THE VALIDATION OF THE CID EVERYDAY SENTENCE TEST FOR USE WITH THE SEVERELY HEARING IMPAIRED

Donald G. Sims, Ph.D.
Coordinator, Audiology Service Section
NTID Communication Center

During the last several years, sentence length test materials have been reexamined by a number of authors as a means of improving the face validity of auditory speech discrimination testing. Sentence length test materials more closely duplicate real-life communication in that connected speech is used (Giolas, 1973, 1966, Davis and Silverman, 1970).

Silverman and Hirsh, 1955, and the Committee on Hearing and Bio-acoustics (CHABA) proposed the CID Everyday Sentence test as an instrument of clinical utility, and in 1972 the National Technical Institute for the Deaf (NTID) elected to utilize these test materials as the basis for a communication profile system (Johnson, 1974). It was hoped that the test could be used to assess skill levels of NTID students for several receptive and expressive communication parameters (speech discrimination, speechreading, manual reception, simultaneous reception, etc.) which were of interest.

Audio-cassette tape recordings were produced for all ten forms of the CID Everyday Sentence (hereafter referred to as the CHABA test). All ten lists were recorded by a male speaker who was chosen as a result of a study of the voices of seven male speakers for comparative intelligibility and a survey of subjects' preferences (Johnson, Subtelny, Andrews, Unpublished NTID Report 1972).

Several questions regarding the validity and usefulness of the CHABA Sentence test were examined in this study. First, was there a difference in the percent correct score as established for these sentence length test materials as compared to the percent correct discrimination score as established by the more commonly used CID W-22, PB word test? Establishing this average difference between the CID W-22 test and the CHABA test would allow comparison of results between clinics utilizing the different tests. In addition, establishing a relatively good correlation with the W-22 test would provide a measure of concurrent validity. That is, it could be assumed that if scores were correlated well, these two tests were measuring the same thing; namely, auditory speech discrimination ability.

Secondly, the relationship between the CHABA test and other frequently used tests of the auditory speech discrimination ability of the deaf was not known. Consequently, the PB-K (Davis and Silverman, 1970) and the Pickett Rhyme Test (Pickett, et al, 1969) were also recorded by the same speaker and studied.

Finally, it was not known whether the NTID recordings of the CHABA test were internally consistent over all ten lists of the test. Obviously, if List 1 always yielded much higher scores than List 2, then some adjustment or normalization of the scores should be made in order to obtain valid results from test to retest such as would be necessary when measuring pre- and post-training achievement in auditory, speech discrimination skills.

METHOD. Thirty-one students enrolled in Auditory Training at NTID were utilized in this study. Auditory speech discrimination ability for these subjects as measured by the CHABA test was distributed as shown in a typical histogram for CHABA test list #7 (Figure 1). The subjects were young adults approximately equally divided between sexes and had an average hearing loss of 85 dB; Range, 63 to 115+ dB HTL, ANSI.

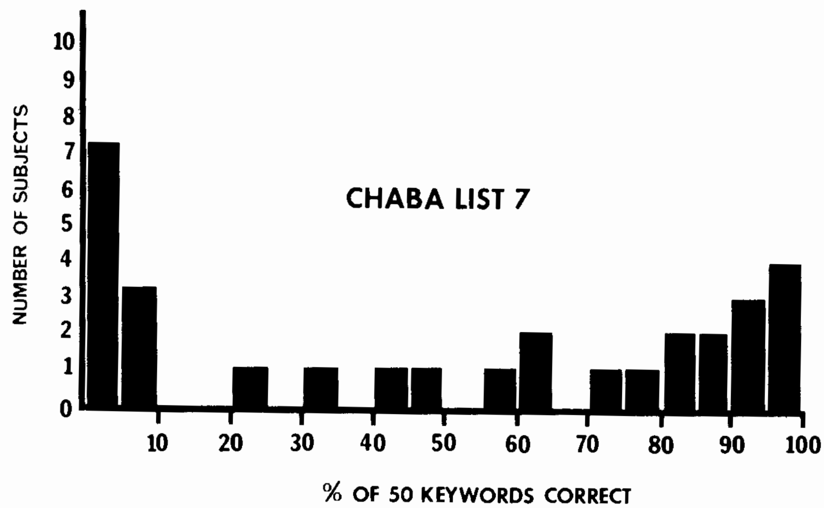

Figure 1. Hearing (Speech) Discrimination ability of NTID students as measured during administration of List 7 of the CID Everyday Sentence Tests (N = 31)

