Journal of Speech and Hearing Research, Volume 32, 347–352, June 1989

DEVELOPMENT OF THE SPEECH INTELLIGIBILITY RATING (SIR) TEST FOR HEARING AID COMPARISONS

ROBYN M. COX
Memphis State University and Veteran's Administration Medical Center, Memphis

D. MICHAEL McDANIEL
Memphis State University

The Speech Intelligibility Rating (SIR) Test has been developed for use in clinical comparisons of hearing aid conditions. After listening to a short passage of connected speech, subjects generate a rating proportional to its intelligibility using an equal-appearing interval scale from 0 to 10. Before test passages are presented, the signal-to-babble ratio (SBR) is adjusted to a level that elicits intelligibility ratings of 7–8 for a "setup" passage. Then, with SBR held constant, three or more test passages are rated and the results averaged for each aided condition. This paper describes the generation of recorded test materials and their investigation using normally hearing listeners. Based on these data, a critical difference of about 2 scale intervals is recommended. A future paper will deal with results for hearing-impaired subjects.

Although many different speech intelligibility tests have been suggested for use in hearing aid evaluation and comparison, most are subject to criticism of their reliability or validity or of their instrumentation demands. To achieve adequate reliability, tests that call for word recognition, such as the Northwestern University Auditory Test #6 (NU-6; Tillman & Carhart, 1966), the Speech Perception in Noise (SPIN; Kalikow, Stevens, & Elliot, 1977), and the Connected Speech Test (CST; Cox, Alexander, & Gilmore, 1987), require administration of so many items that the test may be unacceptably long (see Thornton & Raffin, 1978, for a discussion of these issues). Also, because hearing aid wearers are exposed to connected speech in their everyday lives, validity requirements dictate that the metric that quantifies understanding of hearing-aid-processed speech should have a known relationship to understanding for connected speech. Tests that require recognition of isolated words (such as NU-6 and W-22), although widely used in hearing aid evaluation, have not fared well on this score. Word scores have been shown to be rather poor predictors of scores for connected speech and everyday performance with hearing aids (e.g., Giolas & Epstein, 1963; O'Neill, 1957; Schiavetti, Sitler, Metz, & Houde, 1984; Walden, Schwartz, Williams, Holum-Hardegan, & Crowley, 1983).

Another approach to quantifying speech understanding involves the use of subjective judgments of the intelligibility of speech. Subjective judgment approaches are attractive because they can use connected speech as the stimulus and produce results relatively rapidly. One such approach requires the listener to assign ratings to samples of speech. Typically, the listener is exposed to a brief passage of connected speech and then supplies a rating proportional to its intelligibility on an equal-appearing interval scale. One end of the scale is defined to represent full intelligibility and the other end represents zero intelligibility. This type of procedure has a brief administration time, produces measures of connected speech intelligibility, does not require elaborate instrumentation, and may be performed with the listener wearing real hearing aids. In addition,several investigators have reported data indicating that intelligibility ratings of speech are sensitive, reliable, and valid (e.g., Nakatani & Dukes, 1973; Peters, 1965; Speaks, Parker, Harris, & Kuhl, 1972). Thus, this approach seems potentially well suited to the task of clinical hearing aid comparisons.

Cox and McDaniel (1984) reported an investigation that explored the feasibility of using speed intelligibility ratings in clinical hearing aid evaluation. The study was designed to assess the validity and sensitivity of intelligibility ratings when employed in a context simulating a hearing aid evaluation. Normal-hearing subjects rated the intelligibility of 35-s passages of connected speech produced by 3 talkers and processed by four hearing aids having rather similar frequency responses. Each subject rated each condition three times and the final rating for that subject-condition was the mean of these three. The results indicated that: (a) the mean ratings were valid quantifiers of speech intelligibility in the various conditions; (b) the approach was more sensitive to differences among hearing aids when the signal-to-babble ratio (SBR) was adjusted to produce a moderately challenging (not too difficult) listening condition; (c) when three intelligibility ratings were averaged per hearing aid, the highest-ranked (best) hearing aid was significantly differentiated from the third- and fourth-ranked instruments, but not necessarily from the second-best instrument; (d) hearing aid rankings depended somewhat on the talker so that the best hearing aid was different for different talkers.

