0196/0202/87/0805-119S$02.00/0
EAR AND HEARING

Vol. 8, No. 5 Supplement
Printed in U.S.A.

Development of the Connected Speech Test (CST)*

Robyn M. Cox, Genevieve C. Alexander, and Christine Gilmore

Department of Audiology and Speech Pathology, Memphis State University [R. M. C.], and Memphis V.A. Medical Center [G. C. A., C. G.], Memphis, Tennessee

ABSTRACT

This paper describes the first phase in the development of the Connected Speech Test (CST). This test of intelligibility of everyday speech has been developed primarily for use as a criterion measure in investigations of hearing aid benefit. The test consists of 48 passages of conversationally produced connected speech. Each passage contains 25 key words for scoring. All passages are of equal intelligibility for the average normal hearer. Key words vary in intelligibility within a passage but span the same intelligibility range in all passages. Several passages are administered, and the results averaged, to yield a single intelligibility score. For pairs of scores, each based on mean performance across 4 randomly-chosen passages, the 95% critical difference is estimated to be about 14 rationalized arcsine units (rau). The performance-intensity function for the CST has a slope of 12 rau/dB signal-to-babble ratio. Investigations of the test are continuing with hearing-impaired listeners.

The most ubiquitous problem facing the hearing impaired is diminished understanding of speech (1, 2). As a result, the benefit received from a hearing aid is principally determined by the extent to which the instrument facilitates speech understanding in everyday life. The need to quantify hearing aid benefit is obvious. However, quantification of hearing aid benefit in terms of the difference between aided and unaided speech perception is problematic because no completely satisfactory method has been developed for measuring an individual's understanding of everyday connected speech.

Although tests of intelligibility for isolated words are useful in many diagnostic applications, they do not appear to be a valid means of quantifying intelligibility of connected speech. Intelligibility scores for lists of isolated words have been shown to be poor predictors of scores for connected speech when a linear prediction model was used (3, 4) and only modestly good predictors when a nonlinear model was used (5). In studies of hearing aid benefit, lists of monosyllabic words have not accurately predicted the best-scoring hearing aid on a questionnaire assessment (6) and they have failed to predict the everyday situations in which the greatest self-assessed benefit would be reported (7).

* This work was supported by Veterans Administration Rehabilitation Research and Development, Project No. 344.

This outcome is inconvenient but not surprising. Connected speech (defined here as meaningfully related sentences) contains many cues to intelligibility that are not found in isolated words. In addition to lexical, semantic, and syntactic redundancy, connected speech contains dynamic cues such as relative duration of fricatives and vowels that provide information about the probable meaning of the utterance.

Complete sentences are a much closer approximation to everyday speech than are isolated words. On theoretical grounds, tests employing meaningful, conversationally produced sentences as test items could be expected to result in valid quantification of intelligibility of everyday speech and, therefore, would be suitable for measurement of hearing aid benefit. However, because quantification of hearing aid benefit for a particular individual requires comparison of aided and unaided intelligibility scores, several equivalent test forms are required (to allow for testing of aided and unaided speech understanding under a variety of conditions). In addition, an acceptably small error of measurement is essential. If the error of measurement is too large, intelligibility differences between aided an unaided conditions often cannot confidently be evaluated.

Several sentence intelligibility tests have been devised (8–10). Equivalence of recorded forms has been established for one test, SPIN (11). The error of measurement for an individual score on the SPIN test has not been reported. However, because responses to SPIN test sentences are independent of each other and are scored as correct or incorrect, variability of scores for PL (low probability) and PH (high probability) lists can be estimated using the binomial model described by Thornton and Raffin (12). As these authors have shown, to achieve a small error of measurement for binomially distributed scores, it is necessary to administer a large number of test items per score. Since the SPIN test is composed of 200 PL items and 200 PH items, a large number of equivalent forms with small error of measurement cannot be obtained from this material.

This paper reports the first phase in the development of a test of intelligibility of everyday speech: the Connected Speech Test (CST). Although this test may eventually find numerous applications, it has been developed for the specific purpose of serving as a criterion measure in studies of hearing aid benefit. The overall objective is to produce

a test that has high content validity (i.e., consists of conversationally produced connected speech), a large number of equivalent forms, and an acceptably small error of measurement (sufficient to detect an intelligibility change equivalent to a signal-to-babble ratio change of 1–2 dB).

In the first phase of test development, the prospective test items have been evaluated using normal hearing listeners. In this phase, a version of the test that is suitable for use with normal hearers has been developed. This work is reported in the present paper. Subsequent efforts have been directed toward evaluating and refining the test for use with hearing-impaired persons.

DEVELOPMENT OF INITIAL TEST ITEMS

It was decided a priori that each test unit would be a passage of connected speech about a familiar topic and that the listener would be apprised of the topic in advance. It was expected that a single intelligibility score would be based on mean performance measured for several passages. However, the number of passages to be used per score was not decided in advance.