Subjects listened to a randomized ordering of all ten CHABA lists, CID W-22 lists 1A, 2A, 3A and 4A, the three lists of the PB-K words and two lists (A and B) of the Pickett Modified Rhyme Test. The tests were administered at a rate of two per week over a ten-week period with one or two tests being presented in a given test session of one half hour length. Randomization of the presentation of test lists was accomplished by testing students in groups of eight and presenting to each group a different ordering of all 19 test lists. Ordering of presentations for each of the groups was examined and counterbalancing was used in order to avoid learning and practice effects. A few students, however, missed some test sessions and were required to complete all missing tests in a make-up test session. Therefore, both long-range and short-range test effects were included in this data. Fourteen subjects completed *all* 19 test lists.

Subjects were seated at an auditory training table with hardwire amplification. This amplification system allowed students to adjust the intensity for loudness of the monaural signal directed to both ears. Prior to testing, each student was required to establish a most comfortable listening level (MCL) by adjusting his volume controls. Subjects adjusted their MCL by listening to randomly selected stimulus word presentations of that type utilized in the test which was about to be administered. Thus, sentence length material was used to establish MCL when the CHABA sentences were to be tested, and conversely, PB words were used to establish MCL for testing W-22 words. Test stimuli were presented via cassette tape recorder (Wollensak Model 2550) played through a JL Warren amplifier Model D-1S/S with Model TDH 140 earphones mounted in JL Warren circumaural cushions. The speaker recorded all tests monitoring the vowel peaks at 0 dB VU.

Subjects wrote down all their responses except on the Modified Rhyme Test where the test consisted of a five-alternative, forced-choice task. In that case, the subject's response choice was circled. Responses for CHABA test stimuli were scored by the percentage of 50 key words correct. Except where spelling mistakes were obvious, all responses were scored for exact, phonetic accuracy.

RESULTS AND DISCUSSION. First each of the test lists were correlated with each other to determine inter-test consistency. The Pearson Product Moment Correlation was used for this statistic. Table 1 presents the results of CHABA test administration.

Several conclusions can be drawn from the interpretation of this table. First, the inter-list correlation for the ten CHABA test lists was

TABLE 1: Interlist correlations for ten CHABA tests with missing data (N= 19-29)

List	1	2	3	4	5	6	7	8	9	10
1	1.00									
2	.87	1.00								
3	.92	.98	1.00							
4	.93	.95	.97	1.00						
5	.89	.95	.96	.95	1.00					
6	.88	.93	.95	.97	.96	1.00				
7	.88	.93	.97	.93	.96	.93	1.00			
8	.90	.95	.97	.97	.96	.96	.94	1.00		
9	.84	.89	.91	.97	.96	.96	.93	.95	1.00	
10	.86	.94	.95	.98	.98	.95	.96	.98	.97	1.00

high. The lowest correlation which resulted between List 1 and List 9 was .84; the highest correlation was .98. In general, the correlations were in the high 90's. It should be noted that some missing data may have been responsible for the .84 correlation between CHABA list 1 and list 9. Because, when the fourteen subjects who completed all ten lists were correlated using the Spearman Rank Order procedure, there was a .90 correlation for these two lists.

Table 2 shows that the W-22 test lists were correlated well with an average correlation of .89. The PB-K lists were correlated with an average of .83. The Pickett Rhyme Test lists were correlated at a level of .89.

TABLE 2: Interlist correlations for the W-22, PB-K and Rhyme tests with missing data (N = 18-29)

CID W-22		PB-K		Rhyme Test	
List	Correlation	List	Correlation	List	Correlation
1A x 2A	.83	1 x 2	.79	A x B	.89
1A x 3A	.90	1 x 3	.83		
1A x 4A	.92	2 x 3	.89		
2A x 3A	.88				
2A x 4A	.93				
3A x 4A	.93				

The estimation of concurrent validity between the CID W-22 Test and the CHABA test is depicted in Table 3. For this table, all the correlations between each list of the CHABA with list 1A of the W-22 were averaged, i.e., the correlation of W-22 list 1A with CHABA list 1, plus the correlation of W-22 list 1A with CHABA list 2, etc.

TABLE 3: The averaged correlation for 10 CHABA test lists with each of the W-22 lists.

List	Average Correlation
1A	.70
2A	.78
3A	.84
4A	.85
	Mean = .79

In spite of the different nature of these two tests in terms of stimulus materials, the overall average correlation of .79 appeared to establish good concurrent validity. That is, the CHABA test was apparently ordering subjects' auditory speech discrimination ability in roughly the same way as was the CID, W-22 test. For the CHABA test, however, there was a much greater range of scores as is demonstrated in Table 4.