Overall, this initial investigation produced results that were encouraging with respect to the application of speech intelligibility ratings in clinical hearing aid comparisons. The purpose of the present paper is to report the development of a clinical test employing speech intelligibility ratings—the Speech Intelligibility Rating (SIR) test. The SIR test was developed with the goal of produc-

0022-4685/89/3202-0347$01.00/0

ing a practical, sensitive, speech-based test with high face validity. Development proceeded in the following stages: (a) new recordings were made of a large pool of potential test passages, (b) a group of normal hearers provided intelligibility ratings for each passage, (c) 20 passages were selected as the test passages on the basis of the data for normal hearers, (d) additional passages were selected for validity checking and practice.

METHOD

Test Passages

The 72 passages of connected speech used in the earlier study were equated *a priori* on the basis of length, subject matter, vocabulary, sentence structure, and reading level. However, comments from subjects suggested that these efforts did not result in equivalent intelligibility in the recorded versions of the passages. Consequently, it was decided that the passages used in the clinical test would be empirically equated after recording. The passages were slightly revised to be more uniform in length (108–110 words) and to ensure that the topic of the passage was mentioned in the first four words and again 3–7 times within the passage. In addition, passages were constructed with a logical breakpoint in the middle to allow for administration of half-length passages. When read aloud by the talker described below, mean passage length was about 48 s.

A longer passage (250 words) was also developed using the same guidelines as the shorter passages. This passage was intended for use in setting the SBR for the test as well as for practice with the rating task. It is referred to as the "setup" passage.

Talker

As noted previously, Cox and McDaniel (1984) found that intelligibility ratings of hearing-aid-processed speech were somewhat dependent on the talker. In addition, for some talkers, hearing aid rankings were somewhat dependent on SBR. These findings have important implications for hearing aid comparisons because they indicate that the highest-rated hearing aid may be determined in part by the talker who produced the speech materials and by the SBR for the test. The talker chosen to produce the materials for the SIR test was Talker 3 in the earlier study. This individual was a male professional television broadcaster who produced exceptionally clear speech, without a regional dialect, at a relatively slow rate (about 3 syllables/s). In the earlier study, changes in SBR for Talker 3's recordings did not affect hearing aid rankings. This motivated his choice as the talker for the clinical test: It was anticipated that the setting for SBR would be a relatively minor factor in the obtained hearing aid rankings if the new test used this talker and the same multivoice babble as in the earlier study. In addition, the clear, nonregional speech produced by Talker 3 is at least partly intelligible to almost all hearing aid wearers. Thus, the SIR test is useful for a wide range of individual hearing losses.

Recordings

The 72 experimental connected speech passages and the setup passage were produced by the chosen talker in a sound-treated room with ambient noise level of 43 dB(C)/17 dB(A) and mean (250–4000 Hz) reverberation time (RT_{60}) <50 msec. The talker was instructed to read the passages in a natural manner. The microphone (Bruel and Kjaer, type 4145) was positioned at mouth level, 44 cm from his mouth. The passages were recorded on magnetic tape (Revox A-77 recorder). The recorded passages were subsequently intensity-equalized so that their VU readings (frequent peaks) were within ±0.5 dB. Next, the multivoice babble was added on the second track for use as a competing signal. Finally, a speech-shaped noise was recorded on each track for calibration purposes. Recordings were adjusted so that the level of frequent peaks for passages, babble, and calibration noise were at the same VU level. Figure 1 illustrates the final long-term RMS average 1/3-octave spectra for the speech passage and competing babble recordings.