The initial pool of test items included 72 passages about familiar topics such as common plants, animals, and household objects. Each passage contained 10 syntactically simple sentences, 7 to 10 words in length. The topic word appeared in the first sentence. To control word familiarity, the basic vocabulary used to discuss each topic was derived from a children's educational reading source.

Several investigators (13–15) have reported that relative judgments of hearing-aid processed speech interacted with talker, that is, different hearing aids were optimal for different talkers. This observation presents a problem for a speech intelligibility test that is intended to measure hearing aid benefit because the benefit measured for a particular hearing aid will depend, in part, on the speech characteristics of the talker who produced the speech materials. It was decided, therefore, that the talker selected to produce the CST passages would be a talker having average intelligibility for conversationally produced speech.

To define the characteristics of average intelligibility and to select an appropriate talker, an investigation of the intelligibility of 3 male and 3 female talkers was performed (16). The average talker identified in that study was the talker who produced the CST passages. The chosen talker was a female who lacked a pronounced regional accent and was able to read prepared material in a manner similar to her spontaneous speech. Her long-term RMS speech spectrum is shown in Figure 1. Her articulation rate for the CST passages, measured a posteriori, was 4.8 syllables/sec which is within the range of 4.4 to 5.9 syllables/sec reported by Goldman-Eisler (17) for spontaneous utterances. The passages were recorded audiovisually on professional quality ¾ in videotape. The recording studio was rectangular in shape with a volume of 60.2 m^3. The recording microphone was located 40 cm from the talker's mouth. The effect of room acoustics on the recorded speech was minimal. In subsequent editing, the levels of

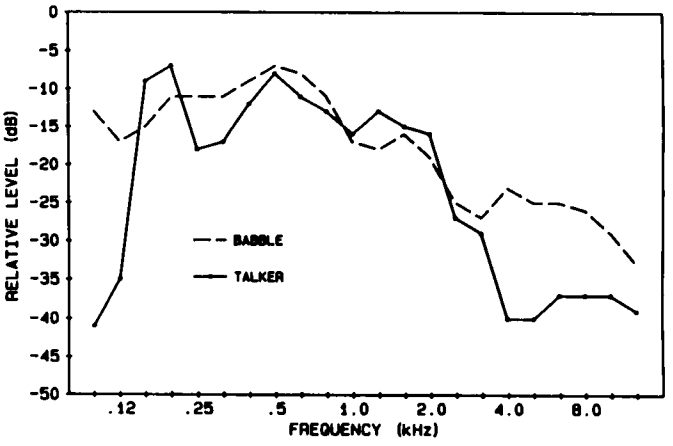

Figure 1. Long-term RMS ⅓-octave spectra of the talker who produced the CST passages and of the multitalker babble used as a competing signal.

the passages, expressed in dB Leq (equivalent continuous A-weighted level), were adjusted so that all were recorded at the same level, ±1 dB. A six-talker speech babble was recorded on the second channel of the master tape at the average level (dB Leq) of the passages (the same babble was used in the previous study of talker intelligibility). The long-term RMS spectrum of the babble is shown in Figure 1. The master recording was rerecorded onto optical disk for playback (Panasonic TQ2024F 2-channel optical disk player). Each disk accommodates 13.3 minutes of continuous speech. Three disks were required for the 72 passages. The frequency response of the optical disk system was flat ±1 dB in the range 150 Hz to 15 kHz.

Determination of Scoring Words (Key Words)

Each passage contained 45 to 55 potential key words that could be used for scoring purposes. Key words were defined as any word that was essential to convey the meaning. No sentence contained more than seven potential key words. As demonstrated by Duffy and Giolas (18), the particular words selected for scoring sentence intelligibility have a significant effect on the outcome of the test: some potential key words are more intelligible than others. This is undoubtedly due to a combination of the context in which the word is found and the acoustic token of the word produced in the particular recording.

Experiment 1: Key Word Selection

In order to select the words for scoring the CST passages, an investigation was performed to determine the probability of identification of each potential key word in the recorded passages. The objective was to provide a basis for selecting 25 key words for each passage, five words in each of five levels of difficulty. Only the audio portion of the test was used in this study; the video signal was not available to subjects. Thus, the scoring words identified in this investigation apply to the audio test only, not to the audiovisual test.

METHOD

Subjects

Thirty college undergraduates who passed a hearing screening (250–8000 Hz) at 15 dB HL audited the test passages. Five subjects served in each of six signal-to-babble (S/B) ratio conditions. ACT test scores (English and Social Science subtests) were available for 22 subjects. They ranged from the 33rd to the 99th percentile with a mean score at the 75th percentile. These scores indicate that the average verbal abilities of the subject group were somewhat above the median for college-bound high school seniors.

Playback Instrumentation

The CST passages and competing babble were replayed by the optical disk player and routed to attenuators to allow independent adjustment of signal-to-babble ratio. The two outputs were then mixed, amplified, and delivered to an insert earphone system (Etymotic Research ER-1) that was coupled to the listener's ear using a compressible foam earplug. In the range 150 Hz to 11 kHz, this playback system delivered the same frequency response to the average eardrum as would have occurred at that location during open-ear listening in a diffuse sound field.

