J Am Acad Audiol 12: 423–432 (2001)

Evaluation of a Revised Speech in Noise (RSIN) Test

Robyn M. Cox*†
Ginger A. Gray†
Genevieve C. Alexander*†

Abstract

A revised version of the Speech in Noise (SIN) test was developed by reallocating the recorded test material on the compact disc into different lists (blocks). The goal was to increase the equivalence and reliability of the test blocks to enhance their usefulness in research settings. The Revised Speech in Noise test has four blocks of sentences. Each block comprises twice as many sentences as in the original SIN test. There are also practice sentences for each condition. Forty-two elderly subjects with normal hearing for their age and gender provided data on the equivalence of the new test blocks. The remaining inequalities in mean scores were mostly eliminated using score weighting. Critical differences were developed to promote interpretation of scores from the same individual under different conditions. The revisions substantially improved the equivalence of test blocks and their sensitivity to performance changes. Increased test time is the associated drawback.

Key Words: Hearing aid, hearing loss, noise, speech understanding

Abbreviations: CD = critical difference, MDB = modified dual block, RAU = rationalized arcsine unit, RSIN = Revised Speech in Noise test, SBR = signal-to-babble ratio, SBR-50 = the signal-to-babble ratio corresponding to a score of 50 percent correct, SIN = Speech in Noise test, WRAU = weighted rationalized arcsine unit

Effective communication in noisy situations continues to be the most significant challenge for hearing aid wearers. Hearing aid designers have attempted to address this problem in many different ways. When a new approach or hearing aid model is put forward to improve speech understanding in noise, it is in the interest of manufacturers and dispensers to assess the efficacy of the new approach to determine whether it indeed improves the speech-in-noise problem and, if so, by how much.

Because technical measurements of a hearing aid's performance cannot provide an accurate prediction of speech understanding in noise with that device, intelligibility of amplified speech in noise must be measured directly. The ideal approach to measuring speech understanding would yield a score that could provide an accurate prediction of abilities in the diverse listening environments of daily life. Standardized tests currently available for measurement of speech understanding fall short of this ideal because of constraints imposed by feasible test administration time and the difficulties in developing valid test procedures. Thus, the search continues for a test that is a more accurate predictor of speech understanding in the noisy situations of everyday living.

The Speech in Noise (SIN) test was first described in 1993 (Fikret-Pasa, 1993; Killion and Villchur, 1993). It was designed to test speech understanding in noise for both soft and loud speech and in a range of signal-to-noise ratios that encompass rather easy to very difficult conditions. The parameters of presentation level and background noise were selected to optimize the SIN test for evaluating the assistance provided by a hearing aid under the types of conditions experienced in daily life. The original test comprises nine blocks (lists) of sentences. Bentler (2000) reported equivalence data for the nine SIN test blocks, obtained from 20 normal-hearing and 20 hearing-impaired listeners. These data indicated that the nine blocks did not yield equivalent scores (especially when evaluated at individual signal-to-noise ratios).

*Department of Veterans Affairs Medical Center, and †School of Audiology and Speech Pathology, The University of Memphis, Memphis, Tennessee

Reprint requests: Robyn M. Cox, Memphis Speech & Hearing Center, 807 Jefferson Ave., Memphis, TN 38105

We used Bentler's data to reallocate the test sentences into "modified dual blocks" (MDBs) that were hypothesized to be more equivalent and reliable than the blocks in the original test. This article reports the details of test revision and evaluation.

METHOD

Original SIN Test

The SIN test sentences are spoken by a female talker in the presence of a four-talker speech babble. The test is recorded on a compact disc (Etymotic Research, 1993). Each of the nine test blocks of the SIN test comprises 40 sentences. It is intended that 20 sentences will be presented at 70 dB HL (83 dB SPL). The other 20 sentences are designated for presentation at 40 dB HL (53 dB SPL). In this investigation, these were called the H-sentences (high level) and the L-sentences (low level), respectively.

The 20 sentences designated for each presentation level are further divided into four signal-to-babble ratios (SBRs): 0 dB, +5 dB, +10 dB, and +15 dB. At each level, 5 sentences are presented at each SBR. Each sentence contains five scoring words. Thus, a percent correct score is obtained for each SBR condition for both H-sentences and L-sentences. In addition, for both H-sentences and L-sentences, the scores for the four SBR conditions can be used to generate two further scores: (1) a mean overall score and (2) the SBR needed for a score of 50 percent correct.

Revised Speech in Noise Test

The original recording of the SIN test sentences (Etymotic Research, 1993) is used for the revised test (RSIN) as well. Bentler (2000) provided mean scores obtained for each SBR condition at each presentation level for all nine original test blocks. These data were used to reallocate the test material into new combinations. For each SBR condition, four combinations of two SIN blocks were designated based on the data obtained by Bentler. The goal was to produce revised, expanded test blocks that would be more equivalent to each other. For example, in the +10-dB SBR condition, the following combinations were formed for the L-sentences: SIN blocks 2 and 6, 4 and 9, 1 and 5, and 3 and 8. The unused SIN block (7) became the practice sentences for this RSIN condition. Bentler (2000) noted that floor and ceiling problems were fairly common in her data. To minimize these effects, the RSIN combinations were based on data from all 20 hearing-impaired subjects for the +5-, +10-, and +15-dB SBR conditions. For the 0-dB SBR condition, combinations were based on data for the normal-hearing subjects who did not have any zero scores.