Subjects

Twenty young adults (3 male, 17 female) with normal hearing in the test ear (≤15 dB HL: 250–8000 Hz) provided intelligibility ratings of each of the recorded

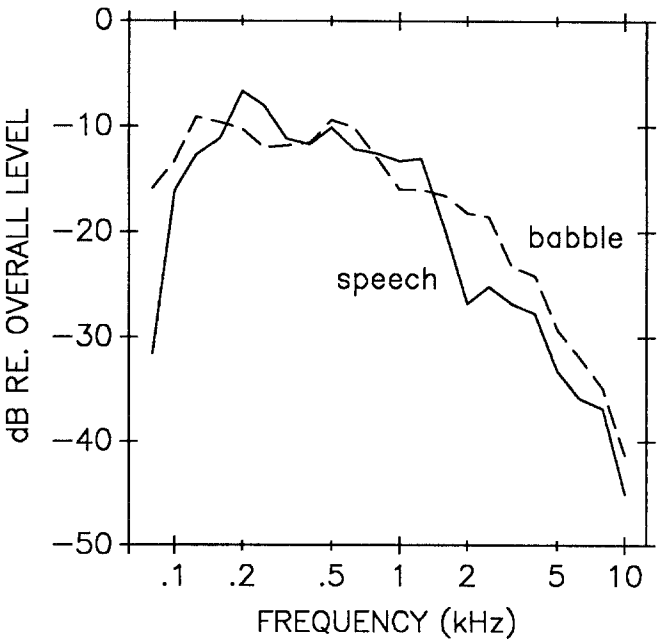

FIGURE 1. Long-term RMS average 1/3-octave band levels of recorded speech passages (solid line) and competing babble (dashed line).

passages. None of the subjects had previous experience in any type of speech intelligibility assessment.

Procedure

Instructions. Subjects were instructed to rate the intelligibility of each passage on an equal-appearing interval scale that ranged from 0 to 10. Figure 2 shows a facsimile of the scale used by the subjects. Instructions were as follows.

> You will be hearing a man's voice through one of the earphones. In the background there will be the sound of several people talking. This background noise will be distracting to you but please try to ignore it as much as possible. Concentrate on listening to the man talking and think about how well you understand the words he is saying. At the end of the passage, I want you to give me a score from 0 to 10 which should reflect how well you understood the words the man said. If you could not understand *any* of the words, you should give the passage a score of zero. If you understood *all* of the words, you should give the passage a score of 10. If you understood some, but not all, of the words you will give a score between 1 and 9, depending on how many words in the passage you understood. For example, if you think you understood about half the words, you should give that passage a score of 5. If you only missed a few words, give the passage a 9. On the other hand, if you only understood a few words, you should give the passage about a 1. If you understood about 30% of the words, give the passage a 3, and so on. Since each passage has about 100 words, you could move up one number in score for every 10 words you understood. Feel free to use any number on the scale that seems appropriate but do not mark between the numbers. Try to be as accurate as you can. Do not worry if you do not understand what the passage is about, I am only interested in how many individual *words* you understood.

Determining test SBR: The recorded speech and competing babble were replayed (Onkyo TA-2080 cassette player), mixed, and presented to the test ear via a supraaural earphone (Telephonics TDH-50). The speech was presented at a fixed level corresponding to 65 dB SPL (RMS, slow) for the calibration noise in a 6 cm^3 coupler. The level of the babble was adjusted individually for each listener. Before presentation of the experimental passages, the setup passage was presented to provide practice with the response task and set the SBR for the test. Because the initial investigation of the rating procedure (Cox & McDaniel, 1984) indicated that the most sensitive ratings were obtained when the listening condition was not too difficult, the target SBR was one that produced ratings of 7–8 for the setup passage. The SBR was determined as follows: (a) a 20-s portion of the setup passage was presented at an SBR of 0 dB and an intelligibility rating was then elicited; (b) if the rating was >8 or <7, the level of competing babble was increased or decreased, respectively, by 5 dB; (c) another 20-s portion of the setup passage was presented and another intelligibility rating elicited; (d) additional ratings were obtained, using new portions of the setup passage, with the competing babble level varied to produce ratings that bracketed the desired value; (e) when two or three ratings of 7 or 8 were obtained at the same SBR, this value was chosen as the test SBR. All experimental passages were presented at this SBR. Chosen SBRs ranged from 0 to −5 dB with a median of −2.5 dB.

Administration of test passages: Following SBR selection, the 72 test passages were presented in pseudorandom order. Subjects were given two breaks to minimize fatigue. A portion of the SBR passage was presented after each break to reorient the listeners to the task. An intelligibility rating was elicited after each passage.