As one may see from this data, the CHABA test scores averaged about 28% greater than the W-22 PB scores.

TABLE 4: Means and standard deviations of each of the CHABA, W-22 test lists, PB-K, and Rhyme lists.

Test	Mean	Standard Deviation
CHABA		
1	44.59	36.39
2	50.36	39.46
3	52.88	40.63
4	44.14	36.86
5	50.41	38.47
6	46.62	38.63
7	48.96	38.60
8	52.22	36.63
9	46.07	38.88
10	39.90	33.94
	Mean 47.61	
W-22		
1A	16.27	17.42
2A	17.73	20.01
3A	17.18	17.09
4A	25.18	21.09
	Mean 19.90	
PB-K		
1	18.3	18.4
2	11.6	14.7
3	14.4	18.0
	Mean 14.8	
Rhyme		
A	47.1	20.7
B	58.8	21.7
	Mean 51.4	

About the same mean differences were seen for the PB-K lists. The PB-K lists were originally developed with the idea that word-familiarity might eliminate language problems encountered by hearing-impaired children. This data, however, did not confirm the predicted effect on test scores with this population.

Pickett Rhyme Test results resembled the CHABA test in terms of mean scores. However, the standard deviation was much smaller.

The distribution of student scores with the 4 tests was as follows: The CHABA test had a bimodal distribution as seen in Figure 1. The W-22 test results (Figure 2) and the PB-K test results were negatively skewed (Figure 3). Because most NTID students scored very low on the PB-K and W-22 test lists, it does not seem desirable to use these tests when examining individual hearing discrimination ability within this population.

The Picket Modified Rhyme Test showed an even distribution of scores (Figure 4). Statistically, significant differences were found among several of the means of the CHABA lists using the t test for correlated samples. These differences between test list means ranged from .2% to 12.3% (Table 4).

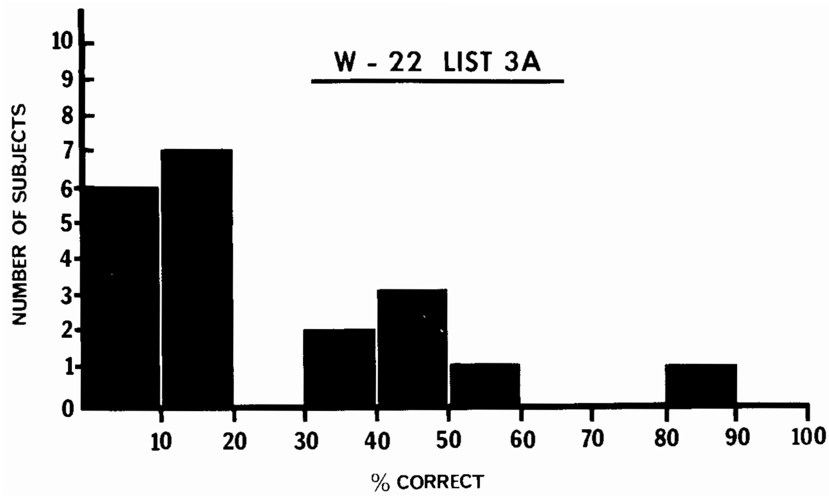

Figure 2. Distribution of Scores for administration of List 3A, W-22
(N = 20)

