ/ was presented as the middle phoneme), and /t/|/t/ (saying /t/ given /t/ in the final phoneme position) were obtained from the confusion matrices. The product of these individual probabilities was then multiplied by a measure of word frequency to give a combined index of overall confusability with the target item.

In Experiment IB, an on-line lexical database was first used to identify all of the words that could be created from a given target item on the MRT by adding, deleting, or substituting a single phoneme. This provided the set of possible alternatives for that target stimulus. The phoneme confusion matrices were then used to determine which five of these alternatives were most confusable with the target, and these items were selected as the response foils for that item. Thus, each of the 100 MRT items tested in the closed-set format were presented with the target word and the five most confusable foils as response alternatives.

**Stimulus Materials and Procedure—**The stimuli and procedures were identical to those in Experiment 1A with the following exceptions. First, only three groups of listeners (NH(Q), NH(+5), and NH(−5)) were tested. Second, only the closed-set format was used. Third, the response foils provided with the MRT were replaced by the five most confusable

items for that target (as identified by the procedure for selecting response alternatives described above). As in the closed-set tests of Experiment 1A, two 50 item blocks (one single-talker, one multiple-talker) were presented. Within each block, half of the items were lexically easy and half were lexically hard.

### Results and Discussion

Figure 4 displays results for both talker variability (top panel) and lexical difficulty (bottom panel). A mixed-design ANOVA with lexical difficulty and talker variability as repeated-measures variables and subject group as a between-subjects factor revealed only a significant main effect of subject group ($F[2, 42] = 357.2$; $p < 0.001$). To examine the effects of using the most confusable response alternatives, the closed-set data from Experiments 1A and IB were combined and analyzed together. Experiment (1A or IB) was treated as a between-subjects factor, and lexical difficulty and talker variability served as repeated-measures variables. The only statistically reliable finding revealed in the analysis was a main effect of experiment ($F[1, 74] = 4.2$; $p < 0.05$); as expected, overall performance was poorer in Experiment IB than in Experiment 1A. Thus, although increasing the difficulty of the closed-set test, by systematically selecting the most confusable response alternatives, reduced identification performance, it did not increase sensitivity to the specific effects of talker variability or lexical difficulty. This finding suggests that even difficult closed-set tests may fail to engage perceptual operations that are used in recognizing spoken words under more natural (open-set) listening conditions. The implications of these findings are discussed in more detail below.

## GENERAL DISCUSSION

Taken together, the results of Experiments 1A and IB suggest that with open-set response formats, NH, NMNH, and CI listeners all exhibit reduced spoken word recognition as a function of increased stimulus variability and greater lexical difficulty. In contrast, these same variables produced no effects on recognition performance when a closed-set response format was used. One implication of the parallel findings across the different subject groups is that, despite considerable differences in absolute sensitivity, these listeners engaged qualitatively similar mechanisms in recognizing spoken words. For example, the significant reduction in open-set identification performance for lexically hard words suggests a similar structural organization of the mental lexicon and comparable use of this organization across subject groups. NH, CI, and NMNH listeners seem to organize words into lexical neighborhoods based on acoustic-phonetic similarity, and their word recognition performance is affected by differences in the number (neighborhood density) and frequency (neighborhood frequency) of the phonetically similar items in those neighborhoods.

The findings from this study are potentially quite important for clinical evaluation of speech perception abilities because they suggest that assessment instruments focusing exclusively on listeners' ability to extract phoneme information are necessary but not sufficient for predicting spoken word recognition performance. If speech perception required only the sequential identification of individual phonemes, word identification and phoneme recognition scores in CI patients should be highly correlated. However, the results of several recent studies are mixed with respect to the association between phoneme and word identification in listeners with cochlear implants (Kirk et al., 1995; Koch et al., Reference Note 1, Miyamoto et al., 1994; Rabinowitz, Eddington, Delhorne, & Cuneo, 1992). Kirk et al. (1995), for example, reported that phoneme recognition scores for pediatric CI patients using the Nucleus-22 channel device were not highly predictive of word identification performance on lexically easy and hard stimuli presented in open-set formats. Similarly, Miyamoto et al. (1994) found that, although phoneme recognition increased systematically over a 5 yr period in children using the Nucleus-22 device, only small performance gains in

word identification were observed during the same period. In contrast, Rabinowitz et al. (1992) reported relatively strong correlations between phoneme and word recognition scores in adult CI patients using the Ineraid cochlear implant (Richards Medical Co.).