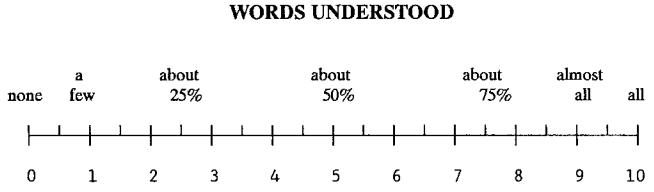

FIGURE 2. Facsimile of scale used by subjects in producing intelligibility ratings of speech passages.

RESULTS

The data consisted of 72 intelligibility ratings from each of 20 subjects. The mean rating and standard deviation of ratings were computed for each passage. Figure 3 shows the outcome. Each symbol in this figure represents a passage. Mean rating scores varied from 2 to 8, indicating a wide range of intelligibility in the recorded passages in spite of the a priori efforts to equate them. Standard deviations of ratings for individual passages ranged from about 1.2 to 2.5 scale intervals. It should be kept in mind that these values incorporate the variability in SBR conditions across subjects as well as the inherent variability of ratings for passages.

Based on these data, 20 passages with mean ratings in the range 6–7 were chosen as the test passages. These are shown as the filled circles. In addition, 6 passages, depicted using open squares, were chosen as "validity" passages. This group encompasses a wide range of mean ratings, indicating substantial variation in intelligibility

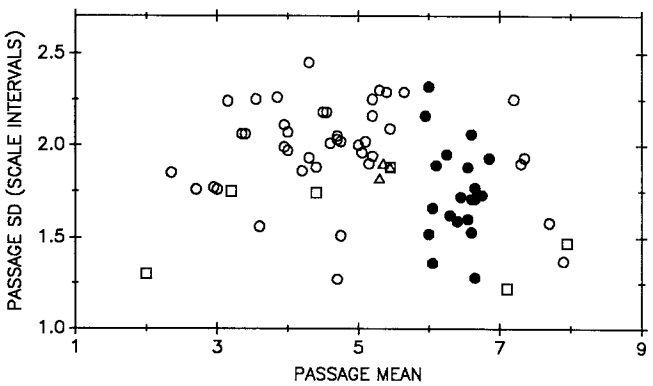

FIGURE 3. Mean rating and standard deviation of ratings for each experimental connected speech passage. Each symbol denotes one passage. Filled circles = test passages, open squares = validity passages, open triangles = practice passages.

across passages (the least-intelligible validity passage was rerecorded backwards to ensure total unintelligibility). These may be presented to verify that a listener is responding on the basis of intelligibility rather than some other aspect of the stimulus. A listener who does not rate the validity passages in the appropriate order, or nearly so, is probably not a good candidate for the test. Finally, 3 passages with similar mean scores were selected for use as practice passages; these are shown with open triangles (one passage appears in both the validity and practice groups).

Determination of Critical Differences

The SIR test is designed to compare intelligibility for a particular individual under different conditions. It is important to estimate how much difference between SIR test scores is necessary before we may reasonably conclude that two listening conditions (e.g., hearing aid A vs. hearing aid B) are significantly different. This requires a determination of the error of estimate that is likely for any single score. Clearly, two conditions cannot be concluded to be different if the difference between their scores (mean ratings) has a high probability of occurring by chance. In general, a critical difference (CD) is defined as the maximum difference between two scores that is likely to occur by chance alone. A difference between two scores that is greater than the critical difference is likely to be due to real differences between listening conditions. For example, a 95% CD will be exceeded by chance on only 5% of comparisons.

If we assume that the intelligibility ratings of the 20 test passages are normally distributed for a fixed listening condition, a critical difference (CD) between two scores may be computed by determining (a) the standard deviation (SD) of the distribution of mean ratings for n passages (where n = number of passages rated and averaged per score) and (b) the SD of the distribution of differences between pairs of scores from the distribution of mean ratings. If the actual difference between 2 scores obtained under different listening conditions, each based on the average of n passage ratings, is near a tail of this distribution of differences it is unlikely that the two scores came from conditions having the same intelligibility.

