Development of a quick speech-in-noise test for measuring signal-to-noise ratio loss in normal-hearing and hearing-impaired listeners[a)]

Mead C. Killion, Patricia A. Niquette, and Gail I. Gudmundsen
Etymotic Research, Inc., 61 Martin Lane, Elk Grove Village, Illinois 60007

Lawrence J. Revit
Revitronix, Brownsville, Vermont 05037

Shilpi Banerjee[b)]
Northwestern University, Evanston, Illinois 60208

(Received 30 September 2003; revised 29 June 2004; accepted 30 June 2004)

This paper describes a shortened and improved version of the Speech in Noise (SIN™) Test (Etymotic Research, 1993). In the first two of four experiments, the level of a female talker relative to that of four-talker babble was adjusted sentence by sentence to produce 50% correct scores for normal-hearing subjects. In the second two experiments, those sentences-in-babble that produced either lack of equivalence or high across-subject variability in scores were discarded. These experiments produced 12 equivalent lists, each containing six sentences, with one sentence at each adjusted signal-to-noise ratio of 25, 20, 15, 10, 5, and 0 dB. Six additional lists were also made equivalent when the scores of particular pairs were averaged. The final lists comprise the "QuickSIN" test that measures the SNR a listener requires to understand 50% of key words in sentences in a background of babble. The standard deviation of single-list scores is 1.4 dB SNR for hearing-impaired subjects, based on test-retest data. A single QuickSIN list takes approximately one minute to administer and provides an estimate of SNR loss accurate to ±2.7 dB at the 95% confidence level. © *2004 Acoustical Society of America.* [DOI: 10.1121/1.1784440]

PACS numbers: 43.71.Ky, 43.71.Gv, 43.72.Dv [KWG] Pages: 2395–2405

I. INTRODUCTION

Hearing aid wearers report that their biggest problem with their hearing aids is that of understanding speech in background noise, and consumer surveys polling approximately 80 000 households have consistently revealed consumer dissatisfaction with hearing aids in noisy environments (Kochkin, 1992, 1993, 1995, 1996, 2000, 2002). Kochkin (2002) reported that only 30% of hearing aid wearers were satisfied with their hearing aids in noisy situations.

As summarized below, recent evidence suggests that the wide range of satisfaction with hearing aids in noise reflects a wide range in the ability of hearing aid wearers to understand speech in a background of noise. By analogy with hearing loss, "SNR loss" (signal-to-noise ratio loss) refers to the increase in signal-to-noise ratio required by a listener to obtain 50% correct words, sentences, or words in sentences, compared to normal performance.[1]

Published reports indicate a wide range of SNR loss in persons with similar pure tone hearing losses (Lyregaard, 1982; Dirks *et al.*, 1982; Killion, 1997; Killion and Niquette, 2000; Taylor, 2003). The standard audiometric test battery does not measure or predict the ability to understand speech in noise (Killion and Niquette, 2000). Figure 1 shows data on 100 hearing-impaired listeners. Those with a 40–60 dB pure-tone-average (PTA) loss, for example, have SNR losses ranging from less than 2 dB (no more trouble hearing in noise than normal-hearing listeners) to greater than 20 dB (severe loss of ability to hear in noise). Without knowledge of a listener's SNR loss, it is virtually impossible to give realistic expectations for their potential improvement in noise with hearing aids. One person with a 50 dB PTA loss but without SNR loss may report little or no difficulty hearing in noise with hearing aids, while another with the same PTA loss, but a severe SNR loss, may require a remote FM microphone in order to understand speech in noise. Just as important, knowing the SNR loss makes it possible for the hearing professional to recommend the appropriate technology (e.g., omnidirectional microphones, directional microphones, array microphones, close-talking FM microphones) required for the listener to function in commonly encountered noisy situations.

SNR testing is only recently becoming common in clinical practice (Strom, 2003). Nilsson *et al.* (1994) described the hearing in noise test (HINT) that uses sentences in continuous speech-spectrum shaped noise and an adaptive procedure that gives the SNR for 50% correct for whole sentences. The use of whole-sentence scoring on the HINT has the advantage that whole sentences are tested, and the disadvantage that a greater number of sentences is required for a given statistical reliability than when key-word scoring is used. Similarly, the use of continuous noise has the advantage of reducing the variability in noise level, and the disad-

[a)]Portions of this work were published in the QuickSIN™ test manual, Etymotic Research, Inc., © 2001.
[b)]Present address: Starkey Laboratories, Eden Prairie, MN 55344.

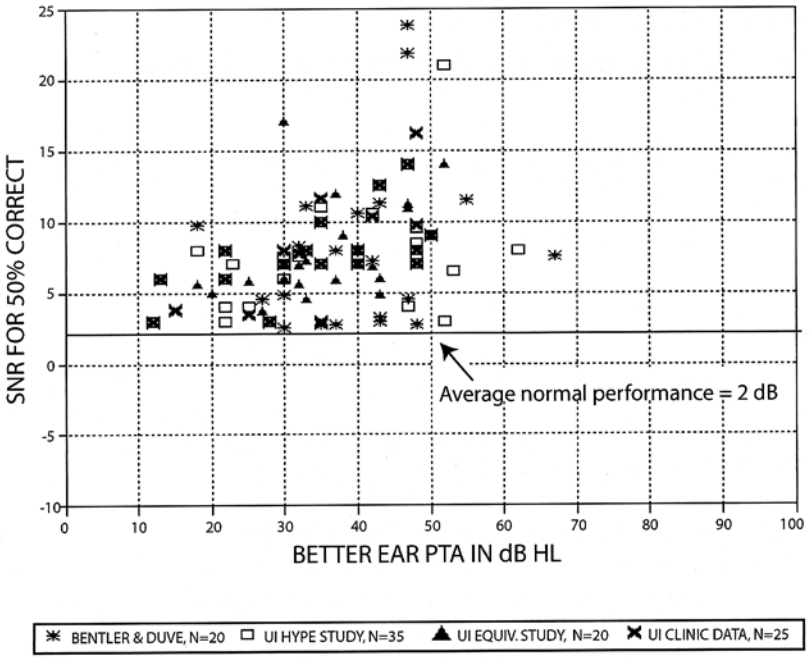

FIG. 1. Signal-to-noise ratio for 50% correct on the SIN test (70 dB HL presentation level) versus three-frequency average pure-tone hearing loss in the better ear (average of 0.5, 1, and 2 kHz). Four data sets obtained at the University of Iowa Speech and Hearing Clinic. From Killion and Niquette (2000), with permission.

vantage that it is less representative of everyday speech-in-noise situations than babble noise.

The purpose of the present experiments was to develop a speech-in-noise test which would (1) estimate SNR loss in one to two minutes, (2) be easy to administer, (3) have good face validity, (4) have simplified scoring, and (5) provide list equivalency for both normal and hearing-impaired subjects.

When designing a speech-in-noise task, the choice of speech and background noise materials is a compromise between realism and reproducibility. Monosyllabic words, recorded and played back at uniform intensity levels, are not representative of speech in the real world. Sentences spoken with natural dynamics have greater dynamic range than monosyllabic words, and are thus a more valid representation of real speech (Villchur, 1982). Furthermore, the effects of coarticulation are not well represented on monosyllabic word lists.