Although factors such as differences in subject populations (e.g., adult versus children) and device type probably are at least partially responsible for the differential findings of studies examining the relationship between phoneme and word recognition, the low correlations between the two tasks that have been reported in some studies (Kirk et al., 1995; Miyamoto et al., 1994) also may reflect fundamental differences in underlying perceptual processes. That is, in addition to extracting phonetic information from the speech signal, spoken word recognition requires listeners to isolate (i.e., discriminate) individual words from phonetically similar sound patterns stored in long-term lexical memory. Thus, any comprehensive evaluation of spoken word recognition abilities in normal-hearing or clinical populations could benefit from incorporating tests designed to measure both phonetic and lexical discrimination as a means of assessing identification performance at several levels of lexical difficulty.

A second implication of the present findings is that measures of speech perception obtained with tests that minimize stimulus variability, by using highly articulated stimuli produced by a single talker, may fail to generalize to more natural listening situations where many different factors combine to produce extensive acoustic-phonetic variability. The significant reduction in identification scores that was observed after a change from single- to multiple-talker word lists indicates that the ability to adjust or normalize for acoustic-phonetic variations due to changes in talker characteristics is an integral aspect of the speech perception system, even for CI patients who receive highly impoverished acoustic information. The ability to predict "real-world" speech perception abilities will, therefore, require the development of new assessment procedures that can evaluate listeners' ability to rapidly accommodate changes in vocal tract characteristics. The current findings indicate that closed-set formats are likely to be inadequate for this purpose because they fundamentally change the task demands imposed on listeners during spoken word recognition.

One reason that closed-set formats may not be sensitive to the effects of either stimulus variability or lexical difficulty is that providing listeners with a set of response alternatives alters the perceptual strategies used to recognize spoken words. For example, in open-set speech discrimination tests, listeners must derive a best match between a phonetic representation obtained from the incoming speech signal and patterns stored in long-term lexical memory. Therefore, the organization of words within the mental lexicon can have a significant influence on the dynamics of the recognition process. In closed-set formats, however, listeners can effectively eliminate the need to access or consult long-term lexical memory by limiting their search to the response alternatives provided on a particular trial. That is, in closed-set formats, the set of potential response candidates is no longer determined by the structure of the mental lexicon. Instead, listeners can restrict their lexical search to the response alternatives that accompany each stimulus. Under such conditions, words are no longer recognized in the context of other phonetically similar words in memory, and the processes of word recognition and lexical discrimination are changed substantially. The differential nature of word recognition under the two response formats may, in part, explain the absence of lexical difficulty and stimulus variability effects in closed-set tests.

Although the preceding account of the influence of response format on spoken word recognition must be considered preliminary until additional evidence is obtained, it nevertheless raises the important theoretical issue that demand characteristics of an

assessment instrument may alter the processes used in spoken word recognition. One difference between open- and closed-set response formats that may partially account for the differential results obtained with talker variability and lexical difficulty is that closed-set tests may fail to mimic the cognitive demands associated with rapid, on-line recognition of spoken words. Consistent with this explanation, Mullennix et al. (1989) reported that reducing cognitive demands, by increasing S/N ratios, reduced the effects of talker variability. In fact, at artificially high S/N ratios, Mullennix et al. failed to observe any effects of talker variability. Considered with the earlier suggestion that providing response alternatives fundamentally alters the normal processes of lexical search and access, the present findings with closed-set formats may be attributable to a combination of reduced task demands and altered perceptual strategies that result from providing listeners with a set of response alternatives.