Using this rationale, the CD may be computed with the following equation (Ferguson, 1966, pp. 140, 147):

$$CD = \frac{(\alpha)\sqrt{2}(SD_p)}{\sqrt{n}} \sqrt{\frac{P-n}{P-1}}$$

Where α = 1.96 or 1.65 for the 95% or 90% CD, respectively; SD_p = SD of the distribution of ratings of all 20 passages under a fixed listening condition; P = total number of passages in the pool; and n = number of ratings averaged per score.

To determine SD_p for the typical listener, within-subject standard deviations were computed across the 20 test passages for each of the subjects. They ranged from 0.6 to 2.0 scale intervals with a mean value (square root of mean variance) of 1.4 scale intervals. Assuming that three separate ratings are obtained and averaged per condition (n = 3), the suggested 95% CD for the SIR test (for normal hearers) may be computed:

$$CD_{(95)} = \frac{1.96\sqrt{2}(1.4)}{\sqrt{3}} \sqrt{\frac{20-3}{19}}$$

$$= 2.1 \text{ scale intervals}$$

Similarly, the 90% CD = 1.8 scale intervals. If 2 scale intervals are used as the CD, this corresponds to a 93% criterion (i.e., this difference is exceeded by chance on 7% of comparisons).

DISCUSSION

The test passages chosen in this investigation have approximately equal overall intelligibility ratings for normal hearers. This does not necessarily ensure that these passages are equally intelligible for individuals with a variety of hearing loss configurations. If the test passages are less equivalent for hearing-impaired persons, this will result in larger critical differences or the necessity of obtaining more than three ratings in each aided condition. Additional study of the test using hearing-impaired listeners is necessary to address this issue.

The method used to estimate critical differences makes the assumption that the within-subject variability in rating scores across passages is independent of a subject's mean rating location on the 0 to 10 scale. However, these intelligibility ratings are essentially estimated proportions; therefore, it seems possible that within-subject variability has a relationship to overall performance level that is qualitatively similar to the analogous relationship observed for percentage scores. For percentage scores, variability is reduced when overall performance is at either end of the scale (see, e.g., Thornton & Raffin, 1978, Figure 1). On the other hand, variability of percentage scores remains relatively constant for overall performance levels between about 20% and 80%, especially when the number of test items is large. If we assume that the characteristic distribution of SIR test ratings is similar to that of percentage scores, it follows that the SBR for administration of the SIR test should be adjusted to produce intelligibility ratings in the 2–8 range of the scale. If ratings are within this range, CDs computed using the method described above should be reasonably accurate.

As mentioned earlier, the test passages were constructed with a logical break in the middle to allow for administration of half-length passages. Pilot work with half-length passages has suggested that subjects tend to give somewhat higher ratings (about 0.5 scale intervals, on the average) to half-length passages than they do to full-length passages. This suggests that ratings for full-length passages should not be directly compared with

ratings for half-length passages. More investigation of this issue is planned. As Figure 3 illustrates, when the SBR was adjusted to produce ratings of 7–8 for the setup passage, the test passages produced ratings of 6–7. This may indicate that the test passages are somewhat more difficult to understand than the setup passage. In addition, it seems possible that this effect is partly due to a tendency for listeners to award higher ratings to shorter passages than to longer ones (recall that in setting the SBR, 20-s segments of speech were used, whereas the test passages are about 48 s long).

Although the instructions for this investigation prohibited subjects from choosing fractional values on the rating scale, several subjects expressed a desire to choose numbers between the integer rating values. In subsequent work with the SIR test, hearing-impaired and normal-hearing subjects have been permitted to choose fractional values on the rating scale. Although most subjects still select integer ratings, a small proportion do exercise this option to choose fractional values.

On the basis of experience with the SIR test in clinical settings with hearing-impaired subjects, a tentative protocol for its use has been devised and is shown in Figure 4. The suggested clinical protocol differs from the one used in this investigation as follows.