Likewise, a constant-level background noise, while easy to control and reproduce, is not typical of that encountered by most people in their everyday environments. Fikret-Pasa (1993) examined the intensity variations as a function of time of the background noise encountered in everyday situations (e.g., shopping malls and crowded restaurants) and found level variations having standard deviations of 2.8 to 8.4 dB, for maximum and minimum sound level meter readings, respectively. In contrast, Fikret-Pasa measured virtually no variation in level in available speech-spectrum-noise maskers, and only a 1-dB variation in level in two examples of multi-talker babble, both of which contained so many talkers that the result was a constant-level murmur. She found that the Auditec four-talker babble (Auditec of St. Louis, 1971) had more level variations than any of the other commercially available noises, presumably because the babble talkers were instructed to speak naturally (Carver, 1991). Use of a background noise with level variations is particularly important for a test used with compression hearing aids, so that the compression circuits are not clamped in a fixed-gain setting by the noise.

Fikret-Pasa (1993) also chose four-talker babble because it represents a realistic simulation of a social gathering, in which the listener may tune out the target talker and tune in one or more of the other nearby talkers using what Broadbent (1958) labeled "selective listening." More subtly, the use of constant-level noise in speech-reception research eliminates the temporary gaps in the noise of real talkers, gaps which those with normal hearing appear able to make use of when listening in noise (Bacon *et al.*, 1998).

The SIN test (Killion and Villchur, 1993; Etymotic Research, 1993) was developed for estimating the signal-to-noise ratio required by hearing impaired subjects to obtain adequate intelligibility under different hearing aid processing conditions. In accordance with standard threshold procedures, the signal-to-noise ratio required for 50% correct words in sentences is abbreviated as "SNR-50."

The SIN test combined a Massachusetts Institute of Technology recording (female talker) of the IEEE sentences [Institute of Electrical and Electronics Engineers (IEEE), 1969] as signal, and a four-talker babble (Auditec of St. Louis, 1971) as competing noise. Each IEEE sentence has five key words, which are scored as correct or incorrect. Rabinowitz *et al.* (1992) reported experiments with cochlear implant subjects indicated that the IEEE sentences are comprised of words that are not highly predictable from the surrounding context, resulting in a performance-intensity function closer to that obtained with word scoring than with whole-sentence scoring.

The ten-sentence IEEE lists were reportedly phonetically balanced (IEEE, 1969). However, no attempt was made to maintain these list groupings on the SIN test. Egan (1948) observed that the distribution of sounds in speech depends on the topic being discussed and who is speaking. The phonetic balancing of a list of words or words-in-sentences is influ-

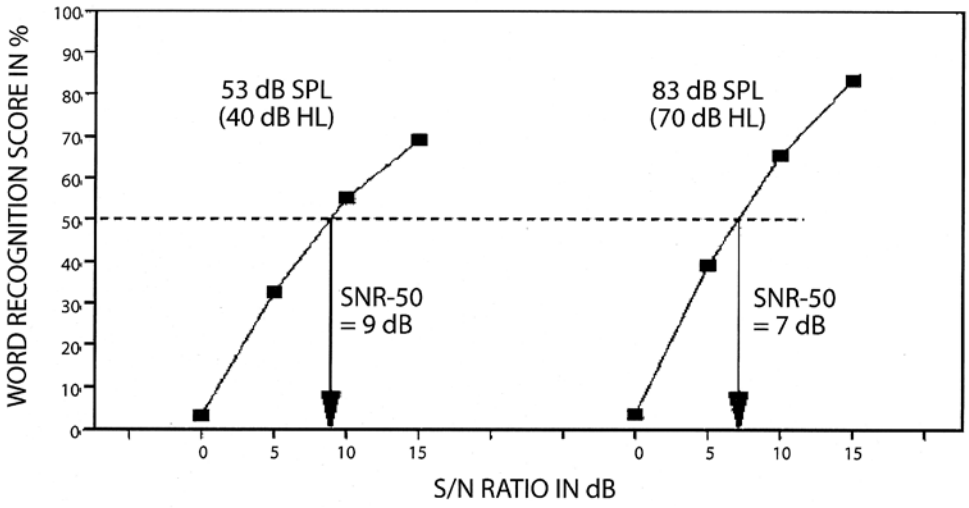

FIG. 2. Illustration of the graphical scoring method used in the SIN test. Each SNR level contains 25 words, so the number correct is multiplied by 4 to arrive at the percent-correct score. The percent correct is then plotted for each SNR. At the point where the 50% line is crossed, a vertical line is drawn down. The SNR-50 is then interpolated by the value given on the *X* axis.

enced by the talker; when the same words or sentences are recorded using a different talker, the original phonetic balancing is not maintained. Hood and Poole (1980) demonstrated this conclusively when they selected "difficult" and "easy" words, and found that "words that were formerly difficult became easy and vice versa" when recorded by a second speaker. Recently, Martin *et al.* (2000) found that words selected randomly from a dictionary could not be statistically distinguished from the phonetically balanced NU-6 words (Tillman and Carhart, 1966). Since the IEEE sentences had no standard talker, it seemed likely that two sets of phonetically balanced sentences might produce significantly different average intelligibility when spoken by different talkers, which is what was found in early SIN test experiments.

The SIN test uses the first 360 of the 720 IEEE sentences (lists 1–36), divided into nine blocks of 40 sentences each. Each SIN test block contains two sections, the first administered at a 70 dB HL presentation level, and the second at 40 dB HL. These levels were chosen to represent the range of typically loud and quiet speech levels encountered by most people in everyday life. Each section contains 20 sentences, five sentences at each signal-to-noise ratio of 15, 10, 5, and 0 dB. Each sentence contains five key words, which are scored as correct or incorrect, resulting in 25 key words at each SNR. The key words repeated correctly are summed for each SNR and multiplied by 4 to obtain a percent-correct score. The percentage scores are manually plotted for each SNR, using a graph with SNR on the abscissa and percent correct on the ordinate. The SNR-50 is interpolated by drawing a horizontal line at the 50% point, and dropping a vertical line at the SNR that intersects the 50% line. As shown in Fig. 2, scores are plotted separately for the 70 and 40 dB HL presentations.

Test time for a single SIN test block (using both of the recommended presentation levels of 70 and 40 dB HL) is approximately 6 min if the subject responds promptly. While the SIN test offered good face validity, some practitioners reported that it was too time-consuming for clinical use and that the graphical scoring for the test was difficult. *Posthoc* analysis of the SIN test (Killion *et al.*, 1996; Bentler, 2000) revealed that not all of the test blocks were equivalent, resulting in too few lists for some clinical comparisons and research purposes. Both floor and ceiling effects were also noted (Bentler, 2000).

Killion *et al.* (1996) reported that on blocks 3, 4, 5, 6, and 8 of the SIN test, normal-hearing listeners, as a group, produced similar SNR-50 estimates across blocks, and hearing-impaired listeners, as a group, also produced similar SNR-50 estimates across blocks. They reported that the standard deviation of a *single* block test score is 0.8 dB, based on test-retest data. Thus the critical difference for a two-way comparison at the 95% confidence level is 2.2 dB, using one half-block at a single presentation level (e.g., 70 dB HL) for each condition, and 1.6 dB using two half-blocks for each condition. To explain, for the average of two lists, the predicted 1.6 dB critical difference at the 95% confidence level is equal to 1.96*(sqrt2*0.8)/sqrt2, where the second sqrt2 refers to the use of two lists.