In the RSIN test, each test block is comprised of 80 sentences. There are 40 H-sentences and 40 L-sentences in each block. For both H-sentences and L-sentences, 10 sentences are administered at each SBR. To formalize these changes from the original test, each test unit of the RSIN test is called an MDB. There are four MDBs. The test material that was not reallocated to an MDB was used as practice material. There are five practice sentences for each combination of level (H or L) and SBR.

Subjects

A total of 42 subjects were randomly divided into two groups of 21. All subjects were older than 60 years, and their hearing sensitivity thresholds were better than the 90th percentile of thresholds for otologically normal individuals of their age and gender (International Standards Organization, 1984). Figure 1 depicts the mean audiograms for the two groups. Each subject's ears were otoscopically normal, and immittance testing confirmed normal middle ear function. There were 29 women (mean age = 66.0 years) and 13 men (mean age = 66.1 years).

Evaluation Procedures

For testing, subjects were seated in a double-walled, sound-treated room. The test sentences were delivered bilaterally using Etymotic ER-1 insert earphones. These earphones produce a signal that is equivalent to what would be received

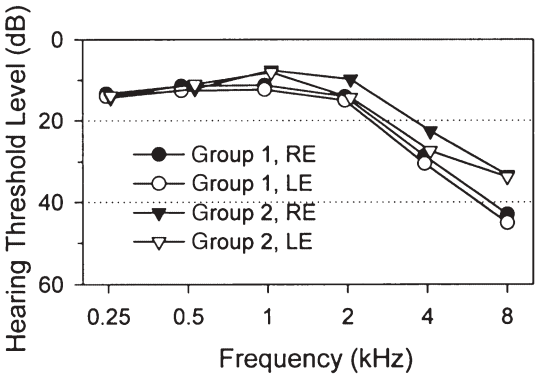

Figure 1 Mean audiograms for the two groups of subjects. RE = right ear, LE = left ear.

if the subjects were listening to the same material in a diffuse sound field. Thus, the test conditions simulated binaural soundfield listening.

In this study, we were evaluating the characteristics of the sentences rather than those of the listeners. To facilitate this goal, all of the sentences (both H-sentences and L-sentences) were presented at a level that was comfortably loud. The rationale was that this would eliminate the potential ambiguity that would result if some of the test material was not fully audible (too soft) or exceeded the subject's undistorted listening levels (too loud). The frequent peaks of the sentences, measured in a 2-cc coupler, were about 73 dB SPL.

Subjects in group 1 listened to the sentences without any frequency shaping. Subjects in group 2 listened to sentences that had been low-pass filtered by about 9 dB per octave from 250 to 4000 Hz to simulate the potential effects of a mild high-frequency sloping hearing loss. The goal of this maneuver was to allow us to evaluate whether this degree of change in audibility would significantly impact the equivalence of the test blocks.

All subjects heard the entire test in one test session. The five practice sentences for the new condition were administered every time the SBR was changed. Administration order of H-sentences and L-sentences and of MDBs was counterbalanced across subjects. Within each MDB, listening conditions progressed from easy (SBR = +15 dB) to difficult (SBR = 0 dB), as in the original SIN test.

The instructions were as follows:

> Imagine that you are at a party. You are going to listen to one female friend with several other people talking in the background. This friend can be easily identified during the first few sentences because her voice is louder than the others.
>
> We want you to repeat the sentence spoken by your friend. The people talking in the background will gradually get louder, making it difficult to identify your friend's voice.
>
> This is a difficult test, and you may not be able to repeat all of the words. Repeat as much of each sentence as possible (even a word or part of a word), even if you need to guess.

The RSIN test was administered using purpose-developed software (Brainerd, 2001). The compact disc was played from the computer's CD-ROM drive under software control. The test sentence was presented, the subject repeated it, and the experimenter selected a score of correct, half-correct, or incorrect for each key word. The next sentence was then played. Key words were scored as correct if they were repeated perfectly: correct word order was not a requirement. A response was scored as half-correct if it was the correct root word but in a different form: for example, CAT instead of CATS was scored half-correct, but PATS instead of CATS was scored incorrect. This rule was adopted to optimize the reliability of partial scoring. Scoring was tracked by the software.

RESULTS

Each MDB of the RSIN test returns 12 scores. For the H-sentences, the six scores are as follows: (1) a percent correct score for each of the four test SBRs, (2) the SBR needed to give a 50 percent correct score (determined from the scores at each test SBR), and (3) an overall percent correct score. A corresponding set of six scores is returned for the L-sentences.

In the present study, the percent correct scores were transformed into rationalized arcsine units (RAUs), as described by Studebaker (1985), to homogenize the variances. RAU scores are similar to the corresponding percentage within the range of scores from about 20 to 80. Outside this range, RAU scores progressively deviate from the corresponding percentages. The total range of RAU scores is –23 to +123.

Effect of Low-Pass Filtering

Figure 2 depicts the mean RAU score at each SBR for each MDB for the H-sentences. Data are shown separately for group 1 (filled

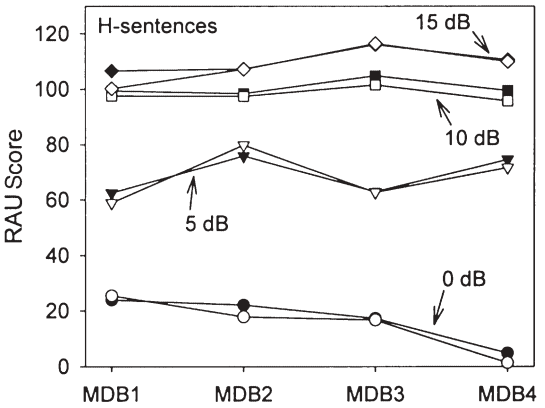

Figure 2 Mean RAU score for each modified dual block (MDB) for the H-sentences. The parameter is signal-to-babble ratio. Data are shown separately for groups 1 (*filled symbols*) and 2 (*open symbols*).