It could be argued that similar reductions in cognitive demands are achieved under natural listening conditions by using semantic context. That is, listeners may be able to limit the number of possible lexical alternatives that they consider by using semantic information. Although increasing semantic predictability has been shown to raise overall identification scores (Nittrouer & Boothroyd, 1990), recent evidence (Karl & Pisoni, 1994) indicates that the effects of talker variability are observed even with highly constrained semantic contexts. Karl and Pisoni (1994) had listeners transcribe Harvard sentences (Egan, 1948) that were produced in either single- or multiple-talker contexts. Their results indicated that multiple-talker transcription performance was significantly poorer than the single-talker condition. Thus, although semantic context can reduce overall task difficulty, the change in cognitive demands resulting from the addition of semantic information is qualitatively different from that produced by providing listeners with response alternatives. This proposal indicates that not all sources of context are equivalent, and we believe it is important to distinguish how different kinds of contextual information affect listeners' performance in a variety of speech perception tasks.

In summary, traditional instruments for assessing speech perception that rely on closed-set formats and single-talker productions can provide invaluable information about the perceptual capacities of clinical populations. These protocols are important for evaluating a number of individual abilities, such as identification and discrimination of phonetic features, that may be necessary for accurate speech perception. In addition, closed-set tests have been useful for examining speech processing strategies in listeners with sensory aids. The present findings suggest, however, that closed-set formats may not be effective for assessing other operations, such as perceptual normalization of talker differences and isolating words in long-term memory, that are also critical for spoken word recognition in natural listening environments. The results of the present study, therefore, represent an initial step in systematically evaluating and understanding the limitations of individual assessment instruments used to measure speech intelligibility in different populations. The goal of this research should be to develop comprehensive test batteries that include a variety of instruments designed to evaluate the broad spectrum of abilities necessary for understanding spoken language under a variety of listening conditions.

Speech perception is an extremely robust process that can quickly adapt to changing listening conditions. To understand the perceptual and neural mechanisms that are responsible for these abilities, we will need to develop a new generation of theoretically motivated tests that assess spoken word recognition across a range of task requirements and listening populations. The results of the present study demonstrate the potential value of this research strategy for gaining a more detailed understanding of speech perception and spoken-language processing.

## Acknowledgments

## REFERENCES

Blamey PJ, Dowell RC, Brown AM, Clark GM, Seligman PM. Vowel and consonant recognition of cochlear implant patients using formant-estimating speech processors. Journal of the Acoustical Society of America. 1987; 82:48–57. [PubMed: 3624640]

Cluff MS, Luce PA. Similarity neighborhoods of spoken two-syllable words: Retroactive effects on multiple activation. Journal of Experimental Psychology: Human Perception & Performance. 1990; 16:551–563. [PubMed: 2144570]

Dorman, MF. Speech perception by adults. In: Tyler, RS., editor. Cochlear implants: Audiological foundations. San Diego: Singular Publishing; 1993. p. 145-190.

Dowell, RC.; Brown, AM.; Mecklenberg, D. Clinical assessment of implanted deaf adults. In: Clark, G.; Tong, Y.; Patrick, J., editors. Cochlear prostheses. Edinburgh, UK: Churchill Livingstone; 1990. p. 193-206.

Egan JP. Articulation testing methods. Laryngoscope. 1948; 58:955–991. [PubMed: 18887435]

Elliott LL, Clifton LA, Servi DG. Word frequency effects for a closed-set word identification task. Audiology. 1983; 22:229–240. [PubMed: 6870665]

Garstecki DC, Wilkin MK. Linguistic background and test material considerations in assessing sentence identification ability in English- and Spanish-English-speaking adolescents. Journal of the American Audiological Society. 1976; 1:263–268.

Geers, AE.; Brenner, C. Speech perception results: Audition and lipreading enhancement. In: Geers, AE.; Moog, JS., editors. The Volta Review Vol. 96 Effectiveness of cochlear implants and tactile aids for deaf children: The sensory aids study at Central Institute for the Deaf. Washington DC: A.G. Bell Association for the Deaf; 1994. p. 97-108.

House AS, Williams CE, Hecker MHL, Kryter KD. Articulation-testing methods: Consonantal differentiation with a closed-response set. Journal of the Acoustical Society of America. 1965; 37:158–166. [PubMed: 14265103]

Johnson K. The role of perceived speaker identity in F0 normalization of vowels. Journal of the Acoustical Society of America. 1990; 88:642–654. [PubMed: 2212287]

Joos MA. Acoustic phonetics. Language. 1948; 24(Suppl. 2):1–136.