Calibration of the playback system was achieved with the output of the insert earphone delivered to a Zwislocki-type ear simulator coupled to a precision sound level meter. The frequency response of the system was monitored daily.

Procedure

Test passages were presented at the level of normal conversational speech in everyday environments reported by Pearsons et al (19). These investigators observed that the level of everyday conversational speech was 55 dB Leq outside the listener's ear. Measurements made by the authors using a Kemar manikin revealed that this corresponded to 61 dB Leq at the average eardrum. Hence, passages were delivered to subjects at a level that produced 61 dB Leq in the Zwislocki-type ear simulator.

The competing babble was presented at S/B ratios of −3, −4, −5, −6, −7, and −8 dB. Each subject heard all 72 passages at a constant S/B ratio. Subjects were randomly assigned to S/B ratio conditions. The use of several S/B ratio conditions allowed more clear definition of the difficulty of potential key words: the more intelligible words were identified at all S/B ratios whereas more difficult words were successively eliminated as S/B ratio worsened.

Subjects were seated in a sound-treated room viewing a video monitor that briefly displayed the passage topic before (but not during) passage presentation. They listened monaurally; the untested ear was occluded with a compressible foam earplug. Delivery and scoring of the CST passages was controlled by an Apple IIe microcomputer. Each passage was presented one sentence at a time. After each sentence, both speech and babble were halted while the subject repeated the sentence or as much of it as he/she had heard. It was emphasized that subjects were to repeat every word exactly as heard.

The examiner sat across from the subject, viewing a second video monitor. The potential key words for the sentence were displayed on this monitor. The examiner scored the words correctly identified by entering the corresponding number on a keypad. Words containing additions, substitutions, or omissions were scored as incorrect.

A practice session, using two passages, was administered before data collection. These two passages were repeated, without feedback, until the subject was familiarized with the task and the talker. Practice passage data were not included in subsequent analyses. Data were collected in two sessions, approximately 1.5 hr each. All experimental variables were counterbalanced or randomized to minimize order effects.

RESULTS

The data consisted of correct/incorrect scores for every potential key word in each CST passage for 30 subjects. Percentage correct scores were derived for these data. These percentages were the basis of passage equalization procedures and are used for descriptive purposes in some portions of the following discussion. However, prior to statistical analyses, all percentage scores were transformed into rationalized arcsine units (rau) as described by Studebaker (20). This had the effect of minimizing the relationship between mean score and variance that is characteristic of percentage scores while at the same time providing a scoring unit that is similar to percentages and is, therefore, readily interpreted. For tests based on 50 or more words, rationalized arcsine units are within 1.3 units of the corresponding percentage value for scores in the range 12 to 88% (for example, 58% = 66.8 rau). As percentage scores increase above 88% or decrease below 12%, the corresponding rau value deviates progressively from the percentage value. Table 1 (adapted from Studebaker's Table 3) shows approximate conversions from percentage to rau within these ranges.

To evaluate the intelligibility of the passages when all potential key words were used for scoring, the overall percentage correct score was computed for each passage. These scores ranged from 36 to 69% (37–68 rau). Obviously, even though the passages had been preequated in difficulty on several criteria, the recorded versions were not all equally intelligible (as anticipated).

Six S/B ratio conditions were used during data collection to more clearly delineate the difficulty of the potential scoring words. It was assumed that changing the S/B ratio would affect the difficulty of the passages about equally, that is, a passage that was relatively intelligible in one S/B ratio would continue to be relatively intelligible in all other S/B ratios. Similarly, a passage that was relatively difficult to understand in one S/B ratio was expected to continue to be relatively difficult under other conditions. To evaluate this assumption, product-moment correlation coefficients were computed between the mean overall score (all S/B ratios) for each passage and the mean score for each passage in the −3 dB S/B ratio condition. The procedure was repeated for the other S/B ratio conditions to give a total of six correlation coefficients. These correlation coefficients ranged from 0.82 to 0.93. All were

Table 1. Approximate conversion from percentage to rationalized arcsine units (rau) for scores based on 50 or more test words.

%	rau	%	rau
100	117	10	8
98	107	8	5
96	102	6	2
94	98	4	−2
92	95	2	−7
90	92	0	−17

statistically significant ($p < 0.001$). This outcome supported the assumption that the relative difficulty of the passages was not changed by the use of different S/B ratio conditions. Consequently, it seemed appropriate to combine data across all S/B ratio conditions to describe the relative intelligibility of the potential key words in each passage.

The objective of determining the intelligibility of each potential key word was to provide a basis for equalizing the passages through the selection of scoring words of equal average intelligibility overall and equal ranges of intelligibility. This was accomplished in the following way: For a given passage, the percent correct score was determined for each potential key word. Words that were correctly repeated more often than 95% or less often than 7% of the time were eliminated as scoring words. The remaining words were divided into five categories based on their intelligibility scores. The categories were: 7.5 to 25%, 25.5 to 42%, 42.5 to 60%, 60.5 to 77%, and 77.5 to 95%. The goal was to select five scoring words in each of these intelligibility categories.