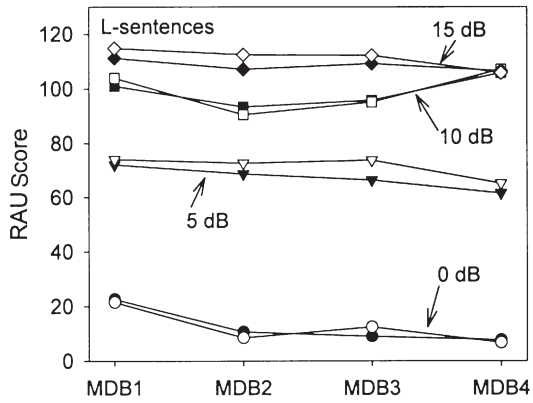

Figure 3 Mean RAU score for each modified dual block (MDB) for the L-sentences. The parameter is signal-to-babble ratio. Data are shown separately for groups 1 (*filled symbols*) and 2 (*open symbols*).

Table 1 Results of Testing the Equivalence of the Four MDBs at Each SBR Using Multivariate Pairwise Comparisons at the p < .05 Level, with Bonferroni Adjustment for Multiple Comparisons

SBR (dB)	*MDB**
H-sentences	
0	4 3 2 1
5	1 3 4 2
10	4 2 1 3
15	1 2 4 3
L-sentences	
0	4 2 3 1
5	4 3 2 1
10	2 3 1 4
15	4 2 3 1

*Modified dual blocks (MDBs) that are joined by an underline are not significantly different from each other.

symbols) and group 2 (open symbols). The corresponding data for the L-sentences are shown in Figure 3.

Figures 2 and 3 both suggest that the moderate low-pass filtering used to simulate mild high-frequency hearing loss did not change the relative equivalence of the test blocks. Consider, for example, the two lowest lines in Figure 3. Mean scores are shown for the four MDBs at the 0-dB SBR. Scores for unfiltered sentences are illustrated with filled circles, whereas those for filtered sentences are shown using open circles. The scores are interweaving and very similar for each MDB. The other SBR conditions in Figures 2 and 3 reveal the same pattern. The largest mean differences of 5 to 10 RAUs are seen for the +5-dB SBR condition in Figure 3. Statistical testing (multivariate analysis of variance) confirmed the impression that there were no significant differences ($p > .05$) between the mean scores for filtered and unfiltered sentences.

Equivalence of Modified Dual Blocks

For a test such as the RSIN test, which is designed to be used to compare different amplification conditions, it is highly desirable for the different test forms (the four MDBs) to be equivalent in difficulty. If this is not the case, then a difference observed across amplification conditions tested with different MDBs cannot be definitively interpreted.

If the MDBs were perfectly equivalent and there was no measurement error, all of the data lines in Figures 2 and 3 would be flat, indicating the same score at a given SBR for each MDB. Although the lines are not flat, suggesting a lack of perfect equivalence across MDBs, statistical testing was needed to determine whether the conditions produced significantly different scores or whether the observed differences should be attributed to measurement error. For these tests, data from groups 1 and 2 were combined, based on the finding of no significant difference between them.

To examine the equivalence of the four MDBs in each SBR condition, preliminary multivariate repeated-measures analyses of variance were performed on the data for the L-sentences and again on the data for the H-sentences. Both analyses revealed significant interactions between the SBR and MDB variables. These interactions were further tested by examining the differences across the four MDBs for each SBR condition. The results are shown in Table 1.

In Table 1, the ideal result for each SBR condition would be a single line connecting all four MDBs. This would indicate that there were no statistically significant differences across the four MDBs, and it would be reasonable to conclude that they are equivalent for test purposes. This ideal result was not seen in any SBR condition. All SBR conditions revealed at least one significant difference across the four MDBs.

Determining Score Weights to Increase Equivalence across MDBs

The analyses reported in Table 1 revealed numerous systematic differences in difficulty across the four MDBs for the eight SBR conditions. One approach to minimizing this problem

Table 2 Weighting Factors Determined for Use with RAU Scores in Each Combination of MDB and SBR

SBR (dB)	H-Sentences				L-Sentences			
	MDB1	MDB2	MDB3	MDB4	MDB1	MDB2	MDB3	MDB4
15	1.010	1.028	0.940	1.015	0.960	1.005	0.993	1.036
10	1.002	1.009	0.960	1.014	0.959	1.071	1.033	0.927
5	1.112	0.875	1.085	0.930	0.939	0.974	0.989	1.081
0	0.705	0.835	0.950	1.110	0.612	1.122	1.013	0.978

is to apply weights to the scores to make them more equivalent. A subject's score for a condition is multiplied by the weighting factor for that condition. Using this strategy, scores from conditions known to be relatively difficult are increased somewhat, whereas scores for conditions known to be relatively easy are decreased somewhat.