Karl J, Pisoni DB. The role of talker-specific information in memory for spoken sentences. Journal of the Acoustical Society of America. 1994; 95:2873.

Kirk KI, Pisoni DB, Osberger MJ. Lexical effects on spoken word recognition by pediatric cochlear implant users. Ear & Hearing. 1995; 16:470–481. [PubMed: 8654902]

Klatt, DH. Review of selected models of speech perception. In: Marslen-Wilson, W., editor. Lexical representation and process. Cambridge, MA: MIT Press; 1989. p. 169-226.

Kucera, F.; Francis, W. Computational analysis of present-day American English. Providence, RI: Brown University Press; 1967.

Luce, PA.; Pisoni, DB.; Goldinger, SD. Similarity neighborhoods of spoken words. In: Altmann, GT., editor. Cognitive models of speech processing: Psycholinguistic and computational perspectives. Cambridge, MA: MIT Press; 1990. p. 122-147.

Luce, PA. Research on Speech Perception, Technical Report No. 6. Bloomington, IN: Indiana University; 1986. Neighborhoods of words in the mental lexicon.

Marslen-Wilson WD. Functional parallelism in spoken word recognition. Cognition. 1987; 25:71–102. [PubMed: 3581730]

Martin CS, Mullennix JW, Pisoni DB, Summers WV. Effects of talker variability on recall of spoken word lists. Journal of Experimental Psychology: Learning, Memory, & Cognition. 1989; 15:676–684.

Miyamoto RT, Osberger MJ, Todd SL, Robbins AM, Stroer BS, Zimmerman-Phillips S, Carney AS. Variables affecting implant performance in children. Laryngoscope. 1994; 104:1120–1124. [PubMed: 8072359]

Mullennix JW, Pisoni DB. Stimulus variability and processing dependencies in speech perception. Perception & Psychophysics. 1990; 47:379–390. [PubMed: 2345691]

Mullennix JW, Pisoni DB, Martin CS. Some effects of talker variability on spoken word recognition. Journal of the Acoustical Society of America. 1989; 85:365–378. [PubMed: 2921419]

Nearey TM. Static, dynamic, and relational properties in vowel perception. Journal of the Acoustical Society of America. 1989; 85:2088–2113. [PubMed: 2659638]

Nittrouer S, Boothroyd A. Context effects in phoneme and word recognition by young children and older adults. Journal of the Acoustical Society of America. 1990; 87:2705–2715. [PubMed: 2373804]

Owens E, Kessler D, Schubert E. The minimal auditory capabilities (MAC) battery. Hearing Aid Journal. 1981; 34:9–34.

Peterson GE, Barney HL. Control methods used in a study of the vowels. Journal of the Acoustical Society of America. 1952; 24:175–184.

Pisoni DB. Speech perception: some new directions in research and theory. Journal of the Acoustical Society of America. 1985; 78:381–388. [PubMed: 4031245]

Pisoni DB. Long-term memory in speech perception: Some new findings on talker variability, speaking rate and perceptual learning. Speech Communication. 1993; 13:109–125. [PubMed: 21461185]

Rabinowitz WM, Eddington DK, Delhorne LA, Cuneo PA. Relations among different measures of speech reception in subjects using a cochlear implant. Journal of the Acoustical Society of America. 1992; 92:1869–1881. [PubMed: 1401531]

Rosen, S.; Fourcin, AJ.; Abberton, E.; Walliker, JR.; Howard, DM.; Moore, BCJ.; Douek, EE.; Frampton, S. Assessing assessment. In: Schindler, RA.; Merzenich, MM., editors. Cochlear implants. New York: Raven Press; 1985. p. 479-498.

Sommers MS. The structural organization of the mental lexicon and its contribution to age-related deficits in spoken word recognition. Psychology & Aging. 1996; 11:333–341. [PubMed: 8795062]

Sommers MS, Nygaard LC, Pisoni DB. Stimulus variability and spoken word recognition. I. Effects of variability in speaking rate and overall amplitude. Journal of the Acous-tical Society of America. 1994; 96:1314–1324.