Of the 72 original passages, 57 were found to have at least five words in each intelligibility category; these 57 passages were selected for continued analysis. For each passage, five words in each category were selected as the scoring words. In selecting the scoring words, an attempt was made to include at least one scoring word from each sentence in the passage; to avoid consecutive words in a sentence; and to avoid using any word more than once in a passage (on occasion, it was necessary to violate these guidelines). These 57 passages were then rescored for all subjects, using only the 25 selected words. Many, but not all, passages yielded an overall score of (50 ± 1)%. For passages that did not meet this criterion, scoring words were changed until an overall score of (50 ± 1)% was obtained. This procedure resulted in 57 CST passages that were equated in terms of average intelligibility and range of key word intelligibility across subjects.

In addition to having equal mean scores, test items should be constructed so that item scores are highly correlated with true score on an individual basis. In other words, the intelligibility score obtained by an individual for one CST passage, should be a reasonably accurate estimate of that individual's true score for connected speech material. To explore this issue, each individual's true score for connected speech material was estimated using that person's overall score for the 57 equated passages. For each passage, a linear correlation coefficient was then computed between each individual's true score and his/her score for that passage. This yielded 57 correlation coefficients, ranging from 0.79 to 0.97.

The 48 passages having the best correlations with true scores were selected as the CST test passages. The 9 passages with the lowest correlations with true score were designated practice passages. The 48 selected passages had correlations with true score of at least 0.88. Table 2 gives the topic, mean score, standard deviation, and correlation with true score for the 48 test passages and 9 practice passages that comprise the CST. Data are given in rationalized arcsine units; deviations in mean score from the specified (50 ± 1)% occurred in the transformation from percentage to rau scores.

Table 2. Topic, mean score, standard deviation (SD), and correlation with true score (r) for the 48 test passages and 9 practice passages of the CST

Topic	Mean (rau)	SD (rau)	r	Topic	Mean (rau)	SD (rau)	r
			Test passages				
cabbage	49.3	21	0.92	lake	50.3	25	0.96
calendar	49.9	26	0.93	lawn	49.8	23	0.89
camel	51.4	22	0.93	leopard	50.0	24	0.95
carrot	48.1	27	0.91	lemon	50.4	24	0.88
crow	52.3	25	0.92	lettuce	49.8	22	0.90
clock	49.2	26	0.92	liver	50.2	26	0.94
dice	49.9	24	0.96	lime	49.4	24	0.94
dictionary	51.3	25	0.90	lion	50.9	22	0.93
door	51.5	23	0.91	lizard	50.5	23	0.90
dove	50.2	24	0.91	lung	51.2	23	0.95
eagle	50.0	27	0.94	nail	50.1	25	0.91
ear	50.5	22	0.90	oak	49.5	28	0.92
egg	50.2	26	0.93	orange	50.0	26	0.91
envelope	51.2	26	0.91	owl	50.6	22	0.91
giraffe	50.4	26	0.94	oyster	49.7	22	0.88
glue	51.2	21	0.92	vegetable	51.2	23	0.90
goose	50.0	26	0.91	violin	50.0	18	0.88
gold	49.7	25	0.93	weed	49.2	23	0.95
grape	49.7	21	0.93	wheat	49.7	21	0.91
grass	49.1	21	0.94	window	49.7	25	0.88
grasshopper	51.4	26	0.88	woodpecker	50.5	23	0.88
kangaroo	51.0	22	0.89	wolf	50.8	25	0.92
kite	49.6	28	0.90	zebra	50.2	25	0.94
knife	49.4	21	0.90	zipper	50.7	22	0.97
			Practice passages				
cactus	48.0	22	0.83	glove	51.2	24	0.87
chimney	50.8	23	0.87	ice	49.3	23	0.79
donkey	49.3	20	0.81	lead	48.4	24	0.83
eye	50.7	23	0.85	umbrella	51.8	26	0.83
guitar	49.8	21	0.85				

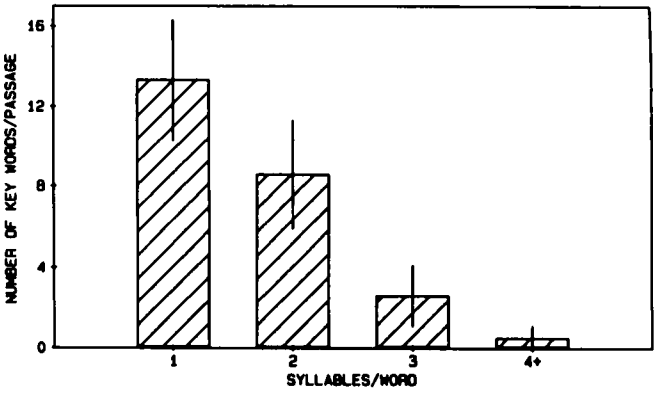

Figure 2. Distribution of syllables/key word in CST test passages. Vertical lines give ±1 SD across the 48 passages.