Bentler (2000) reported that SIN test blocks 1, 2, and 9 were equivalent to each other and that blocks 3, 4, and 5 were equivalent to each other. Test-retest correlations were high, with a 95% confidence-level critical difference between two single-block scores (single levels) of 2.4 dB (equivalent to a single-block test score standard deviation of 0.87 dB). Cox *et al.* (2001) developed a revised version of the SIN test, the RSIN, which reallocated the recorded SIN test materials into four "modified dual blocks" based on the data from Bentler (2000). Cox *et al.* reported a critical difference of 1.7 dB SNR for these RSIN dual blocks. The standard deviation inferred from their report for a *single-block* test score is thus also 0.87 dB.

From these three reports, it would appear that a conservative value for the standard deviation of one single-presentation-level SIN test block would be 0.85 dB. This value was used in the design of the four experiments conducted during the development of the QuickSIN test. Since the SIN test employed five sentences at each SNR and the QuickSIN was to employ only one, the predicted standard deviation for a single QuickSIN test score was 0.85*sqrt5 or 1.9 dB.

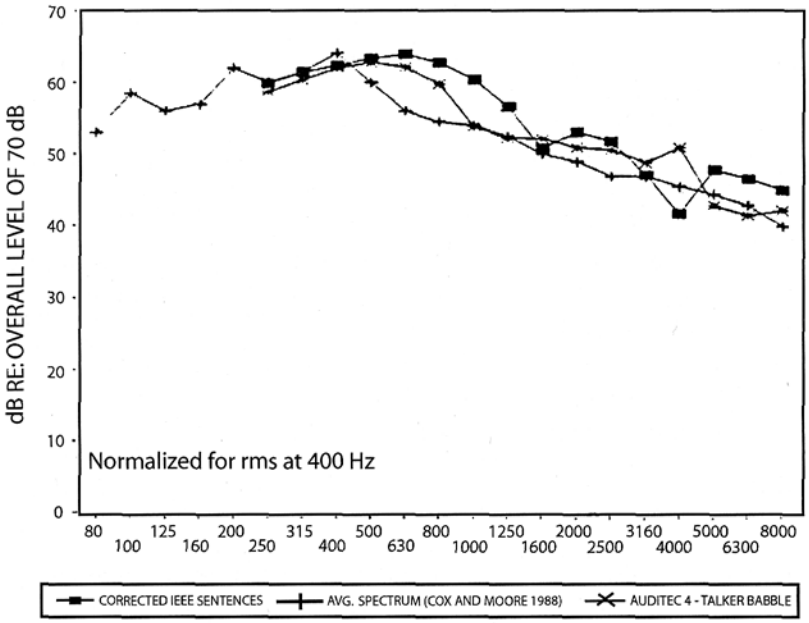

FIG. 3. Long-term average one-third-octave spectrum of IEEE sentences after filtering for microphone placement artifact, compared to measured long-term average, one-third-octave spectrum of Auditec four-talker babble and Cox and Moore (1988) average spectrum for continuous male speech. From Fikret-Pasa (1993), with permission.

II. EXPERIMENT 1

The purpose of experiment 1 was to normalize the SNR of the sentence-babble pairs, so that after normalization each pair had the same SNR-50 for normal-hearing subjects.

A. Method

1. Subjects

Sixteen adult subjects with normal hearing participated in experiment 1.[2] Normal hearing was defined as pure tone thresholds equal to or better than 20 dB HL for octave frequencies 250 to 4000 Hz. Subject ages ranged from 18 to 51 years, with an average age of 24 years and a median age of 20 years.

2. Stimuli

The same equalized MIT recording of a female talker and the same four-talker babble described above were used here, except the sentences were the second group of 360 IEEE sentences, lists 37–72 (recall that the first group of 360 sentences was used in the SIN test). A spectral analysis of the equalized female talker, the four-talker babble, and Cox and Moore's (1988) average spectrum for continuous male speech is shown in Fig. 3.

An "Alpha 1" master digital recording of the sentences was made, with the MIT female talker on channel 1 and the four-talker Auditec babble on channel 2. The instantaneous levels of both the MIT female talker and the babble talkers ebb and flow at any presentation level; as a result, the SNR required for 50% correct in a given sentence depends on the babble with which that sentence is paired. In the two-channel master recording of the sentences and babble, each pair was time locked, meaning that the time relationship between each sentence and its corresponding babble segment was fixed. The master recording was made so that all subsequent rerecordings of a given sentence had the same time-locked relationship between speaker and babble segments.

In previous studies using the SIN test, Killion *et al.* (1996) and Bentler (2000) found that normal-hearing subjects score 50% correct at approximately a 2-dB SNR. In experiment 1 the prerecorded calibration tones found on the MIT sentence recordings and on the Auditec babble recordings were used to set a nominal 2-dB signal-to-noise ratio for each sentence-babble combination.[3] Additional recordings of all 360 sentences were made at −1 dB SNR and +5 dB SNR in the same manner. The result was three Alpha 1 CDs, one each at nominal average signal-to-noise ratio of −1, +2, and +5 dB, which covered the range of likely SNR-50 values for each sentence-babble pair.

3. Stimulus presentation

In all experiments, speech materials were routed through the speech channel of standard clinical audiometers, and testing took place in sound-treated test booths. Each CD contained a calibration track, and VU meters were set for "0" using the calibration tone prior to testing.

The Alpha 1 sentences were presented at 70 dB HL via ER-3A insert earphones. Twelve subjects were tested monaurally, and four subjects were tested binaurally. Prior to the test session, three prototype lists were administered at 70 dB HL to familiarize the subjects with the task. Each prototype list consisted of six IEEE sentences (IEEE, 1969), female talker, in four-talker babble (Auditec of St. Louis, 1971) with one sentence at each SNR of 25, 20, 15, 10, 5, and 0 dB. The 360 Alpha 1 sentences were then presented at −1, +2, and +5 dB SNR, in that order.

4. Scoring

One point was given for each of five key words repeated correctly in each sentence. Half credit was given for words close to the target word, e.g., "cat" for "cats." The SNR-50 was calculated for each sentence using a formula based on the Tillman–Olsen (1973) recommended method for obtaining spondee thresholds.[4] These scoring methods were used in all four experiments.

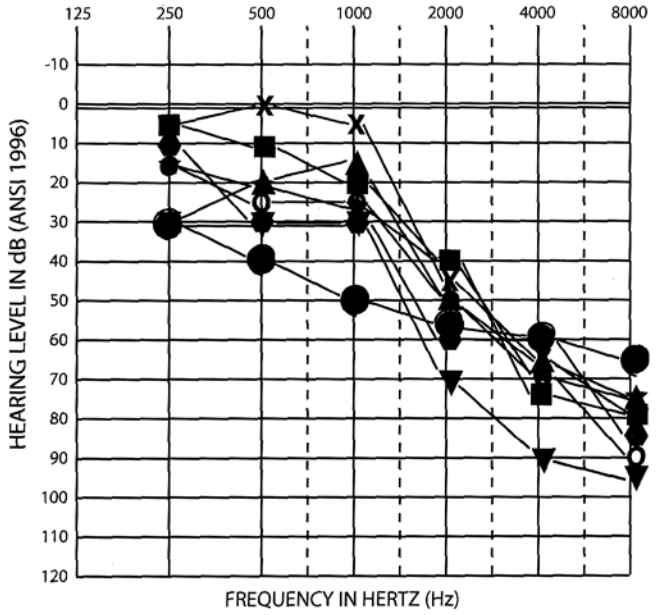

FIG. 4. Composite audiogram for eight hearing-impaired adult subjects from experiment 2.