Tyler RS, Lowder MW, Otto SR, Preece JP, Gantz BJ, McCabe BF. Initial Iowa results with the multichannel cochlear implant from Melbourne. Journal of Speech & Hearing Research. 1984; 27:596–604. [PubMed: 6549200]
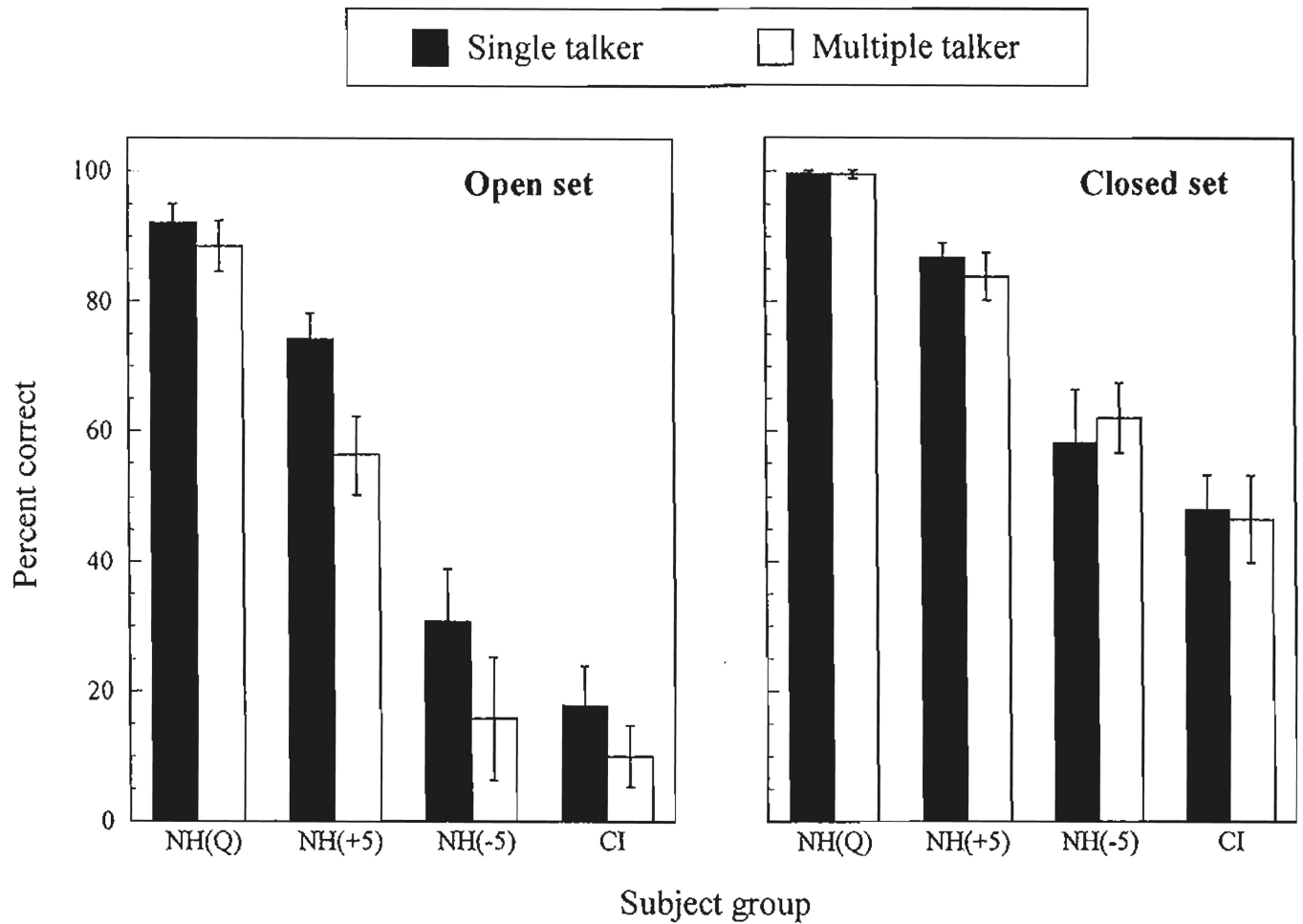
**Figure 1.**
Comparison of identification scores for words produced by single (dark bars) and multiple (open