Figures 2 and 3 describe the key words in the test passages. Figure 2 gives the distribution of syllables/word. In the typical passage, about half the key words are monosyllables and one-third are two-syllable words. Figure 3 shows the phonetic analysis of the key word consonants.

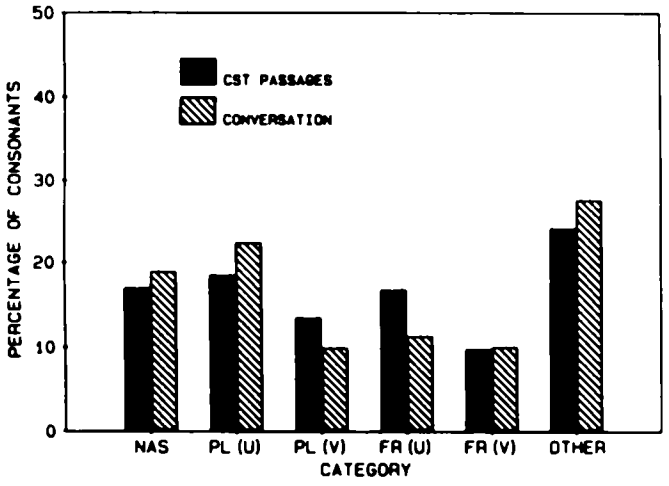

Figure 3. Distribution of key word consonants in various phonetic categories compared with the corresponding distribution of consonants in conversations (21). Nas, nasals; pl(u), voiceless plosives; pl(v), voiced plosives; fr(u), voiceless fricatives; fr(v), voiced fricatives; other, l, r, j, w, h.

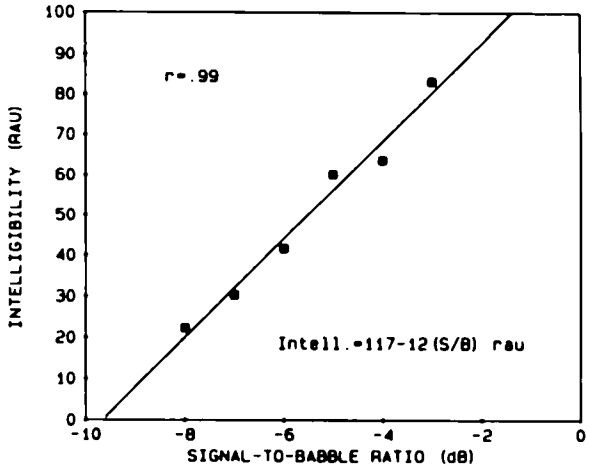

Figure 4. Mean CST score at each S/B ratio and linear regression analysis of relationship between mean CST score and S/B ratio.

As the figure reveals, the mean percentage in each phonetic category is quite similar to the corresponding percentage in conversations reported by Fletcher (21). The standard deviations in each phonetic category across passages were within the range 3.4 to 5.4%.

Data for the 6 S/B ratio conditions were used to determine the slope of the performance-intensity function for the CST. For each S/B ratio condition, the average score across all subjects for the 48 passages was computed. Linear regression analysis was then performed to determine the slope of the line relating mean CST score and S/B ratio. The result is shown in Figure 4. The slope of the function was 12 rau/dB signal-to-babble ratio.

Experiment 2: Validation

To check the difficulty and equivalence of the final CST passages, the 48 test passages and the 9 practice passages were administered to a second group of subjects.

METHOD

Subjects

Ten normal hearers, ranging in age from 12 to 39 yr.

Procedure

The test was administered to all subjects at a nominal S/B ratio of −4 dB (because of an instrumentation problem, the passages were delivered to one subject at −5 dB). This new set of data was collected using the same procedures and instrumentation as described for experiment 1.

RESULTS

Figure 4 gives the regression equation derived from experiment 1 for the prediction of CST score from S/B ratio. The standard error estimate (SE) associated with this equation was 3.9 rau. From this equation, we would predict a mean score in the range 60.0 to 75.6 rau (predicted mean score ± 2 SE) for the normal hearers in experiment 2. In fact, the group mean score for the test passages was 60.1 rau, a value barely within the expected range (individual scores ranged from 48 to 75 rau; the average score for the practice passages was about equal to that for the test passages). This relatively extreme mean score may have occurred by chance or perhaps it suggests that the test is more difficult than anticipated. Although there were two 12-yr-old subjects in this group, there did not appear to be a clear relationship between score and subject age. In addition, the children had no difficulty with the task.

To determine whether there were systematic differences in the difficulty of the test passages, a correlation matrix was derived in which the scores (for the 48 test passages) for each subject were correlated with the corresponding scores for each other subject. It was anticipated that if there were systematic differences in passage difficulty, significant intersubject correlations would be observed. On the other hand, if passage score differences were due mostly to measurement error, significant intersubject correlations would not be expected. Of the 45 linear correlation coefficients derived, one was significant ($r = 0.47$, $p < 0.001$). The remaining 44, nonsignificant, correlation coefficients ranged from −0.32 to 0.30 with a mean value of 0.04. In view of the large number of nonsignificant correlations. It is reasonable to conclude that the single significant correlation occurred by chance. These results confirm that, for normal hearers, there are no systematic differences in difficulty across the 48 test passages.