B. Results and discussion

Experiment 1 provided, for each sentence-babble combination, SNR-50 values for each subject. The across-subject average SNR-50 for the 360 Alpha1 IEEE sentences was +2.5 dB, and the standard error of the mean for the 16 subjects was 0.3 dB. The experimentally derived SNR-50 value for each of the time-locked sentence-babble combinations ranged from −1.3 to +5.9 dB across the 360 sentences. The subject-data-based SNR-50 values thus often differed substantially from the nominal values based on the prerecorded calibration tones. In addition to the ebb and flow in the talker and babble, some sentences were simply easier than others. The individual-sentence subject-data-based SNR-50 values formed the basis for the recordings in experiment 2.

III. EXPERIMENT 2

The purpose of experiment 2 was to assess the equivalence of the 360 time-locked sentence-babble combinations after they had been readjusted for SNR equivalence based on the data from experiment 1.

A. Method

1. Subjects

Six normal-hearing adult subjects and eight hearing-impaired adult subjects with symmetrical sloping, severe high-frequency sensorineural hearing loss participated in experiment 2. Normal hearing was defined as pure tone thresholds equal to or better than 20 dB HL for octave frequencies 250 to 8000 Hz. The normal-hearing subjects ranged in age from 20 to 23 years, with an average age of 22 years and a median age of 22 years. Figure 4 shows the audiograms of the eight hearing-impaired subjects; thresholds represent the test ear for the seven subjects who were tested monaurally, and the better ear for the subject who was tested binaurally. The hearing-impaired subjects ranged in age from 60 to 78 years, with an average age of 69 years and a median age of 66 years.

2. Stimuli

A new set of recordings was made based on the SNR-50 values from experiment 1. For these recordings, the values from experiment 1 were used to readjust the recorded SNR of each sentence-babble combination to bring it to an expected value of 2 dB [the average SNR-50 for normal-hearing adults at a 70 dB HL presentation level on the SIN test (Bentler, 2000)]. For example, sentence 1 on list 37 had an SNR-50 of 3.5 dB in experiment 1, so the level of the babble associated with this sentence was reduced by 1.5 dB to produce an expected SNR-50 of 2 dB. A new master recording was made using the readjusted babble levels, preserving the previous time-locked relationship between sentences and babble. From this master, a set of "Alpha 2" CD recordings was made of all sentences, with the sentence-to-babble ratio readjusted to nominal SNR values of 0, +5, +10, and +15 dB.

3. Stimulus presentation

The Alpha 2 recordings were presented monaurally at 70 dB HL as described in experiment 1 for subjects with normal hearing, and at a presentation level judged to be "Loud, but O.K." (just below discomfort) for subjects with hearing loss (Valente and Van Vliet, 1997). These presentation levels were chosen to ensure that nearly all speech cues were audible. Frequency response shaping was not used for hearing-impaired subjects, since previous research (Skinner, 1976; Dirks, 1982; van Buuren *et al.*, 1995) showed negligible differences in SNR at these high presentation levels as the frequency response was changed. Monaural presentation was used for seven of the eight hearing-impaired subjects.

The normal-hearing subjects listened to the lists at the 0 and +5 dB SNRs, since they were expected to obtain nearly 100% correct at +10 dB SNR and above. To determine the appropriate SNRs to use for testing the hearing-impaired subjects, two half-blocks of the SIN test were administered to each subject. The resulting SNR-50 value was used to determine which two Alpha 2 CDs to use. For example, the +5 and +10 dB SNR recordings were used for a subject with an SNR-50 of 7 dB on the SIN test. In all cases, the lists with the lower (most difficult) SNR were administered first.

B. Results and discussion

The across-subject average SNR-50 for the six normal-hearing subjects was 2.4 dB, and the across-sentence average of the across-subject standard deviations was 1.6 dB, giving a standard error of the mean for the six normal-hearing subjects of 0.65 dB. The six-subject average value of 2.4 dB was not significantly different than the expected value of 2 dB (for this small sample, the ±95% confidence interval around the average value of 2.4 dB was 1.1 to 3.7 dB). The across-subject standard deviation was small, indicating good equivalence for the readjusted sentence-babble combinations.

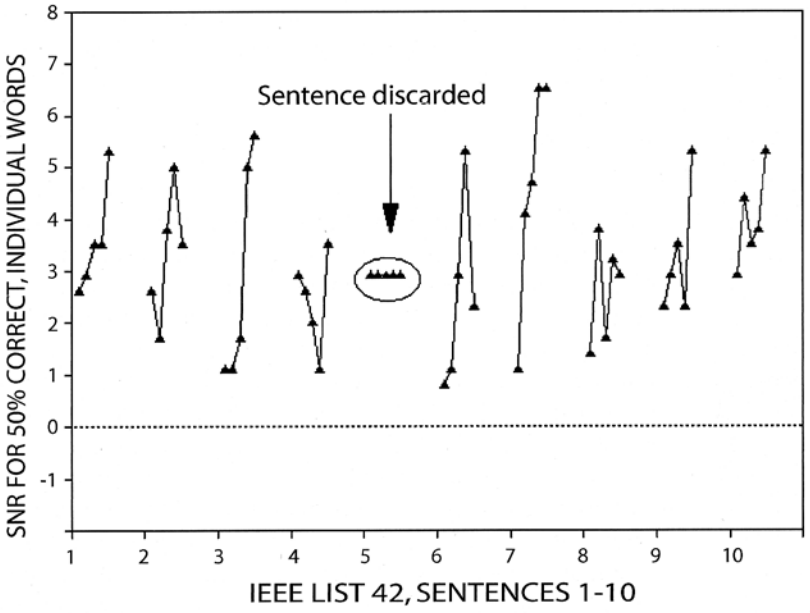

FIG. 5. SNR-50 data for individual words in IEEE list 42, sentences 1–10, derived from an analysis of data from five normal-hearing subjects from experiment 1. Sentences in which the SNR-50 for each of the five key words varied by less than or equal to 2 dB were discarded. Sentence 5 illustrated an unusual case in which all words had identical SNR-50 values.

The across-subject SNR-50 for the eight sloping-hearing-loss subjects was 7.4 dB, and the across-sentence average of the across-subject standard deviations was 2.9 dB. The higher across-subject standard deviation for the hearing impaired subjects resulted from the wide range of apparent SNR losses they exhibited. These data were used in the sentence selection process of experiment 3.

IV. EXPERIMENT 3

The purpose of experiment 3 was to assess the equivalence of selected time-locked sentence-babble combinations, using subjects with normal hearing and subjects with hearing loss. The selection of these sentences was made based on data from experiments 1 and 2.