1. In the clinical setting, instructions should be memorized and delivered verbally. Subjects can be encouraged to think in terms of percentages if they are comfortable with this concept.
2. A test level of 60 dB SPL (RMS, integrate) measured in the sound field for a single talker, closely simulates the level

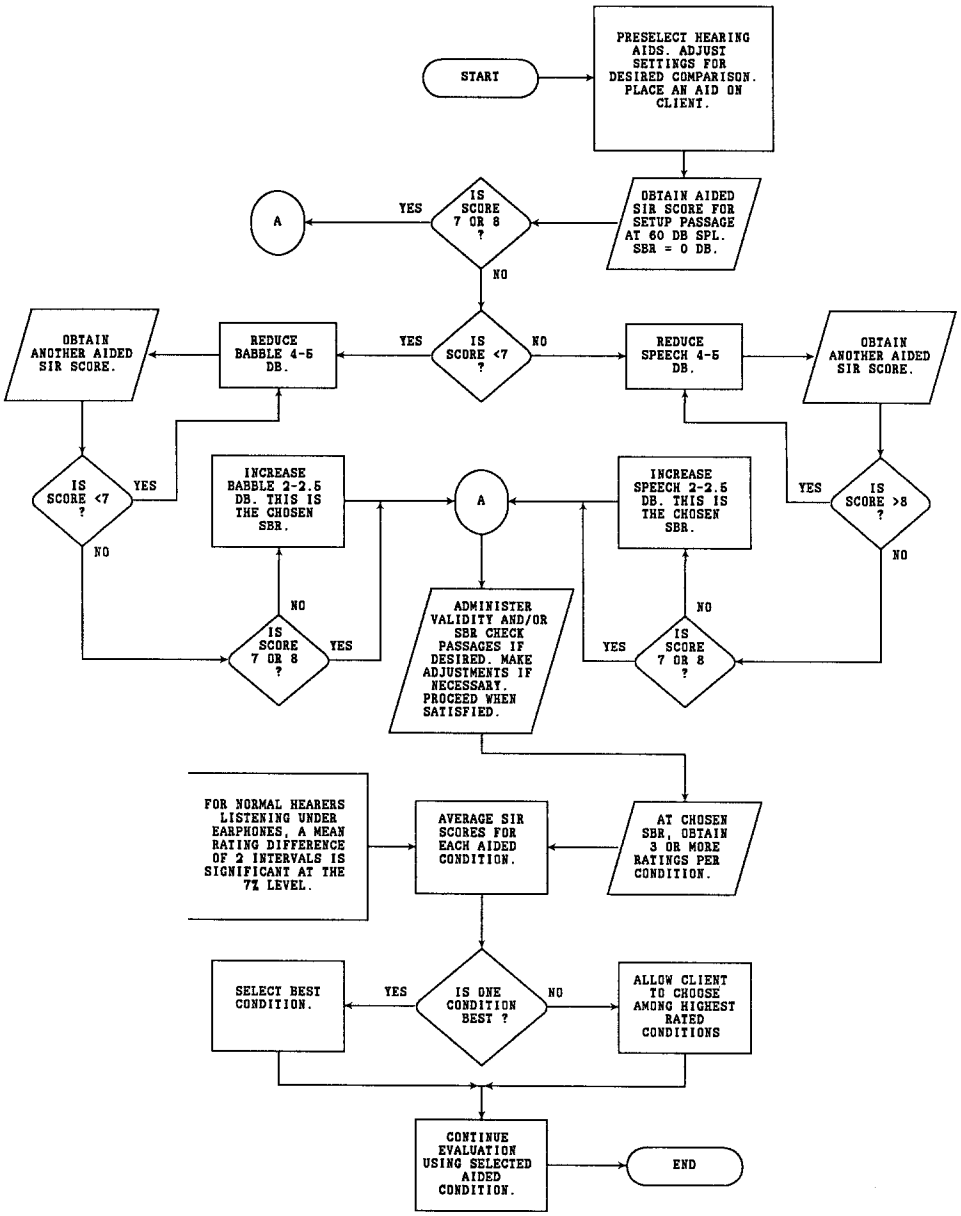

FIGURE 4. Flow diagram showing tentative protocol for administration of the SIR test for clinical hearing aid comparisons. SBR = signal-to-babble ratio.

of everyday conversations (Pearsons, Bennett, & Fidell, 1977). This level can also be approximated by 65 dB SPL (RMS, fast), measured for speech *peaks* in the sound field.
3. To adjust the SBR appropriately, either the speech or the babble may be lowered, but should not be raised above the starting level. This maintains the overall intensity at about the same level regardless of SBR.
4. Aided conditions should be varied randomly between ratings.