A weighting factor was determined for each SBR of each MDB. Identical procedures were followed for both L-sentences and H-sentences. First, for each subject, the mean RAU score across the four MDBs was computed for each SBR condition. The mean score was assumed to be the best estimate of the subject's true score for that SBR condition. Second, for each MDB, regression analyses were performed between observed and true scores for each SBR. The intercept was specified to equal zero to maximize the applicability of the weights to a variety of audibility and hearing loss conditions. These analyses determined the weighting factor for each SBR that would most accurately transform the observed RAU scores into the true RAU scores. To test the robustness of the obtained weights across different audibility conditions, the analyses were run independently on group 1 and group 2 data. In addition, if a subject obtained either 0 or 100 percent correct for a condition before transformation into RAUs, the data were not used to determine the weight for that condition.

This process resulted in a set of weights for each group of subjects. Each set contained a separate weighting factor for each SBR of each MDB for both L-sentences and H-sentences. Examination of the weights for the two groups revealed them to be within 0.1 of each other in 30 of the 32 conditions. In the other 2 conditions, the weights for the two groups differed by 0.11 and 0.21. Because they were extremely similar overall, the two sets of weights were averaged to produce the final weighting factors. The final weights for RAU scores are reported in Table 2.

Equivalence of Weighted MDB Scores

Each subject's RAU scores were weighted using the derived weighting factors. Figures 4 and 5 show the mean weighted RAU (WRAU) scores for the H-sentences and L-sentences, respectively. In Figure 4 (H-sentences), the mean scores for the four MDBs essentially overlap in the 5- and 10-dB SBR conditions. This outcome indicates that the application of score weighting has equated the MDBs in these conditions. In the 0- and 15-dB SBR conditions, the scores for the four MDBs are fairly similar but not perfectly overlapping, indicating some persistent mean differences among the four MDBs. This occurred because the method used to derive the weights produced less than perfect corrections for the most deviant MDB conditions. Nevertheless, multivariate pairwise comparisons at the $p < .05$ level (with Bonferroni adjustment) revealed that only two significant mean differences remained after weighting: MDB4 at 0-dB SBR and MDB1 at 15-dB SBR.

In Figure 5 (L-sentences), the four MDBs give essentially the same mean weighted scores in the 5-, 10-, and 15-dB SBR conditions. In the

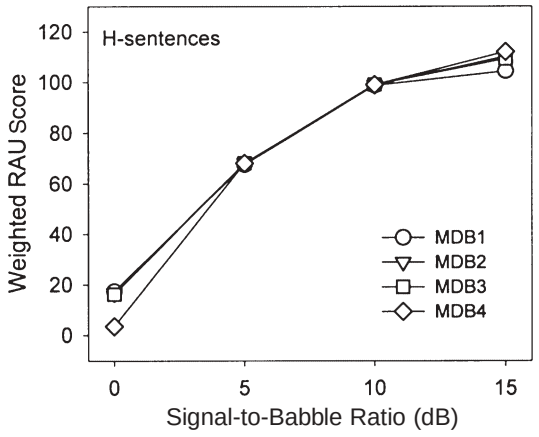

Figure 4 Mean weighted RAU scores for the H-sentences. Data are shown for each MDB in each SBR condition.

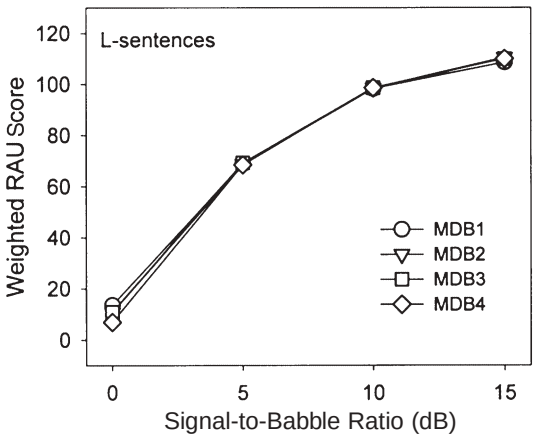

Figure 5 Mean weighted RAU scores for the L-sentences. Data are shown for each MDB in each SBR condition .

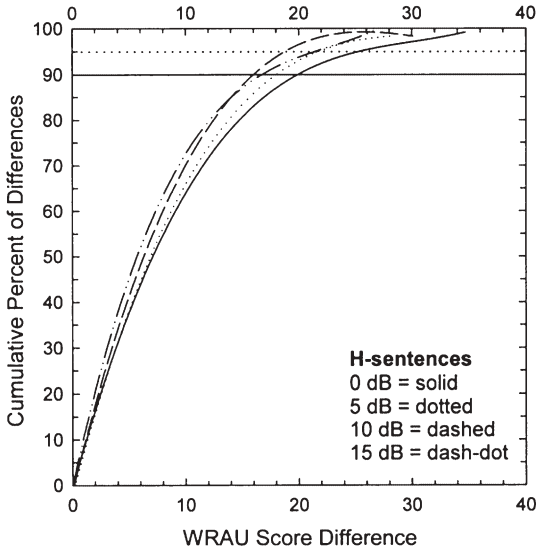

Figure 6 Cumulative distributions of differences in scores between pairs of modified dual blocks (MDBs) for the H-sentences. Each distribution includes differences derived from all six pairs of MDBs for all 42 subjects. A separate distribution is shown for each SBR condition. Third-order polynomials have been fitted to the data. WRAU = weighted rationalized arcsine unit.

0-dB SBR condition, however, some inequalities remain among the four MDBs. The only statistically significant differences are between MDB1 and MDB4 at the 0-dB SBR condition. Again, this result is attributable to the effects of the method used to derive the weights.

Overall, the application of weights to the RSIN test scores greatly improved the equivalence of the four MDBs.