bars) talkers collapsed across lexical difficulty. The left side of the figure displays data obtained with an open-set response format, and the right side shows results obtained with a closed-set test. The four subject groups are normal-hearing tested in quiet (NH(Q)), normal-hearing tested at +5 (NH(+5)) and −5 (NH(−5)) S/N ratios, and cochlear implant patients (CI). Error bars in this and all subsequent figures indicate standard errors of the mean.
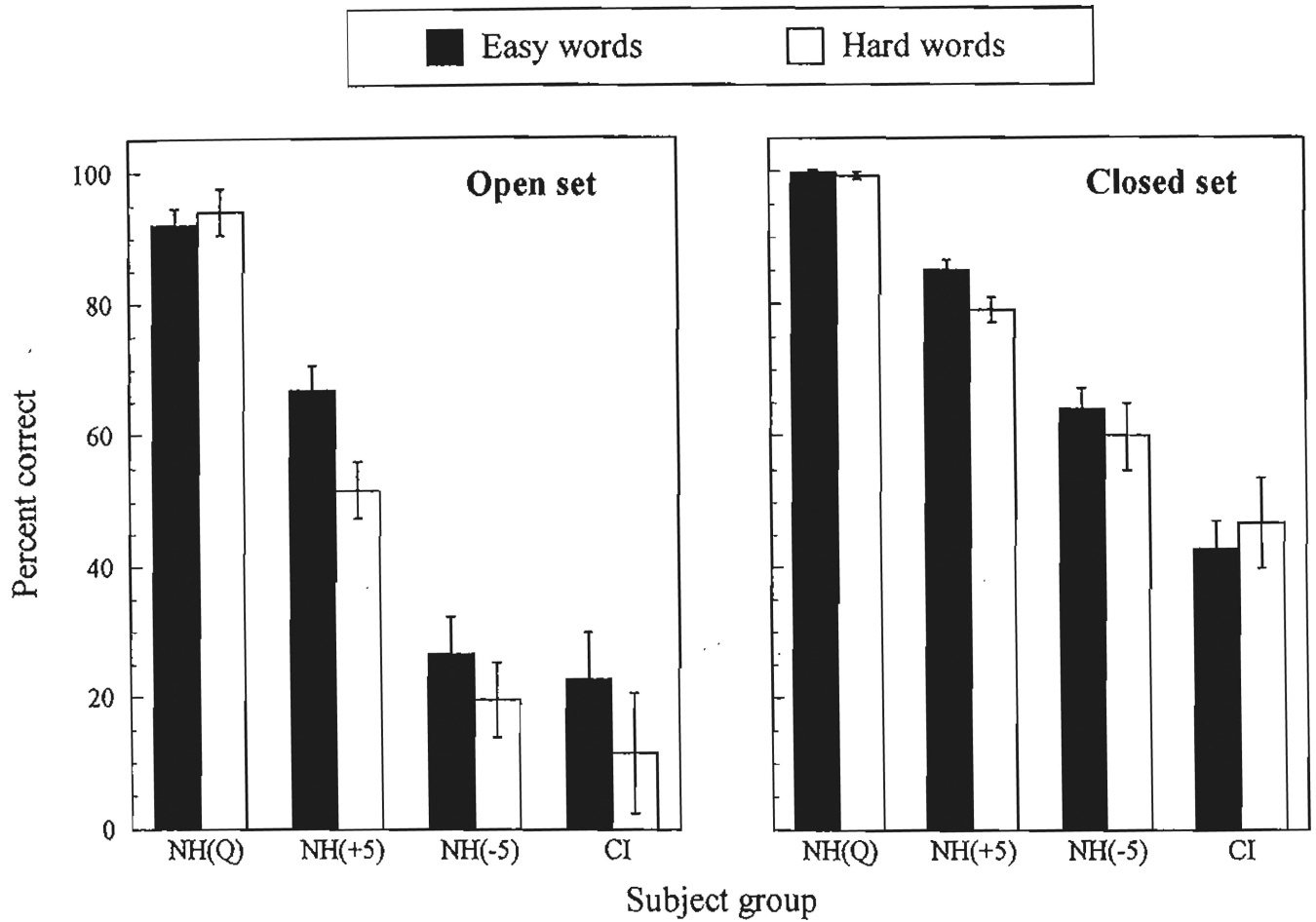
**Figure 2.**
Same as Figure 1 except the data show the effects of lexical difficulty (easy versus hard words) collapsed across single and multiple talkers.
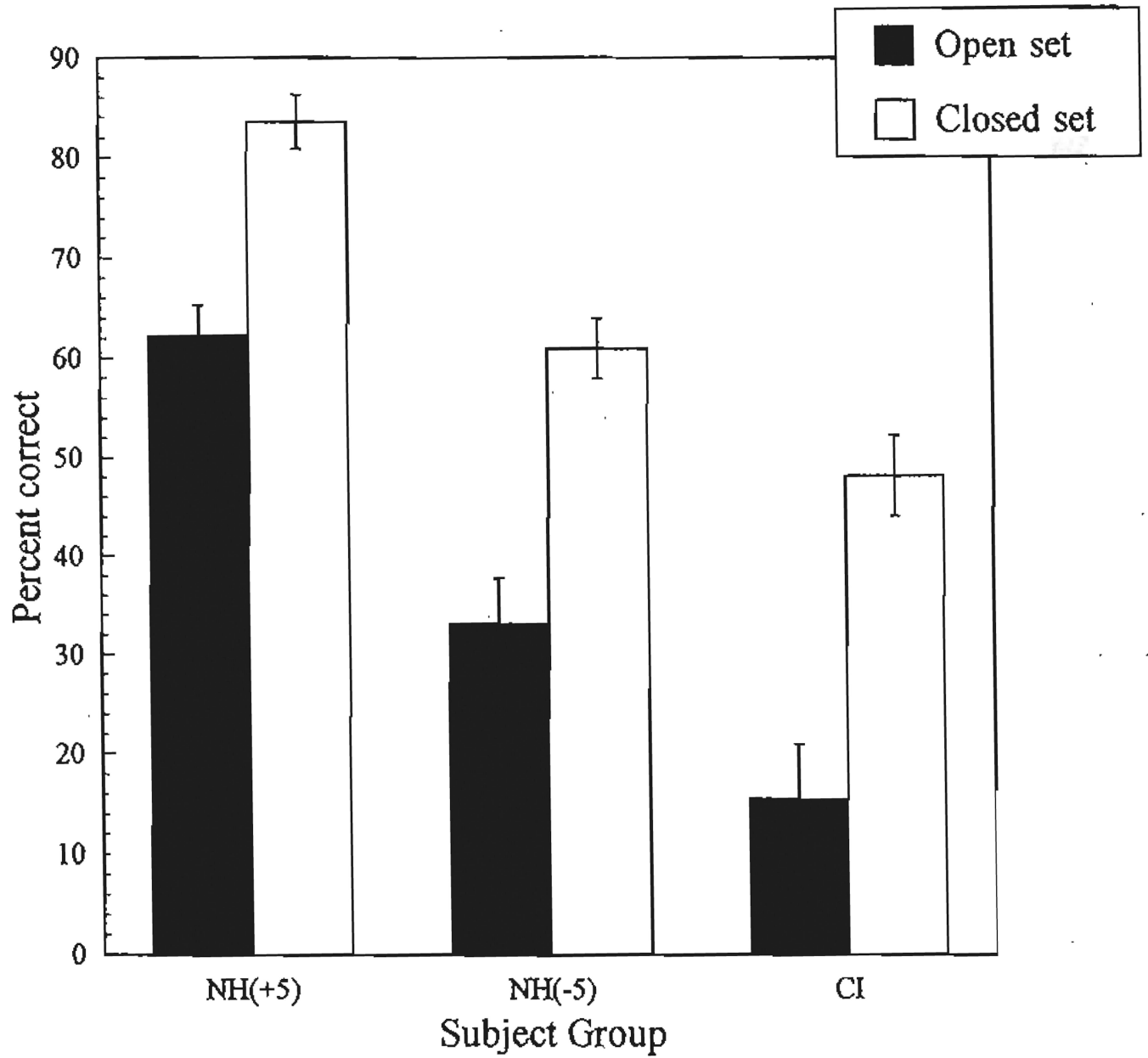