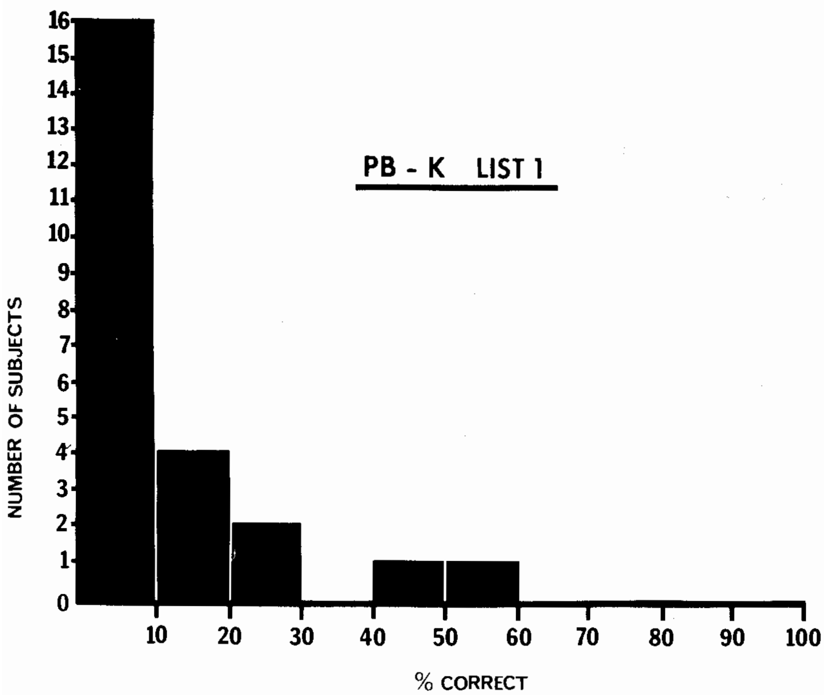

Figure 3. Distribution of Scores for administration of List 1, PB-K
(N = 24)

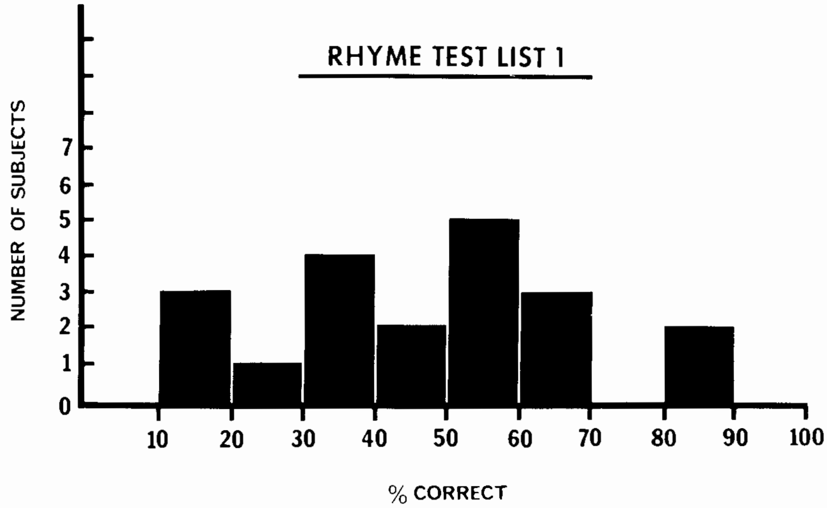

Figure 4. Distribution of Scores for administration of List 1, Rhyme Test (N = 29)

One may conclude that students were ranked in the same order from test list to test list within the CHABA test, i.e., the test lists were highly correlated. However, students' mean scores differed somewhat from list to list. As a consequence, some statistical manipulation was necessary to "normalize and equate" the mean scores of the various test lists. A Z-statistic transformation was used to provide this correction. The computation for this correction factor was as follows:

1. For all scores on all lists a Z score was obtained

$$Z = \frac{\text{Raw score - mean}}{\text{S. D.}}$$

2. The lowest and highest Z scores for all 10 lists were set equal to 0% and 100% (Figure 5).

Then all other Z scores were plotted on this 2 point function and related to a percent score from 0 to 100%. A sample of the correction table is seen in Table 5.

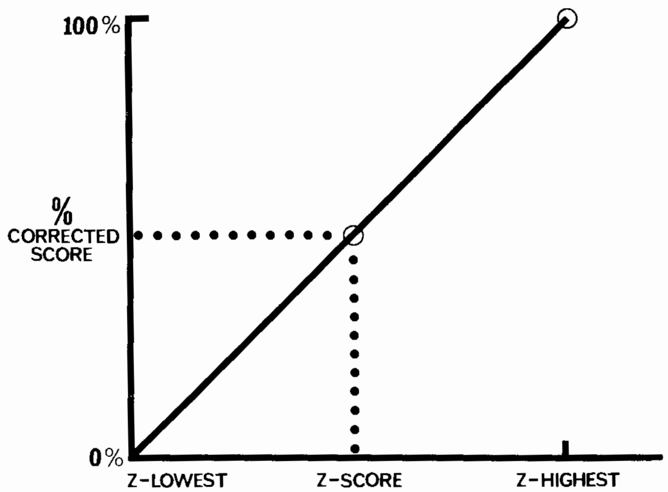

Figure 5. Two point function to correct for intertest list differences using the Z statistic

TABLE 5: CHABA list number 2 correction table for raw scores of from 51-100%.