Comparing RSIN Test Scores for Individual Subjects

As Figures 4 and 5 demonstrate, the four MDBs of the RSIN test are almost equivalent, after weighting, when compared in terms of group mean scores. This indicates that the RSIN test should be well suited to delineating differences between listening conditions for research purposes when group data are used. However, there are many applications when it is necessary to compare scores for different listening conditions within a single subject. For example, a clinician might wish to use the test to compare hearing aid A with hearing aid B for a particular patient. In this kind of application, it is necessary to have guidelines for interpreting a difference between two scores observed on the same person. This type of guideline is called a critical difference (CD).

A CD can be used to determine the probability that a given difference between two scores was obtained by chance owing to measurement error. If this probability is sufficiently small, it is reasonable to conclude that the observed difference between two scores illustrates a genuine disparity between the tested conditions. In the present study, CDs for the RSIN test were estimated as described below.

For each SBR condition in the H-sentences data, the absolute difference between the weighted scores for each pair of MDBs was derived for each subject. Since the four MDBs can be combined into six different pairs, each subject contributed six difference scores. Thus, for each SBR condition, there were 252 differences (6 pairs $\times$ 42 subjects) between pairs of weighted scores. A corresponding set of data was derived for each SBR condition in the L-sentence data.

A cumulative distribution of absolute differences was generated for each SBR condition. Analysis of each cumulative distribution revealed that a third-order polynomial could be derived that described more than 99.5 percent of the variance in the data. Figures 6 and 7 depict the polynomials describing the between-MDB differences for each SBR condition in the H-sentences and L-sentences, respectively.

These figures can be used to estimate the CDs for the various conditions. To facilitate the procedure of deriving a CD, each figure has horizontal lines at the 90 and 95 percent cumulative levels. A polynomial intersects with the 90 percent line at the x-axis value that is the 90 percent CD for that SBR condition. All values that lie above the 90 percent line represent differences

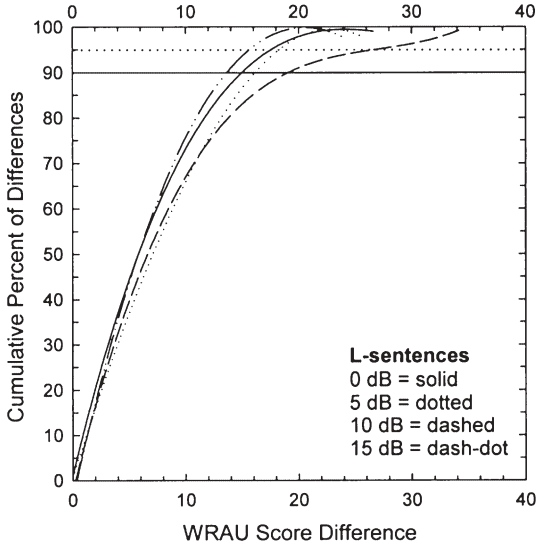

Figure 7 Cumulative distributions of differences in scores between pairs of modified dual blocks (MDBs) for the L-sentences. Each distribution includes differences derived from all six pairs of MDBs for all 42 subjects. A separate distribution is shown for each SBR condition. Third-order polynomials have been fitted to the data. WRAU = weighted rationalized arcsine unit.

in scores between two MDBs that occur *by chance* 10 percent of the time or less. For example, consider the solid polynomial (0-dB SBR curve) in Figure 6. It intersects the 90 percent line at an x-axis value of 20 WRAUs. Thus, 20 WRAUs are the 90 percent CD for the H-sentences' 0-dB SBR condition. Table 3 gives the 90 percent CD for each condition.

CDs for other probability levels can be determined using other levels of the cumulative distribution. For example, many research applications might use a 95 percent CD. This would be determined by using the line at the 95 percent level of the distribution.

Table 3 Ninety Percent Critical Differences, in Weighted RAUs, for H-Sentences and L-Sentences in Each SBR Condition

SBR (dB)	*H-Sentences*	*L-Sentences*
0	20	15
5	18	16
10	16	19
15	17	13

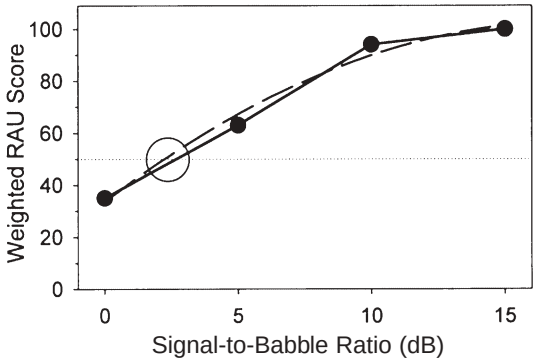

Figure 8 Example for one subject and one modified dual block showing the typical difference observed when SBR data were fitted with straight lines (*solid line*) versus a second-order polynomial (*dashed line*). The dotted line depicting a WRAU score of 50 intersects the two data functions at slightly different SBR values (shown in the circle).

Using the RSIN Test to Measure Changes in Understanding in Noise

A principal application of the SIN test, as envisioned by its developers, is the determination of the SBR corresponding to a score of 50 percent correct understanding. Following Killion and Niquette (2000), we refer to this metric as the SBR-50. The SBR-50 is often used to quantify a listener's ability to understand speech in a noisy environment. For example, two listening conditions (aided and unaided) or two different hearing aids might be compared in terms of the SBR needed to produce a 50 percent score. A change in the SBR-50 can signify that the ability to understand speech in noise has been either improved or degraded.