A. Method

1. Subjects

a) Group 1. Twenty-six normal hearing adult subjects, 15 subjects from 12 Beta sites and 11 subjects from the University of Iowa Speech and Hearing Clinic, participated in experiment 3. Normal hearing was defined as pure tone thresholds equal to or better than 20 dB HL for octave frequencies 250 to 4000 Hz. All subjects met this criterion except two; one subject had a 25 dB HL threshold at 250 Hz in one ear, and another subject had a threshold of 30 dB HL at 4000 Hz in one ear. Subjects ranged in age from 20 to 58 years, with an average age of 32 years and a median age of 27 years.

b) Group 2. Eighteen hearing-impaired adult subjects from ten Beta sites participated in experiment 3. Hearing impaired subjects had symmetrical sensorineural hearing losses. Audiometric criteria for the hearing-impaired subjects were (*i*) mild to moderate sloping loss, with a minimum 20-dB drop from 500 to 4000 Hz; (*ii*) mild to moderate flat loss, with less than a 15-dB drop from 500 to 4000 Hz; or (*iii*) severe loss, with thresholds from 60 to 90 dB for frequencies 500 to 4000 Hz. Subjects ranged in age from 20 to 77 years, with an average age of 57 years and a median age of 58 years.

2. Stimuli

Data from experiments 1 and 2 were used to select sentences that were equivalent for normal-hearing and hearing-impaired subjects. Selection criteria were the following.

(i) Better than average across-subject variability in sentence scores: Only sentences with an across-subject standard deviation of less than 1.5 dB were retained, based on experiment 2 data from six normal-hearing subjects. Since the criterion was 0.1 dB below the *average* of 1.6 dB, slightly more than half of the sentences were discarded as having greater than average across-subject standard deviation.[5]

(ii) The across-subject average SNR-50 value for the normal-hearing subjects was within 1.5 dB of the grand average value for each retained sentence, based on experiment 2 data.

(iii) The across-subject average SNR-50 value for the hearing-impaired subjects fell within 2 dB of the grand average value for each selected sentence, based on experiment 2 data from eight subjects with sloping, high-frequency sensorineural hearing loss.

(iv) The SNR-50 value was calculated for each of the 1800 key words found in the 360 sentences, using data from experiment 1 for five randomly selected subjects from the 16 normal-hearing subject data pool. Data from five subjects were considered adequate, since it resulted in 15 data points for each word, or 27 000 total data points. For a sentence to be retained, the range of SNR-50 values across the five key words in the sentence had to be greater than 2 dB (see Fig. 5).

(v) Sentences containing language that is no longer contemporary were eliminated. For example, "The smell

of burned rags itches my nose" and "The vamp of the shoe had an old buckle" were eliminated.

The purpose of the above steps was to select time-locked sentence-babble combinations that had good reliability across subjects and sentences. Some variation in SNR of words within sentences was desirable so that the sentences were more representative of everyday speech than sentences with nearly the same SNR for each word. These combined procedures eliminated 75% of the original 360 sentences, leaving 89 sentences meeting the criteria.

Beta recordings were made using 84 of the 89 selected sentence-babble combinations. This yielded 14 lists of six sentences each, with one sentence at each of the following signal-to-noise ratios: 25, 20, 15, 10, 5, and 0 dB. These were labeled Beta lists 1–14. Since additional lists were desired, the standard deviation limit in step *ii* (above) was increased from 1.5 to 2.0 dB. This change yielded enough additional sentences for seven more lists of six sentences each, labeled Beta lists 15–21.

3. Stimulus presentation

Beta lists were presented binaurally at 70 dB HL to normal-hearing subjects. Most subjects (19 of 26) were tested with ER-3A insert earphones; four subjects were tested in sound field, and three subjects with TDH headphones. At a presentation level of 70 dB HL, the quietest speech cues should have been at 40 dB HL. Even taking into account the field-referenced high-frequency rolloff of the TDH-39 and ER-3A earphones, no speech cues should have fallen below 30 dB HL. van Buuren *et al.* (1995) found no difference in intelligibility in noise for 18 of 25 frequency responses, covering most of the dynamic range of their hearing-impaired listeners, providing further evidence suggesting that the effect of the earphones should have been minimal. Hearing-impaired subjects were tested at 70 dB HL except when their three-frequency pure tone average (PTA) exceeded 50 dB HL, in which case they were tested at a level they judged to be "Loud, but OK." The Beta list presentations were counterbalanced to control for potential order effects.

B. Results and discussion

The across-subject, across-list SNR-50 average for the 26 normal-hearing subjects in group 1 was 1.9 dB, nearly identical to the original SIN test average of 2 dB. As in experiment 2, the average fell well within the 95% confidence interval of the expected value: The single-list test-score standard deviation of SNR-50 was 1.25 dB for the normal-hearing subjects, giving a standard error of the mean of 0.25 dB.

The performance of the hearing-impaired subjects in group 2 was so diverse that it precluded extraction of the desired normative data. Our motivation for creating an efficient SNR-loss test was the finding that SNR performance is not predictable from hearing thresholds (Killion and Niquette, 2000). Only a few of our subjects had SNR-50 scores at the same SNRs, and none had SNR-50 scores at the higher SNRs (15, 20, and 25 dB). Thus, it was not possible to assess the SNR equivalence of the lists for the higher SNRs. The single-list test-score standard deviation (derived from individual test-retest scores) was 1.4 dB, however, which was an encouraging result since we had predicted 1.9 dB from SIN test data. Nonetheless, the results from the 18 hearing-impaired subjects in experiment 3 made it clear that an exceedingly large number of hearing-impaired subjects would need to be screened in order to obtain enough data for statistically useful tests of list equivalence at the higher signal-to-noise ratios. We therefore chose a different approach in experiment 4.

V. EXPERIMENT 4

The purpose of experiment 4 was to assess list equivalence for five of the six signal-to-noise ratios used in the test (25, 20, 15, 10, and 5 dB).

A. Method

1. Subjects

Twenty-five normal-hearing adult subjects (22 from the University of Iowa Speech and Hearing Clinic and 3 from the Northwestern University Speech and Hearing Clinic) participated in experiment 4. None of these subjects participated in experiments 1, 2, or 3. Normal hearing was defined as pure tone thresholds equal to or better than 20 dB HL for octave frequencies 250 to 4000 Hz. One subject failed this criterion in one ear, with thresholds of 30 and 25 dB HL at 2000 and 3000 Hz, respectively. Subjects ranged in age from 18 to 49 years, with an average age of 26 years and a median age of 24 years.

2. Stimuli

To simulate a wide range of SNR losses, the Beta-list recordings were low-pass filtered. Low-pass cutoff frequencies of 2000, 1400, 1100, 850, and 750 Hz were chosen to provide expected SNR-50 scores of 5, 10, 15, 20, and 25 dB. The choice of cutoff frequencies was made using the count-the-dot articulation index method (Mueller and Killion, 1990) and the relationship between AI and SNR-50 (Killion and Christensen, 1998, Fig. 7). The procedure was as follows: (1) a desired SNR-50 value was chosen; (2) the Killion and Christensen (1998) data were used to find the corresponding AI value; and (3) a low-pass filter cutoff frequency was selected using a trial and error method to filter out the required number of "dots." For example, filtering out all speech sounds above 1100 Hz with a fairly steep filter slope will remove 63% of the speech cues (leaving 37% audible, or an articulation index of 0.37). Figure 7 of Killion and Christensen (1998) indicates that 63 "missing dots" should correspond to a 15 dB SNR-50, whether the cues are inaudible as a result of filtering or hearing loss. As shown in their Fig. 7, the average measured SNR-50 of our subjects under the various filtering conditions was closely correlated with the predicted value. By using these filtered sentences with normal-hearing subjects, only the lack of audible speech cues could be expected to affect the measured SNR-50 values.