DETERMINATION OF CRITICAL DIFFERENCES

As noted earlier, it is essential to know the error of estimate that is likely to be associated with scores obtained for a particular individual on any speech intelligibility test. This information can be used to determine critical values for differences between test scores obtained for the same individual under different conditions (such as aided and unaided). In this context, a 95% critical difference (CD) is defined as the difference between two test scores that will

be exceeded by chance alone on only 5% of comparisons. Hence, an observed difference greater than the 95% CD between aided and unaided scores is probably attributable to the effects of the hearing aid rather than due to measurement error.

It has been shown (12) that critical differences for monosyllabic word intelligibility scores can accurately be estimated using a binomial model based on sampling theory. It would be convenient if the same theory could be used to generate critical differences for the CST test. However, because the CST test items (keywords) are embedded in meaningful sentences, they cannot be considered to be independent of each other and, therefore, they do not satisfy the requirements for the binomial model described by Thornton and Raffin (12). With this limitation in mind, the intrasubject variability of scores across the 48 CST passages was compared with the variability that would be predicted from the binomial model. The result is shown in Figure 5. Note that the data displayed in this Figure are percentage values. The *dashed curve* gives the relationship predicted by the binomial model between true score (percent correct) and variability for individual CST passages (i.e., 25 test items per score). The *open circles* give the relationship between the (unbiased) standard deviation of passage scores and true score (estimated as the mean score across the 48 passages) for each subject in experiment 1. The *filled circles* give the analogous data for each subject in experiment 2. The *solid curve* was fitted to the 40 data points using a least squares method. This figure has two noteworthy features: First, the curve fitted to the CST data is very similar to the dashed curve derived from the binomial model. This suggests that the nonindependence of the test items is not a major factor in the variability of the CST passage scores; CST score variability is quite similar to that observed for scores based on independent items. Second, the curve fitted to the CST data deviates progressively more from the dashed curve as true scores improve. This suggests that the nonindependence of the test items has its greatest effect on variability of passage scores when the overall score is high. This outcome is consistent with the observation by numerous investigators that the effect of sentence context on word intelligibility is greater for high scores than for low scores (3, 22). Overall, the data illustrated in Figure 5 indicate that the variability of CST passage scores is slightly greater than would be predicted using a binomial model. However, because the difference is small, critical differences derived for monosyllabic word tests as described by Thornton and Raffin (12) would probably be reasonably accurate for the CST test also. If the test is scored in percentage correct, published values for critical differences could be used to evaluate the significance of differences between scores. If percentage correct scores are transformed to rau values, the following formula, adapted from Studebaker (20), may be used to approximate the 95% CD based on the binomial model:

$$\text{CD (rau)} = 1.96 \sqrt{\frac{173}{n}}$$

where n = number of CST passages used per score. If a 90% confidence interval is desired, 1.65 should be substituted for 1.96 in the equation.

A second, empirical, approach to estimation of critical differences for the CST was based on the measured variability of scores for the 48 passages for each individual. Figure 6 shows the cumulative distribution of the standard deviations (SDs) of the 48 scores for each subject (note that these SDs are for the population of 48 test passages, not unbiased estimates of the SDs for all possible similar passages). As this Figure reveals, the standard deviations of scores for the 48 passages were 11 rau or less for 95% of the subjects. Hence, 11 rau was adopted as a conservative estimate of the standard distribution of passage scores for normal-hearing subjects. Using this value, the 95% CD for CST scores can be computed using the following equation:

$$\text{CD (rau} = \frac{2.7(\text{SD})}{\sqrt{n}} \sqrt{\frac{48 - n}{47}}$$

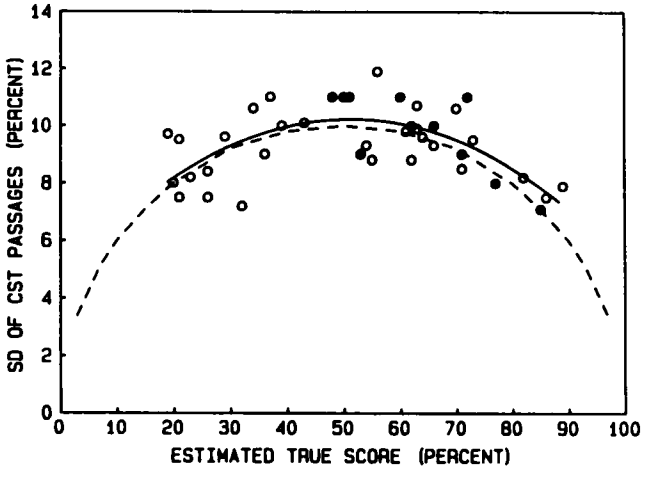

Figure 5. Standard deviations of CST passage scores plotted as a function of overall true scores for: experiment 1 subjects (*open circles*); experiment 2 subjects (*filled circles*); and a binomial model (*dashed curve*). The *solid curve* was fitted to the circles using a least squares method.