Raw Score %	Corrected Score %	Raw Score %	Corrected Score %
51	48	76	68
52	49	77	69
53	50	78	70
54	51	79	71
55	51	80	71
56	52	81	72
57	53	82	73
58	54	83	74
59	55	84	75
60	55	85	75
61	56	86	76
62	57	87	77
63	58	88	78
64	59	89	79
65	59	90	80
66	60	91	80
67	61	92	81
68	62	93	82
69	63	94	83
70	63	95	84
71	64	96	84
72	65	97	85
73	66	98	86
74	67	99	87
75	67	100	Corrected Score is between 88 and 100

To facilitate scoring when group testing is done, results were also scored by computer and the normalization formula was applied to the raw scores to yield the corrected score automatically.

Finally, split-half reliability of the CHABA test results was analyzed with the Pearson Product Moment Correlation in order to determine if a half list could be administered with valid results. This was necessary in order to conserve time and to preserve the number of test lists available. Harris, 1970 indicated that the memory factor is quite high with these Everyday Sentences, and therefore, once they are used, they cannot be re-used with the same subject. The results can be seen in Table 6. The data indicates that the split-half correlations are extremely high. Also, the correlations are high for the first half of the test and the total score and for the second half of the test and the total score.

TABLE 6: CHABA Test Split-Half Correlation.

CHABA List	Split Half	1st Half W/Total	2nd Half W/Total
1	.96	.98	.98
2	.95	.98	.99
3	.96	.99	.99
4	.92	.98	.98
5	.90	.98	.97
6	.87	.98	.95
7	.94	.99	.98
8	.94	.98	.99
9	.97	.99	.99
10	.91	.98	.98

Test list 6 was somewhat poorer in split-half reliability but examination of the list, it was found that the first and second halves of the list were not equated in terms of sentence length, therefore a rearrangement of the sentences to balance length between the first and second halves would probably improve split-half reliability. However, on balance, it appears to be appropriate, when necessary, to utilize the CHABA Test Lists administered in split-half form.

SUMMARY. The CHABA CID Everyday Sentences Test was administered to 29 subjects over 20 experimental test sessions. Lists 1, 2, 3, and 4A of the CID W-22 Test, the 3 lists of the PB-K Test and the 2 lists of Pickett Rhyme Test were also administered. Subjects listened to the same speaker present all test stimuli at a most comfortable listening level in both ears.

Results were analyzed in terms of interlist consistency within each test, and the differences in mean scores between each of the tests. Also, a normalizing Z score transformation was developed to equate the means of CHABA test lists. Finally, the CHABA test results were studied for split-half reliability.

CONCLUSIONS. The CHABA (CID Everyday Sentence) test of auditory speech discrimination appeared to have concurrent validity with the CID-W-22 PB word test as evidenced by the correlations obtained between each of the tests. The means and standard deviations were different for the same subjects with raw scores differing by approximately 28%. However, apparently the subjects were rank-ordered similarly on the two tests—hence the good correlations.

Within the CHABA test, some of lists were found to be statistically different; thus, for this population it was necessary to use the Z-score normalization statistic to correct for inter-list differences.

A study of the split-half results indicated that good results could be obtained using half-lists.

Finally, the data reported herein with the CID Sentences do not resemble those obtained by Giolas on normal listeners with the sentences in noise. The difference was basically in the lower interlist correlations associated with his data. This result was puzzling in view of his good experimental design and the results which showed no mean differences between lists and tighter SD's. It is possible that the bimodal distribution evidenced in this population of severely hearing impaired subjects resulted in higher inter-test correlations.

REFERENCES

Davis, H., and S. R. Silverman, *Hearing and Deafness,* Revised 3rd Edition, (New York, N. Y.: Holt, Rinehart and Winston, Inc., 1970).

Ferguson, G. A., *Statistical Analysis in Psychology and Education,* ch. II (New York, N. Y.: McGraw-Hill), 1966.

Giolas, T. G., "Comparative Intelligibility Scores of Sentence Tests and Continuous Discourse." *J. Aud. Res.,* 6 p. 31, 1966.

Giolas, T. G., and J. R., Duffy, "Equavalency of CID and Revised CID Sentence Lists". *J. of Speech and Hearing Disorders,* 16, 4, 549, 1973.

Johnson, D. D., "Communication Characteristics of NTID Students" Proceedings ARA Convention, July, 1974.

Pickett, J. M., Martin, E. and Johnson, D., et al., "On Patterns of Speech Feature Reception by Deaf Listeners," Graduate School, Gallaudet College, Washington, D. C. (1969).

Silverman, S. F., and I. J. Hirsh, "Problems related to the use of speech in clinical audiometry." *Annals Otology, Rhin., & Laryng.,* 64, 1234, 1955.