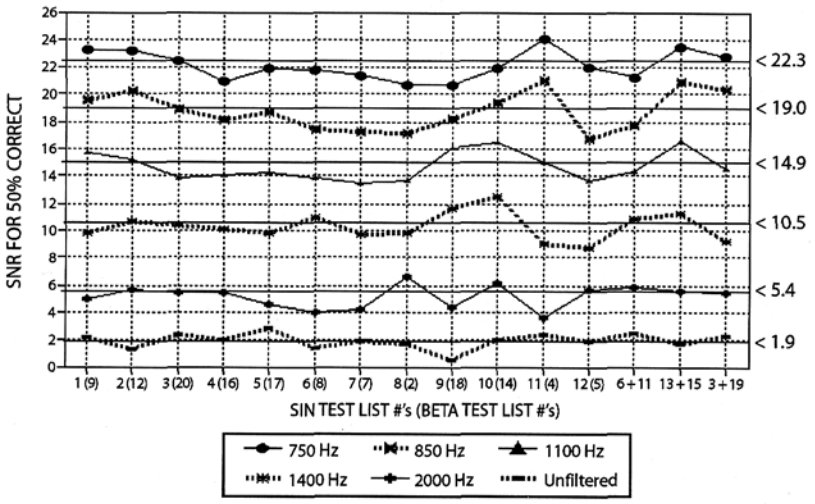

FIG. 6. Across-subject average SNR-50 data for lists included in the final QuickSIN CD. The original list numbers are given in parentheses next to the new list numbers shown on the abscissa.

3. Stimulus presentation

The filtered recordings were presented binaurally at 70 dB HL via ER-3A insert earphones. Subjects were tested in three sessions over several days, and list presentation order was randomized to minimize potential order effects. The most difficult condition (750 Hz low pass) was presented first, followed in order by the less difficult conditions (850, 1100, 1400, and 2000 Hz). Learning effects were not expected to significantly affect the subjects' performance because (a) the testing took place over several days and (b) few words were understood in the more difficult filtering conditions. When similar lists (unfiltered) were retested on the *same* day, Chung (2001) reported learning effects of 1.6 to 3.6 dB, which suggests that learning effects should have little effect on the present results, where each successive list presentation was made at a 5-dB-greater SNR.

B. Results and discussion

1. Selection of lists

The final selection of equivalent lists was made using normal-hearing data from experiment 3 and the filtered data from experiment 4. Twelve lists had experimental SNR values that fell within ±2.2 dB of the average in each condition (unfiltered and all filtered conditions) and became lists 1–12 on the final QuickSIN CD. Three pairs of lists were also found whose *pair averages* met the above criteria. Typically one list score in a pair would be high and the other would be low under similar conditions. These became list pairs 13/14, 15/16, and 17/18 on the final QuickSIN CD. Three lists (labeled A, B, and C) failed the equivalency criteria and were included on the final QuickSIN CD for practice only. These lists can be used to familiarize the listener with the task prior to testing.

Figure 6 shows the across-subject average SNR-50 data for the lists included on the final QuickSIN CD. The original list numbers are given in parentheses next to the new list numbers shown on the abscissa.

2. Application of test

SNR loss is defined as the difference between the test subject's threshold, and average-normal threshold: More precisely SNR loss is equal to the test subject's SNR-50 in dB minus the average-normal SNR-50 in dB. Since the average normal SNR-50 on the QuickSIN test is 2 dB, the simple formula for SNR loss is SNR loss = 25.5 − Total Correct.

The time interval between sentences in the final recording is 5 s, providing adequate response time for most subjects. Each QuickSIN list takes approximately 1 min to administer if the subject responds promptly. If the subject requires more time to respond, the CD can be paused between sentences.

3. AI predictions versus subject performance

The correlation between articulation-index calculated results and 25-subject average results was good. Recall that the expected values of SNR-50 for the filtered lists were 25, 20, 15, 10, and 5 dB, based on the AI predictions. The results from experiment 4 were similar to the predicted values: 22.5, 19.0 14.9, 10.5, and 5.4 dB. Figure 7 shows the correlation between articulation-index calculated results and 25-subject average results.

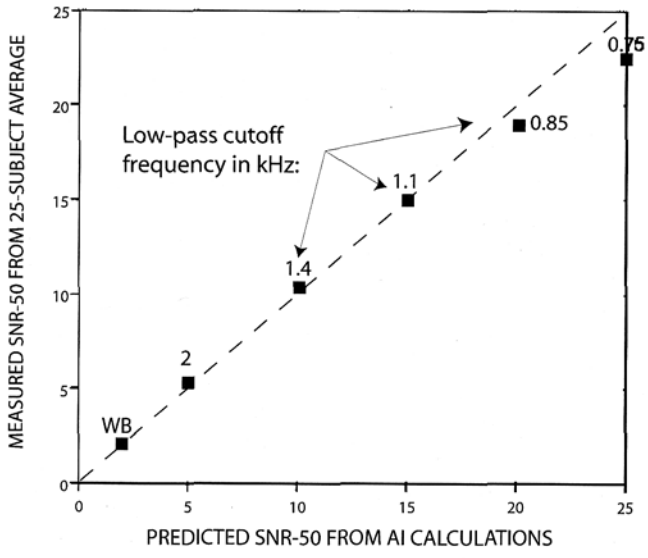

FIG. 7. Comparison between average SNR-50 for normal-hearing subjects listening to filtered speech in noise and predicted values based on the count-the-dots version of the articulation index.

TABLE I. Reliability: number of QuickSIN lists required for a given accuracy. An 80% confidence interval is normally adequate for clinical testing, where the results of any one test are used in context with other factors. A single QuickSIN list is accurate to ±2.2 dB at the 80% confidence interval. A 95% confidence interval is common for research reporting, where a reduced risk of error is normally required. In this case, a single QuickSIN list is accurate to ±2.7 dB at the 95% confidence interval.

	No. of lists								
	1	2	3	4	5	6	7	8	9
95% C.I.+(dB)	2.7	1.9	1.6	1.4	1.2	1.1	1.0	1.0	0.9
80% C.I.+(dB)	2.2	1.6	1.3	1.1	1.0	0.9	0.8	0.8	0.7

VI. GENERAL DISCUSSION

A pure tone audiometric threshold, measured using the standard Hughson–Westlake technique, is accurate to about 5 dB at the 80% confidence level. (1.28×SDEV; Witting and Hughson, 1940). An 80% confidence level is normally adequate for clinical audiometric testing, where the results of any one test are used in context with other factors. A 95% confidence level is common for research reporting, where a reduced risk of error is normally required. A QuickSIN score obtained in 1 min from a single list is accurate to ±2.2 dB at the 80% confidence level and ±2.7 dB at the 95% confidence level.