**Figure 3.**
Comparison of open- and closed-set test performance. Group names are the same as in Figure 1.
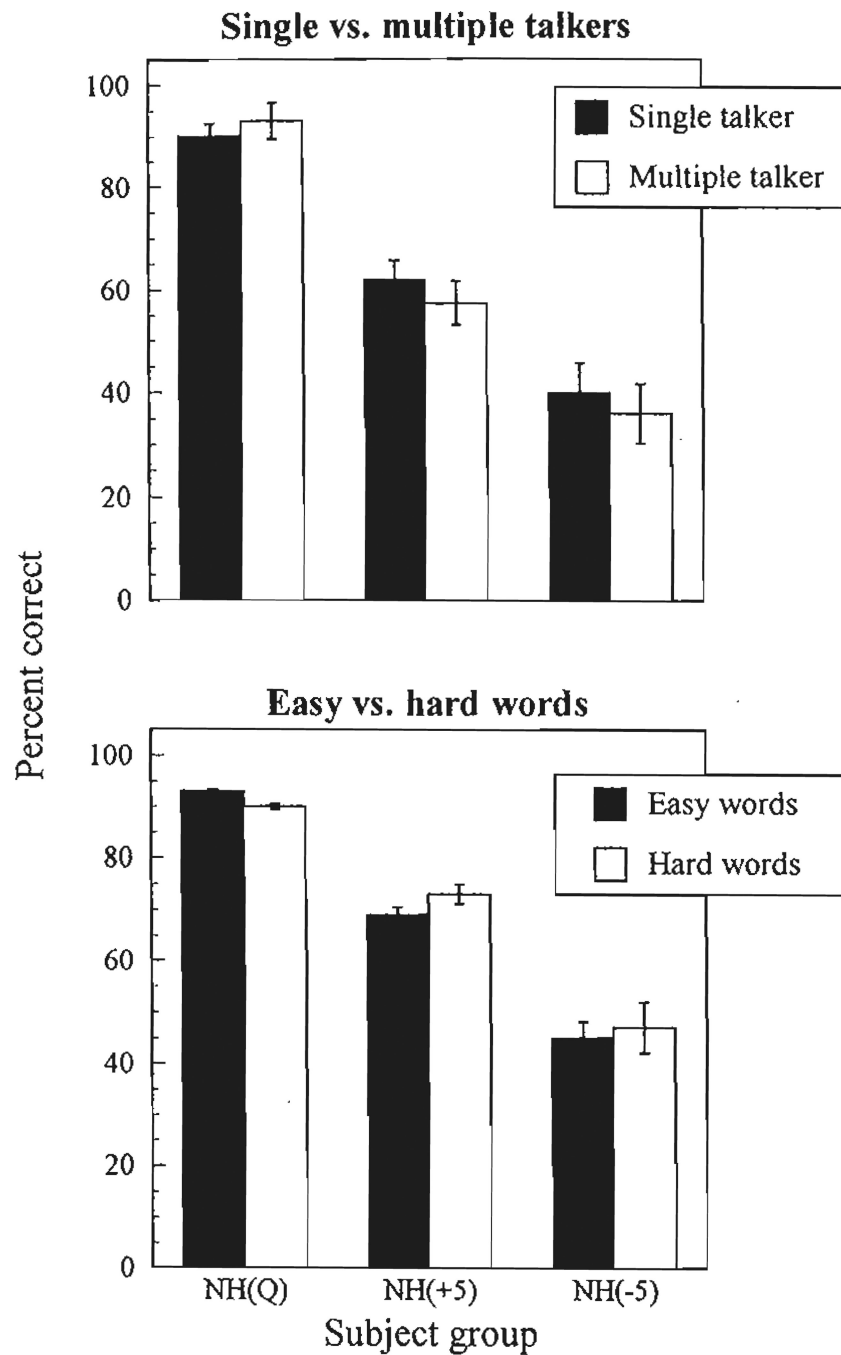
**Figure 4.**
Effects of talker variability (top) and lexical difficulty (bottom) as measured in a closed-set test designed to maximize confusability between response alternatives and the target item. The group names are the same as in Figure 1.