For such comparisons to be valid, it is important to be aware of the inherent equivalence of the test forms in terms of the SBR needed to produce 50 percent scores. To evaluate this aspect of the four MDBs in the RSIN test, the SBR-50 was determined for each MDB for each subject. Note that after the scores have been weighted to optimize equivalence across forms, a score of 50 WRAUs is still close to the middle of the potential scoring range, even though it does not denote the precise middle of the range. To maintain comparability with other tests, the analyses of the weighted scores focused on the SBR needed for a score of 50 WRAUs correct.

For a given MDB and sentence level, the SBR-50 can be determined in two ways, as depicted in Figure 8: (1) to connect the WRAU scores for each of the four SBR conditions with straight lines, draw a line from the ordinate at

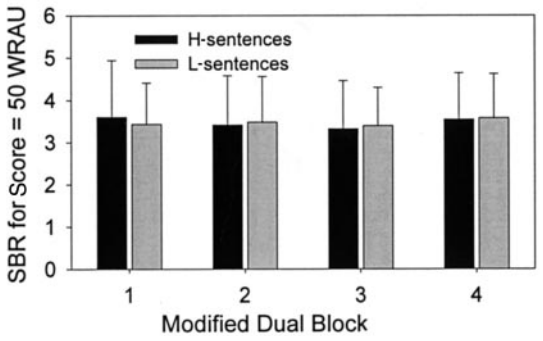

Figure 9 The mean SBR-50 score for each sentence type and each modified dual block. These data were determined using the straight-line method of fitting the data illustrated in Figure 8.

50 WRAUs to intersect the data function and note the SBR value at which this intersection occurs and (2) to fit a second-order polynomial to the four data points using a least squares method and solve the resulting equation for SBR when the WRAU score equals 50. We used both methods and found them to produce almost identical results for many subjects. When there were differences, the typical result for one subject and one MDB is illustrated in Figure 8. The dotted line depicting a 50-WRAU score intersects the two data functions at slightly different SBR values (shown in the circle). The SBR-50 value computed for the polynomial fit is at 2.25 dB, which is slightly lower than the SBR-50 of 2.75 dB observed for the straight-line fit.

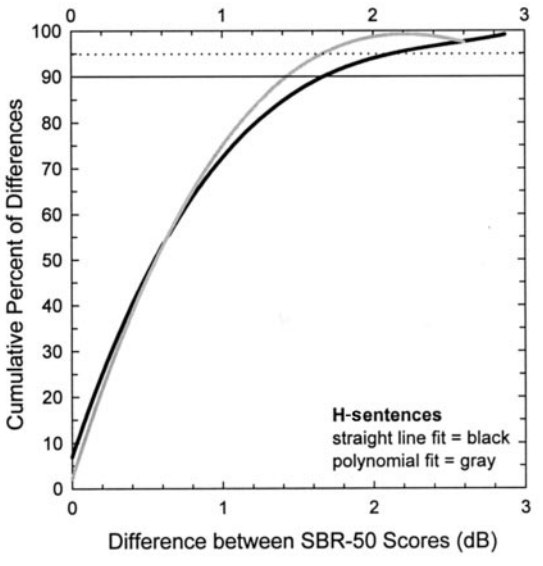

Figure 10 Cumulative distributions of differences between pairs of SBR-50 scores for the H-sentences. Each distribution includes differences derived from all six pairs of modified dual blocks for all 42 subjects. The black curve gives the result when the SBR-50 scores were obtained using the straight-line method. The gray curve gives the result when the SBR-50 scores were obtained using the polynomial fit method. Third-order polynomials have been fitted to the data.

Figure 9 indicates the mean SBR-50 for each MDB. These data were determined using the straight-line method of fitting the data. As predictable from the example shown in Figure 9, mean SBR-50 scores derived using the polynomial fitting method were slightly smaller than these. The means in Figure 9 varied from 3.3 to 3.6 dB across the eight conditions. Multivariate analysis determined that, on average, the eight conditions were not significantly different from each other ($p > .05$). This result supports the equivalence of the four MDBs when used for group data.

When we wish to compare two SBR-50 scores obtained *from the same individual* under different conditions such as two different hearing aids, a CD value is needed. The method used to determine CDs was the same as used above for the WRAU scores for individual SBR conditions. That is, a cumulative distribution was constructed from the differences between all pairs of SBR-50 scores. We postulated that because the polynomial method of determining the SBR-50 uses all four data points instead of only two, it might produce more reliable results and thus smaller CDs. This was found to be the case, as illustrated in Figures 10 and 11.

Figures 10 and 11 illustrate the cumulative distributions of differences between pairs of SBR-50 scores when those scores were obtained using the straight-line method (black curve) and the polynomial fit method (gray curve). As in Figures 6 and 7, a third-order polynomial, fitted to the data, is displayed for each distribution. The polynomials explained 96 to 98 percent of the data in the cumulative distributions. Once again, to determine the 90 percent CD, we find the x-axis score that corresponds to the solid horizontal line in the figures. Both figures show that the 90 percent CD is smaller for the polynomial-fitted method than for the straight-line method. Table 4 gives the values for 90 and 95 percent CDs. The CD values range from 1.1 to 2.1 dB.