Table I shows the number of lists required for a given accuracy (such as ±2.2 dB) for confidence levels of 80% and 95%. The numbers in Table I are based on the across-subject root-mean-square standard deviation of 1.4 dB SNR for single-list test scores found for the hearing-impaired subjects in experiment 2. This value comes from two numbers: (a) the 1.3 dB standard deviation derived from the combined individual *test-retest* scores (441 comparisons), and (b) the *across-list* standard deviation of 0.6 dB. If only normal-hearing subjects are used, the single list test score standard deviation drops from 1.4 to 1.25 dB.

A standard deviation of 1.4 dB is slightly better than the standard deviation that would have been expected based on the SIN test, which employs five sentences at each SNR level. As discussed above, the single-block SIN test score standard deviation of 0.85 dB multiplied by the square root of 5 would predict a standard deviation of 1.9 dB for the QuickSIN test, because the QuickSIN test uses only one sentence at each SNR level. The more careful preselection of sentences used in the QuickSIN test may have contributed to the slightly better result.

Averaging the results of several QuickSIN lists improves reliability compared to a single list. For example, one list with the assumed standard deviation of 1.4 dB gives an 80% confidence level of 1.6×1.4= ±2.2 dB. A 50% increase in testing is required to improve from a confidence level of 80% to 95% at a given criterion, e.g., two lists give an 80% confidence interval of ±1.6 dB; three lists provide ±1.6 dB at the 95% confidence level.

The use of multiple lists is particularly important when QuickSIN lists are used to compare two conditions (often two hearing aids or hearing aid adjustments). In this case, the real differences may not be large. A difference of 3.9 dB can be expected 5% of the time when two successive lists are administered, for example, *even if the lists are perfectly equivalent.*

TABLE II. QuickSIN critical differences for comparing two conditions (e.g., two hearing aids or hearing aid adjustments). If one list per condition is used, results must be greater than 3.2 dB different to be considered statistically significant at the 80% confidence interval, and 3.9 dB at the 95% confidence interval.

	Lists per condition								
	1	2	3	4	5	6	7	8	9
95% C.D.+(dB)	3.9	2.7	2.2	1.9	1.7	1.6	1.5	1.4	1.3
80% C.D.+(dB)	3.2	2.2	1.8	1.6	1.4	1.3	1.2	1.1	1.1

The preceding result illustrates the difficulty in validating the equivalence of lists to better than 2 dB. With a standard deviation of 1.4 dB for a single list, even if all 12 lists are administered to ten subjects and their SNR results are averaged, the lists with the highest and lowest scores across the 12 lists will differ by 1.8 dB 5% of the time. If all 12 lists are administered to a single individual, then by the Bonferoni inequality for multiple comparisons (Miller, 1966), a difference of 6.1 dB between the two lists with the highest and lowest measured SNR can be expected 5% of the time.

Table II shows the number of lists required for the comparison between two conditions at an 80% or 95% confidence level. For a critical difference of 1.9 dB, for example, four lists are required for each condition at the 95% confidence level. For a critical difference of 1.4 dB, eight lists are required for each condition at the 95% confidence level.

The use of half-word credit and phoneme scoring were debated during QuickSIN test development. Whole-word scoring was preferred in order to simplify the scoring procedure and to reduce the variability in scoring among clinicians. Achieving consistency in scoring without compromising test accuracy was the goal. Instructions for scoring the SIN test promoted awarding half-credit for partially correct answers because an analysis of the relative independence of IEEE words indicated that 25 words in five sentences, using half-word scoring, gives the equivalent of 27 independent words with whole-word scoring (Fikret-Pasa, 1993). Bentler (2000), however, observed poor interobserver reliability using partial-word scoring. By reexamining the test-retest data from the 26 normal-hearing subjects from experiment 3, whole-word and half-word scoring were compared. Results indicated only a slight improvement (from a standard deviation of 1.25 to a standard deviation of 1.21) when half-word scoring was used. These findings suggest that whole-word scoring on the QuickSIN test is adequate for clinical testing.

VII. CONCLUSIONS

The QuickSIN test provides 12 equivalent lists for testing normal-hearing and hearing-impaired subjects. The test is time efficient; the administration of a single list takes approximately 1 min. The standard deviation for an SNR estimate using a single list is 1.4 dB. Averaging multiple lists results in a lower standard deviation. In addition to the 12 equivalent lists, another six lists are equivalent to the 12 when used in designated list pairs.

ACKNOWLEDGMENTS

The four-talker babble recording was incorporated under license from Auditec of St. Louis. The IEEE recordings were incorporated under license from Massachusetts Institute of Technology. Ruth Bentler and Donna Devine provided much of the subject data for experiments 3 and 4. The following individuals contributed Beta-site data: Nancy Aarts, Harvey Abrams, Rose Allen, Pauline Bailey, Dawn-Marie Bass, Pamela Buethe, Jodi Cook, Paul Efros, David Hawkins, Erica Johnson, Krista Johnson, Melissa Kluck, Barbara Kruger, Lisa Lamson, Stacey Matson, Kristi Mohr, Susan Phillips, Tricia Roh, Juliette Sterkens, Becky Warner, and audiologists at the Eye and Ear Institute, Pittsburgh, PA.

[1]A 50% correct score on words in sentences typically corresponds to 90% correct on complete sentences (e.g., Killion and Christensen, 1998).

[2]After these experiments were completed, we discovered that English was not the first language of one of the subjects, who nonetheless spoke English fluently. After examining the data, we found that this subject's ability to understand speech in noise was indistinguishable from that of the other subjects and, more importantly, it would have made no difference in the results if that subject had been excluded. We thus retained her data.

[3]The calibration tones on the Auditec four-talker babble recording and on the MIT female-talker recording had been adjusted in accordance with Section 6.2.11 of ANSI S3.6-1996 (specification for audiometers): "...the level of the rms sound pressure of a 1000 Hz signal [is to be] adjusted so that the deflection of the volume level indicator produced by the 1000 Hz signal is equal to the average peak deflection produced by the speech signal" (ANSI, 1996). The same calibration method was employed in the NU-4 and NU-6 speech tests (Tillman *et al.*, 1963; Tillman and Carhart, 1966) and the SIN test (Etymotic Research, 1993).

[4]The Tillman–Olsen (1973) recommended method for obtaining spondee thresholds provides a simple method for estimating SNR-50 using the total number of words repeated correctly. In the Tillman–Olsen method, two spondees are presented at each level, starting at a level where all spondees are repeated correctly, then decreasing in 2-dB steps until no responses are obtained for several words. The starting level plus 1 dB, minus the total number of spondees repeated correctly, is the spondee threshold. The simple arithmetic comes from the use of 2-dB steps and two words per step. If the audiometer has 5-dB steps, the corresponding method would use five words per step and take the starting level plus 2.5 dB (half the step size), minus the total number of spondees repeated correctly.

[5]Although this procedure worked well on the average, some sentences with low-variability in sentence score were discarded and some sentences with high-variability were included by this procedure because the statistical variability in standard deviation estimates is moderately large.

ANSI (**1996**). ANSI S3.6-1996 "American National Standard Specification for Audiometers" (American National Standards Institute, New York).

Auditec of St. Louis (**1971**). "Four-talker babble," 2515 S. Big Bend Boulevard, St. Louis, Missouri 63143-2105.