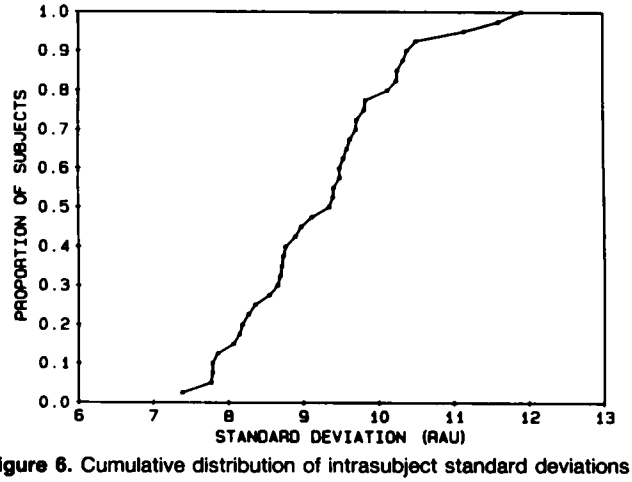

Figure 6. Cumulative distribution of intrasubject standard deviations of scores for 48 CST passages.

where SD = 11 rau and n = number of CST passages used per score. Equation 2 is based on computation of the standard deviation of the distribution of differences between pairs of mean scores for n passages selected randomly from a population of 48 passages (23). A different value of SD can be used in this equation if desired [for example, one may wish to use the average standard deviation (9.3 rau) rather than a value that embraces essentially all subjects]. Also, if calculation of a 90% CD seems more appropriate for a particular application, 2.3 should be substituted for 2.7 in the above equation.

The critical differences produced by equation 1 are only slightly smaller than those produced by equation 2, assuming SD = 11 rau. For clinical applications, equation 1 would be more convenient. For research applications, the more conservative equation 2 might be required.

DISCUSSION

Using equation 1 or 2, it is possible to determine the number of CST passages required per score to meet the goals of this intelligibility test instrument. As noted earlier, it was considered desirable for the error of measurement of the test to be small enough to allow detection of intelligibility differences between conditions equivalent to S/B ratio change of 1 to 2 dB. As shown in Figure 4, the slope of the function relating intelligibility and S/B ratio was 12 rau/dB. On average, an intelligibility change equivalent to 1 to 2 dB S/B ratio would produce a score difference between conditions of 12 to 24 rau. From equation 1, the 95% CD for 2 scores, each based on the mean of four randomly chosen, different, CST passages, is about 13 rau. Equation 2 gives 14 rau. Hence, intelligibility scores based on mean performance for four CST passages should be reliable enough to detect performance differences equivalent to a change of 1 to 2 dB S/B ratio. If six CST passages are used per score, the 95% CD is 10.5 rau by equation 1, 11.5 rau by equation 2, values less than that equivalent to a 1 dB change in S/B ratio.

Also noted earlier, it was considered essential to have a large number of equivalent forms of the CST to allow for testing intelligibility in several conditions without repetition of passages. Because the 48 passages are essentially equally intelligible for normal hearers, it is not necessary to designate particular combinations of passages to comprise equivalent forms. If four randomly chosen passages are used per score, this will result in 12 equivalent CST forms for the average normal hearer.

The slope of the performance-intensity function, 12 rau dB, (essentially equal to 12%/dB for scores in the range 20–80%) is considerably greater than the 4 to 6%/dB usually reported for monosyllabic word intelligibility for normal hearers (24) but less than the 16 to 20%/dB noted by Cox and McDaniel [unpublished data associated with (15)] for subjectively assessed intelligibility of connected speech embedded in competing multitalker babble. A steep performance-intensity function has both positive and negative aspects. On the one hand, the test is sensitive to small changes in S/B ratio so that a small change in conditions produces a large change in scores. On the other hand, the range of S/B ratio values within which the test is useful is narrow—about 9 dB. As Figure 4 shows, a change in S/B ratio from 0 to −9 dB will typically result in a performance change from full intelligibility to essentially zero intelligibility. The implication is that S/B ratio must be carefully chosen for the CST, probably on an individual basis.

As described above, listeners were apprised of the passage topic before the passage was presented. This procedure was chosen in an attempt to maintain the content validity of the test: in everyday speech, the topic of conversation is frequently known. Experience with the CST with normal hearers and a limited number of hearing-impaired individuals suggests, not surprisingly, that information on passage topic has a significant effect on intelligibility of the passage. Hence, presentation of passage topic must be considered an integral part of the CST.

The time required to administer a test is an important consideration in clinical applications. Although the CST has been developed primarily for use as a criterion measure in investigations of hearing aid benefit, it could be used in clinical settings. Administration of four CST passages consumes less than 10 minutes. One obvious application for this test would be to estimate the everyday speech intelligibility that might be expected with a newly fitted hearing aid. Ideally, the CST would be administered after a real ear technique has been used to ascertain the presence of the desired frequency-gain characteristic.