DISCUSSION

The process of reallocating sentences, expanding blocks, and weighting scores resulted in much greater average equivalence across the four MDBs of the RSIN test than across the

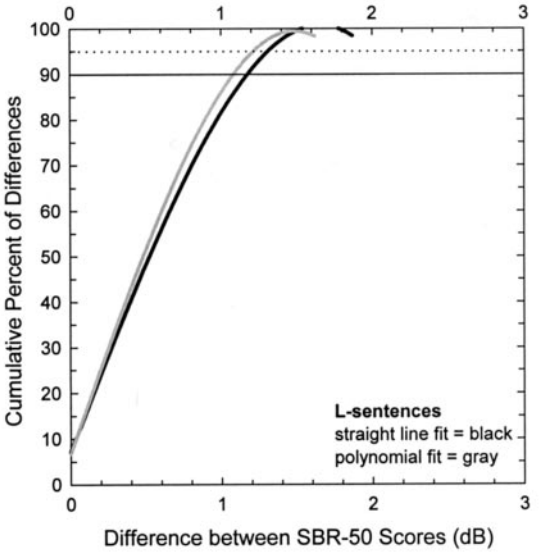

Figure 11 Cumulative distributions of differences between pairs of SBR-50 scores for the L-sentences. Each distribution includes differences derived from all six pairs of modified dual blocks for all 42 subjects. The black curve gives the result when the SBR-50 scores were obtained using the straight-line method. The gray curve gives the result when the SBR-50 scores were obtained using the polynomial fit method. Third-order polynomials have been fitted to the data.

nine blocks of the SIN test. This can be seen by comparing Figures 4 and 5 in this study with Figures 1 and 2 in Bentler's data (Bentler, 2000). This was achieved at a cost of greater testing time per score (more sentences) and a smaller number of different test forms. Each researcher must decide whether this is an acceptable tradeoff for his or her particular application.

Because all of these data were obtained from listeners with normal hearing for their age and gender, it is reasonable to ask whether the equivalence results can be expected to apply to listeners with hearing loss. Theoretical considerations suggest that these results can probably be generalized with acceptable accuracy, at least to individuals who have mild to moderate hearing losses. Data indicate that mild to moderate hearing impairments are mostly the result of outer hair cell loss (e.g., Killion and Niquette, 2000). This type of impairment primarily produces a loss of audibility, which can be rather accurately simulated in normal-hearing listeners by low-pass filtering, as used in this study (e.g., Zurek and Delhorne, 1987; Dubno and Schaefer, 1992). Numerous studies support the conclusion that most of the variance in speech recognition scores for hearing-impaired listeners can be explained by differences in audibility (e.g., Humes et al, 1994).

Table 4 90 and 95 Percent Critical Differences for SBR-50 Scores (dB SBR)

Analysis Method	*H-Sentences*		*L-Sentences*	
	90%	*95%*	*90%*	*95%*
Straight line	1.7	2.1	1.2	1.3
Polynomial	1.4	1.7	1.1	1.2

Two further aspects of our findings suggest that the equivalence data presented here should hold up well for many individuals with sloping high-frequency hearing loss. First, the pattern of differences among MDBs was very consistent across the different audibility conditions used for groups 1 and 2. Second, the weights derived to improve equivalence of scores across MDBs were essentially identical for the two groups of subjects. Nevertheless, the MDBs might not be equivalent, on average, for individuals who have sharply falling or rising audiogram configurations or for those whose hearing impairment exceeds about 60 dB. Caution should be used in interpreting results in these cases.

The CDs shown in Figures 6 and 7 and Table 3 merit some consideration. The 90 percent CDs for both H-sentences and L-sentences range from about 13 to 20 WRAUs. This means that when comparing two scores from the same individual *in only one SBR condition,* a difference of 13 to 20 WRAUs (depending on SBR) is needed before it can be concluded with reasonable confidence that the scores were obtained under genuinely different conditions. For example, if hearing aid program 1 is compared with hearing aid program 2 by testing each one with a different MDB at the 10-dB SBR condition and the H-sentences, a score difference of 16 WRAUs or more is needed to justify a conclusion that one program was better. This would be a total of 20 test sentences, or about 5 minutes of testing, not counting the time to change the hearing aid program.

However, the uniqueness of the SIN and RSIN tests lies in their potential to delineate speech understanding ability under a variety of conditions that are similar to those of daily life. This should include testing at more than one SBR and possibly at more than one presentation level. The conditions tested (e.g., programs 1 and 2) can then be compared across a set of listening conditions. How can we determine a

CD for this type of comparison? A reasonable approach is to assume that the appropriate CD is reflected by the joint probability of obtaining certain simultaneous differences between scores *by chance alone*. The joint probability of several independent events occurring simultaneously by chance is equal to the product of the separate probabilities of each separate event occurring by chance. Thus, an estimate of the probability of observing, by chance, a pattern of scores in which one condition is consistently superior to the other condition can be derived by multiplying the separate probabilities of observing each difference alone. These separate probabilities can be derived from Figures 6 and 7.

Let us consider an example. Suppose hearing aid program 1 is being compared with program 2 using the H-sentences. Program 1 is tested at 5- and 10-dB SBR, and the scores are 42 and 60 WRAUs, respectively. Program 2 is tested at 5- and 10-dB SBR, and the scores are 54 and 75 WRAUs, respectively. Figure 6 indicates that the probability of obtaining a score difference owing to a measurement error of 12 WRAUs at 5 dB SBR is about .27 ([100 – 73]/100). The corresponding probability of obtaining a score difference of 15 WRAUs at 10-dB SBR is about .14 ([100 – 86]/100). Neither of these exceeds a 90 percent CD. However, the probability of obtaining this combination of independent differences by chance alone is .27 × .14, which is roughly .04. Thus, the likelihood that this pattern of differences between programs 1 and 2 occurred by chance is about 4 percent—a small enough probability to justify a reasonably confident conclusion that program 2 is superior to program.[1]

Killion and Villchur (1993) reported that for the SIN test, the SBR-50 was 1 dB for normal-hearing individuals. Despite their nominally normal hearing, our subjects needed, on average, about 3-dB SBR to achieve this level of performance, as shown in Figure 9. This can probably be attributed to a combination of effects owing to age and high-frequency hearing sensitivity (see Fig. 1). It is noteworthy that a 3-dB SBR-50 is consistent with the performance of the best of Bentler's hearing-impaired subjects (Bentler, 2000, Fig. 5).