Bacon, S. P., Opie, J. M., and Montoya, D. Y. (**1998**). "The effects of hearing loss and noise masking on the masking release for speech in temporally complex backgrounds," J. Speech Lang. Hear. Res. **41**, 549–563.

Bentler, R. A. (**2000**). "List equivalency and test-retest reliability of the Speech in Noise Test," Am. J. Audiol. **9**, 84–100.

Broadbent, D. E. (**1958**). *Perception and Communication* (Pergamon, New York).

Carver, W. (**1991**). Auditec of St. Louis, St. Louis, MO. Personal communication.

Chung, K. (**2001**). "Effects of input-output functions on speech recognition and preference ratings (for subjects with normal hearing and with selected hearing losses)," doctoral dissertation, Northwestern University, University Microfilms, Ann Arbor, MI.

Cox, R. M., and Moore, J. N. (**1988**). "Composite speech spectrum for hearing aid gain prescriptions," J. Speech Hear. Res. **31**, 102–107.

Cox, R. M., Gray, G. A., and Alexander, G. C. (**2001**). "Evaluation of a revised Speech in Noise (RSIN) test," J. Am. Acad. Audiol **12**, 423–432.

Dirks, D. D. (**1982**). "Comments Regarding 'Speech Discrimination Ability in the Hearing-Impaired,' " in *The Vanderbilt Hearing Aid Report*, edited by G. Studebaker and F. Bess (Monographs in Contemporary Audiology, Upper Darby, PA), pp. 44–50.

Dirks, D. D., Morgan, D. E., and Dubno, J. R. (**1982**). "A procedure for quantifying the effects of noise on speech recognition," J. Speech Hear Disord. **47**, 114–123.

Egan, J. (**1948**). "Articulation testing methods," Laryngoscope **61**, 891–909.

Etymotic Research (**1993**). "The SIN Test," (Compact Disk) 61 Martin Lane, Elk Grove Village, IL 60007.

Fikret-Pasa, S. (**1993**). "The effects of compression ratio on speech intelligibility and quality," doctoral dissertation, Northwestern University, University Microfilms, Ann Arbor, MI.

Hood, J. D., and Poole, J. P. (**1980**). "Influence of the Speaker and Other Factors Affecting Speech Intelligibility," Audiology **19**, 434–455.

Institute of Electrical and Electronics Engineers (**1969**). "IEEE recommended practice for speech quality measurements," Appendix C. Global Engineering Documents, Boulder, CO.

Killion, M. C. (**1997**). "The SIN report: Circuits haven't solved the hearing-in-noise problem," Hear. J. **50**(10), 28–32.

Killion, M. C., and Christensen, L. A. (**1998**). "The case of the missing dots: AI and SNR loss," Hear. J. **51**, 32–47.

Killion, M. C., and Niquette, P. A. (**2000**). "What can the pure-tone audiogram tell us about a patient's SNR loss?" Hear. J. **53**, 46–53.

Killion, M. C., and Villchur, E. (**1993**). "Kessler Was Right—Partly: But SIN Test Shows Some Aids Improve Hearing in Noise," Hear. J. **46**(9), 31–35.

Killion, M. C., Olsen, W. O., Clifford, C. L., VanVliet, D. D., Rose, D. E., Bensen, D. E., Marion, M. W., Tillman, P. A., Hawkins, D. B., Dalzell, S. M., and Fabry, D. A. (**1996**). "Preliminary data on the SIN Test," presented at the annual convention of the American Academy of Audiology, Salt Lake City, UT.

Kochkin, S. (**1992**). "MarkeTrak III identifies key factors in determining consumer satisfaction," Hear. J. **45**(8), 39–44.

Kochkin, S. (**1993**). "MarkeTrak III: Why 20 million in US don't use hearing aids for their hearing loss," Hear. J. **46**(1), 20–27.

Kochkin, S. (**1995**). "Customer satisfaction and benefit with CIC hearing instruments," Hear. Rev. **2**(4), 16–24.

Kochkin, S. (**1996**). "Customer satisfaction and subjective benefit with high performance hearing aids," Hear. Rev. **3**(12), 16–26.

Kochkin, S. (**2000**). "MarkeTrak V: 'Why my hearing aids are in the drawer: The consumers' perspective," Hear. J. **53**(2), 34–42.

Kochkin, S. (**2002**). "10-Year customer satisfaction trends in the US hearing instrument market," Hear. Rev. **9**(10), 14–46.

Lyregaard, P. E. (**1982**). "Frequency selectivity and speech intelligibility in noise," Scand. Audiol. Suppl. **15**, 113–122.

Martin, F. N., Champlin, C. A., and Perez, D. D. (**2000**). "The Question of Phonetic Balance in Word Recognition Testing," J. Am. Acad. Audiol. **11**, 489–493.

Miller, R. (**1966**). *Simultaneous Statistical Inference* (McGraw–Hill, New York).

Mueller, H. G., and Killion, M. C. (**1990**). "An easy method for calculating the articulation index," Hear. J. **43**(9), 14–17.

Nilsson, M., Soli, S. D., and Sullivan, J. A. (**1994**). "Development of the hearing in noise test for the measurement of speech reception thresholds in quiet and in noise," J. Acoust. Soc. Am. **95**, 1085–1099.

Rabinowitz, W. M., Eddington, D. K., Delhorne, L. A., and Cuneo, P. A. (**1992**). "Relations among different measures of speech reception in subjects using a cochlear implant," J. Acoust. Soc. Am. **92**, 1869–1881.

Skinner, M. W. (**1976**). "Speech intelligibility in noise-induced hearing loss: Effects of high frequency compensation," doctoral dissertation, Washington University, University Microfilms, Ann Arbor, MI.

Strom, K. E. (**2003**). "The HR 2003 dispenser survey," Hear. Rev. **10**(6), 522–538.

Taylor, B. J. (**2003**). "Speech-in-noise tests: How and why to include them in your basic test battery," Hear. J. **56**(1), 40–46.

Tillman, T. W., and Carhart, R. (**1966**). "An expanded test for speech discrimination utilizing CNC monosyllabic words: Northwestern University auditory test No. 6," SAM-TR-66-55.

Tillman, T. W., and Olsen, W. O. (**1973**). "Speech audiometry," in *Modern Developments in Audiology (Second Edition)*, edited by J. Jerger (Academic, New York), pp. 37–74.

Tillman, T. W., Carhart, R., and Wilber, L. (**1963**). "A test for speech discrimination composed of CNC monosyllabic words (N.U. auditory test No. 4)," SAM-TDR-62.
Valente, M., and Van Vliet, D. D. (**1997**). "The independent hearing aid fitting forum (IHAFF) protocol," Trends Amplif. **2**(1), 6–35.
van Buuren, R. A., Festen, J. M., and Plomp, R. (**1995**). "Evaluation of a wide range of amplitude-frequency responses for the hearing impaired," J. Speech Hear. Res. **38**, 211–221.
Villchur, E. (**1982**). "The evaluation of amplitude-compression processing for hearing aids," in *The Vanderbilt Hearing Aid Report*, edited by G. Studebaker and F. Bess (Monographs in Contemporary Audiology, Upper Darby, PA), pp. 141–143.
Witting, E. G., and Hughson, W. (**1940**). "Inherent accuracy of a series of repeated clinical audiograms," Laryngoscope **50**, 259–269.