It is important to keep in mind that the critical differences computed in this study were for normal hearers listening under earphones. For hearing-impaired persons listening in the sound field, the critical differences may be larger. Thus, the CDs reported in this study do not provide clear guidelines for interpreting the significance of differences between hearing aid ratings obtained with hearing-impaired-subjects. Continued development of the SIR test has included work with hearing-impaired subjects to (a) assess the sensitivity of the ratings to differences among aided conditions, (b) determine the number of ratings necessary to produce reliable hearing aid rankings, (c) estimate critical differences appropriate for hearing-impaired subjects, and (d) explore the use of half-length passages. The results will be the subject of a future paper.

ACKNOWLEDGMENTS

Supported in part by VA Rehabilitation Research and Development funds. Also supported in part by the Center for Research Initiatives and Strategies for the Communicatively Impaired (CRISCI), Memphis State University.

Part of this work was reported at the National Convention of the American Speech-Language-Hearing convention, Detroit, MI, 1986.

REFERENCES

COX, R. M., ALEXANDER, G. C., & GILMORE, C. (1987). Development of the Connected Speech Test (CST). *Ear and Hearing, 8,* 119S–126S.

COX, R. M., & MCDANIEL, D.M. (1984). Intelligibility ratings of continuous discourse: Application to hearing aid selection. *Journal of the Acoustical Society of America, 76,* 758–766.

FERGUSON, G. A. (1966). *Statistical analysis in psychology and education.* New York: McGraw-Hill Book Company.

GIOLAS, T. G., & EPSTEIN, A. (1963). Comparative intelligibility of word lists and continuous discourse. *Journal of Speech and Hearing Research, 6,* 349–358.

KALIKOW, D. N., STEVENS, K. N., & ELLIOT, L. L. (1977). Development of a test of speech intelligibility in noise using sentence materials with controlled word predictability. *Journal of the Acoustical Society of America, 61,* 1337–1361.

NAKATANI, L. H., & DUKES, K. D. (1973). A sensitive test of speech communication quality. *Journal of the Acoustical Society of America, 53,* 1083–1092.

O'NEILL, J. J. (1957). Recognition of intelligibility test materials in context and in isolation. *Journal of Speech and Hearing Disorders, 22,* 87–90.

PEARSONS, K. S., BENNETT, R. L., & FIDELL, S. (1977). *Speech levels in various noise environments.* (Report EPA 600/1-77-025). United States Environmental Protection Agency.

PETERS, R. W. (1965). *A rating scale technique for the measurement of speaker intelligibility* (AD-629 308). Springfield VA: Clearinghouse for Federal Scientific and Technical Information, U.S. Department of Commerce.

SCHIAVETTI, N., SITLER, R. W., METZ, D. E., & HOUDE, R. A. (1984). Prediction of contextual speech intelligibility from isolated word intelligibility measures. *Journal of Speech and Hearing Research, 27,* 623–626.

SPEAKS, C., PARKER, B., HARRIS, C., & KUHL, P. (1972). Intelligibility of connected discourse. *Journal of Speech and Hearing Research, 15,* 590–602.

TILLMAN, T. W., & CARHART, R. (1966). *An expanded test for speech discrimination utilizing CNC monosyllabic words.* Brooks Air Force Base, TX: Northwestern University Auditory Test No. 6. USAF School of Aerospace Medicine Technical Report.

THORNTON, A.R., & RAFFIN, M.J.M. (1978). Speech-discrimination scores modelled as a binomial variable. *Journal of Speech and Hearing Research, 21,* 507–518.

WALDEN, B.E., SCHWARTZ, D.M., WILLIAMS, D.L., HOLUM-HARDEGAN, L.L., & CROWLEY, J.M. (1983). Test of the assumptions underlying comparative hearing aid evaluations. *Journal of Speech and Hearing Disorders, 48,* 264–273.

Received June 16, 1988
Accepted September 26, 1988

Requests for reprints should be sent to Robyn M. Cox, Memphis State University, Memphis Speech & Hearing Center, 807 Jefferson Ave., Memphis, TN 38105.