As shown in Figures 10 and 11, the CDs for SBR-50 are minimized when the intelligibility scores at the four SBRs are fitted with a second-order polynomial rather than simple straight lines. The 95 percent CDs are 1.7 dB or less, and 90 percent CDs are 1.4 dB or less. This is a considerable improvement in sensitivity over the SIN test, based on Bentler's data for normal listeners. She reported the 95 percent CDs for SBR-50 to be 2.4 to 2.6 dB (Bentler, 2000). These results mean that if the RSIN test is used to compare two conditions tested on the same individual, and SBR-50 scores are determined by fitting a second-order polynomial to the data, a difference of 1.4 (H-sentences) or 1.1 dB (L-sentences) is sufficient to support a reasonably confident (90% certainty) decision that the two conditions are different.

It is interesting to note in Table 4 that the CDs for the L-sentences are quite a bit smaller than those for the H-sentences. This does not reflect anything about the presentation levels because all of the sentences were actually presented at essentially the same levels in this study. Instead, these differences probably reflect inherent, unintentional differences in the test materials. This result suggests that, in cases where it is important to maximize the sensitivity of the data, the L-sentences would be the optimal test material.

Research and Clinical Applications of the RSIN Test

The MDBs of the RSIN test offer the option of capitalizing on the advantages of the SIN test while obtaining more reliable and sensitive data. However, these advantages can be obtained only at the price of additional testing time. Thus, the RSIN test is probably more suited to research applications than to clinical ones. On the other hand, the automated scoring of the software-driven RSIN test saves time while promoting accuracy.

One problem that can be encountered with the SIN and RSIN tests is that of performance extremes. When the L-sentences are administered at 40 dB HL and the H-sentences are administered at 70 dB HL (as the test was designed), many hearing-impaired subjects score 0 percent on some conditions and/or 100 percent on others (see Bentler, 2000, for an example). That problem was mostly avoided in this study by adjusting the presentation levels of the sen-

[1]Note that this reasoning is valid only when a pattern of consistent differences is observed between the tested conditions and you wish to determine the likelihood that this pattern of differences is owing to chance. Also, the joint probability computed with this method is an estimate rather than a precise value.

tences so that all of them were heard at a comfortable level.

In principle, the RSIN test sentences can be presented at any levels that seem appropriate to answer the questions of interest. However, it is important to be aware of the potential effects of extreme scores and to have a plan for treating them. The plan should be based on consideration of the relative validity, in the particular application, of discarding versus keeping the extreme scores.

It is important to keep in mind that all of the data reported here for the RSIN test were obtained with the practice sentences administered whenever the SBR was changed. Past research has suggested to us that providing practice listening to the new SBR is important in promoting consistent data. Thus, use of the practice passages is highly recommended.

Acknowledgment. This article is based on work supported by the Office of Research and Development, Rehabilitation R&D Service, Department of Veterans Affairs. Vanessa Kendrick and Melissa Franklin helped with data collection and Greg Flamme assisted with test revision. Appreciation is owing to Ruth Bentler for allowing us to use her data as the basis for test revision.

REFERENCES

Bentler RA. (2000). List equivalency and test-retest reliability of the Speech in Noise test. *Am J Audiol* 9:84–100.

Brainerd LE. (2001). *Software to Administer and Score the Revised Speech in Noise (RSIN) Test* [Web Page]. www.ausp.memphis.edu/harl. Accessed Sept. 2000.

Dubno JR, Schaefer AB. (1992). Comparison of frequency selectivity and consonant recognition among hearing-impaired and masked normal-hearing listeners. *J Acoust Soc Am* 91:2110–2121.

Etymotic Research. (1993). *The SIN Test Compact Disc.* Elk Grove, IL: Martin Lane.

Fikret-Pasa S. (1993). *The Effects of Compression Ratio on Speech Intelligibility and Quality.* Ann Arbor, MI: University Microfilms, Northwestern University.

Humes LE, Watson BU, Christensen LA, Cokely CG, Halling DC, Lee L. (1994). Factors associated with individual differences in clinical measures of speech recognition among the elderly. *J Speech Hear Res* 37:465–474.

International Standard Organization. (1984). *Acoustics—Threshold of Hearing by Air Conduction as a Function of Age and Sex for Otologically Normal Persons.* ISO 7029: Global Engineering Documents.

Killion MC, Niquette PA. (2000). What can the pure-tone audiogram tell us about a patient's SNR loss? *Hear J* 53(3):46–53.

Killion MC, Villchur E. (1993). Kessler was right—partly: but SIN test shows some aids improve hearing in noise. *Hear J* 46(9):31–5.

Studebaker GA. (1985). A "rationalized" arcsine transform. *J Speech Hear Res* 28:255–262.

Zurek PM, Delhorne LA. (1987). Consonant reception in noise by listeners with mild and moderate sensorineural hearing impairment. *J Acoust Soc Am* 82:1548–1559.