Presentation of the test using a microcomputer-controlled optical disk player is not essential. However, this method has significant advantages because scoring and random selection of passages can be handled by the computer and the optical disk format allows essentially instantaneous random access to any sentence of any passage.

Future work with the CST will include investigation of the contribution of the visual component of the CST passages. It is anticipated that the CST will be useful in measurements of hearing aid benefit in audiovisual communication. Evaluation of the CST with hearing-impaired persons is in progress. Although hearing-impaired listeners may not be inherently less reliable than normal hearers [see Dillon (25) for a discussion of this issue], it is probable that the intelligibility of passages equalized for normal hearers will not be equal for hearing-impaired listeners. It is anticipated that it may be appropriate to designate specific combinations of passages to assemble equivalent CST forms for hearing-impaired persons. Furthermore, the particular passages comprising equivalent forms may be different for persons having different audiogram configurations. This work will be reported in a future paper.

References

1. Barcham LJ, Stephens SDG. The use of an open-ended problems questionnaire in auditory rehabilitation. Br J Audiol 1980;14:49–51.
2. Hagerman B, Gabrielsson A. Questionnaires on desirable properties of hearing aids. Karolinska Institute 1984; Report TA109.
3. O'Neill JJ. Recognition of intelligibility test materials in context and isolation. J Speech Hear Disord 1957;22:87–90.
4. Giolas TG, Epstein A. Comparative intelligibility of word lists and continuous discourse. J Speech Hear Disord 1963;6:349–58.
5. Schiavetti N, Sitler RW, Metz DE, Houde RA. Prediction of contextual speech intelligibility from isolated word intelligibility measures. J Speech Hear Res

1984;27:623–6.
6. Walden BE, Schwartz DM, Williams DL, Holum-Hardegan LL, Crowley JM. Test of the assumptions underlying comparative hearing aid evaluations. J Speech Hear Disord 1983;48:264–73.
7. Scherr CK, Schwartz DM, Montgomery AA. Follow-up survey of new hearing aid users. J Acad Rehabil Audiol 1983;202–9.
8. Silverman SR, Hirsh IJ. Problems related to the use of speech in clinical audiometry. Ann Otol Rhinol Laryngol 1955;64:1234–44.
9. Anonymous. IEEE recommended practice for speech quality measurements. IEEE Trans. Audio and Electroacoustics 1969; AU-17:225–47
10. Kalikow DN, Stevens KN, Elliot LL. Development of a test of speech intelligibility in noise using sentence materials with controlled word predictability. J Acoust Soc Am 1977;61:1337–51.
11. Bilger RC, Neutzel JM, Rabinowitz WM, Rzeczkowski C. Standardization of a test of speech perception in noise. J Speech Hear Res 1984;27:32–48.
12. Thornton AR, Raffin MJM. Speech-discrimination scores modeled as a binomial variable. J Speech Hear Res 1978;21:507–18.
13. Punch JL. Quality judgements of hearing aid processed speech and music by normal and otopathologic listeners. J Am Aud Soc 1978;3:179–88.
14. Witter HL, Goldstein DP. Quality judgements of hearing aid transduced speech. J Speech Hear Disord 1971;14:312–22.
15. Cox RM, McDaniel DM. Intelligibility ratings of continuous discourse: application to hearing aid selection. J Acoust Soc Am 1984;76:758–66.
16. Cox RM, Alexander GC, Gilmore C. Intelligibility of average talkers in typical listening environments. J Acoust Soc Am 1987;81:1598–1608.
17. Goldman-Eisler F. Experiments in Spontaneous Speech. New York: Academic Press, 1968:25.
18. Duffy JR, Giolas TG. Sentence intelligibility as a function of key word selection. J Speech Hear Res 1974;17:631–7.
19. Pearsons KS, Bennett RL, Fidell S. Speech levels in various noise environments. United States Environmental Protection Agency, 1977; Report EPA 600/1-77-025.
20. Studebaker GA. A "rationalized" arcsine transform. J Speech Hear Res 1985;28:255–62.
21. Fletcher H. Speech and Hearing in Communication. New York: Van Nostrand Reinhold Company, 1953:96.
22. Sitler RW, Schiavetti N, Metz DE. Contextual effects in the measurement of hearing-impaired speakers' intelligibility. J Speech Hear Res 1983;26:22–30.
23. Ferguson GA. Statistical Analysis in Psychology and Education. New York: McGraw-Hill Book Company, 1969:140, 147.
24. Olsen WO, Matkin ND. Speech Audiometry. In: Rintlemann WF, ed. Hearing Assessment. Baltimore: University Park Press, 1979:175.
25. Dillon H. A quantitative examination of the sources of speech discrimination test score variability. Ear Hear 1982;3:51–8.

Acknowledgments: Robert Joyce wrote the software to administer and score the CST. The administrative assistance of Kay M. Pusakulich is gratefully acknowledged.

Address reprint requests to Robyn M. Cox, Ph.D., Memphis Speech & Hearing Center, 807 Jefferson Ave., Memphis, TN 38105.