JASA ARTICLE

Evaluation of binomial distribution estimates of confidence intervals of speech-recognition test scores[a]

Robert H. Margolis[1,b] and Richard H. Wilson[2]

[1]*Audiology Incorporated, Arden Hills, Minnesota 55112, USA*

[2]*Speech and Hearing Science Program, Arizona State University, Tempe, Arizona 85287, USA*

ABSTRACT:
Speech-recognition tests are a routine component of the clinical hearing evaluation. The most common type of test uses recorded monosyllabic words presented in quiet. The interpretation of test scores relies on an understanding of the variance of repeated tests. Confidence intervals are useful for determining if two scores are significantly different or if the difference is due to the variability of test scores. Because the response to each test item is binary, either correct or incorrect, the binomial distribution has been used to estimate confidence intervals. This method requires that test scores be independent. If the scores are not independent, the binomial distribution will not accurately estimate the variance of repeated scores. A previously published dataset with repeated scores from normal-hearing and hearing-impaired listeners was used to derive confidence intervals from actual test scores in contrast to the predicted confidence intervals in earlier reports. This analysis indicates that confidence intervals predicted by the binomial distribution substantially overestimate the variance of repeated scores resulting in erroneously broad confidence intervals. High correlations were found for repeated scores, indicating that scores are not independent. The interdependence of repeated scores invalidates confidence intervals predicted by the binomial distribution. Confidence intervals and confidence levels for repeated measures were determined empirically from measured test scores to assist in interpreting differences between repeat scores.
https://doi.org/10.1121/10.0013826

(Received 8 February 2022; revised 2 August 2022; accepted 10 August 2022; published online 6 September 2022)

[Editor: Richard A. Wright] Pages: 1404–1415

I. INTRODUCTION

Word-recognition testing is routinely included in the diagnostic hearing evaluation to assess the patient's speech communication ability. There are many methods and tests that have been investigated and used in the clinic, and there is little consensus on the choice of methods. Tests that employ monosyllabic words in quiet are the most frequently used. Although most word lists were constructed with 50 items (Egan, 1948; Hirsh *et al.*, 1952), it is common to use 25-item lists for time savings (Elpern, 1961; Martin and Sides, 1985; Martin and Morris, 1989; Martin *et al.*, 1994; Wiley *et al.*, 1995). The confidence intervals around percent-correct test scores for 25- and 50-word tests are important to the clinician for interpreting differences between scores obtained on subsequent visits and in different listening conditions, such as tests performed with and without amplification or before and after treatment.

Hagerman (1976) derived confidence intervals for 25- and 50-word tests based on the binomial distribution. This approach was employed in subsequent studies by Thornton and Raffin (1978), Raffin and Thornton (1980), Raffin and Schafer (1980), and Carney and Schlauch (2007). Thornton and Raffin (1978) and Raffin and Schafer (1980) tested the confidence intervals against repeated measures. Their subject sample, however, was probably skewed toward high scores (more on this in Sec. IV). Many audiology textbooks have suggested that confidence intervals predicted by the binomial distribution are helpful for interpreting differences between test scores (Bess, 1983; Penrod, 1994; Thibodeau, 2000; Gelfand, 2018; Kramer and Brown, 2019; Martin and Clark, 2019).

Thornton and Raffin (1978, p. 508) point out that "If the responses to test stimuli are assumed to be independent of each other, then test results can be treated as binomial distributions and the statistics of proportions can be used to describe their characteristics." The assumption of independence is critical. If responses are not independent, then confidence intervals based on the binomial distribution will not accurately estimate the variance of repeated measures. None of the reports that derived confidence intervals with the binomial distribution examined the independence of speech-recognition test results.

Two principles of independence of data are particularly relevant to speech-recognition testing. First, repeated measures from the same individual are not independent (Cohen *et al.*, 2003; Witte and Witte, 2007; Bijma *et al.*, 2017). Second, test scores that are highly correlated are not independent (Bertsekas and Tsitsiklis, 2008).

In this report, the independence of word-recognition scores is critically examined from these two perspectives.

[a]This paper is part of a special issue on Reconsidering Classic Ideas in Speech Communication.

[b]Electronic mail: rhmargo001@gmail.com

On the individual response level, we consider the interdependence of multiple responses from the same subject. On the test score level, where test and retest scores are highly correlated, we discuss the implication of the high correlations for the independence of test scores and provide statistical evidence that correlated scores are not independent. These considerations lead to the conclusion that repeated word-recognition scores violate the independence assumption of the binomial distribution resulting in erroneous predictions of confidence intervals.

Because there has not been a rigorous comparison of confidence intervals predicted by the binomial distribution and variability of actual test scores, this investigation was undertaken to determine the variance and confidence intervals of repeated speech-recognition scores. The dependence of repeated test scores results in substantially lower variance and narrower confidence intervals compared to predictions derived with the binomial distribution.

II. METHODS

A. Word lists

Margolis *et al.* (2021) reported the development of automated, forced-choice word-recognition tests. The results from the automated methods were compared to conventional open-set word-recognition testing with responses scored manually by the tester and closed-set forced-choice scores obtained by computer scoring. The study produced a dataset with repeated measures of monosyllabic-word tests over a wide range of presentation levels for normal-hearing and hearing-impaired listeners. These open- and closed-set data were employed in the current study to derive confidence intervals and test-retest correlations for repeated measures.

The test materials were Northwestern University Auditory Test No. 6 (NU-6; Tillman and Carhart, 1966), spoken by a female talker (Causey *et al.*, 1983; Department of Veterans Affairs, 2006). One hundred words were organized into 4 equivalent 25-word lists using the item-difficulty data reported by Wilson and McArdle (2015). (See Margolis *et al.*, 2021 for details.) All 100 words were presented to 10 normal-hearing listeners at each of the 6 levels, ranging from 11 to 41 dB re pure-tone average (500, 1000, and 2000 Hz; ANSI, 2018) and 16 listeners with sensorineural hearing loss at each of the 5 levels ranging from 22 to 46 dB re pure-tone average. The word order was randomized at each level and responses were parsed to produce scores for each 25-word list. For the analysis of 50-word scores, responses to the words in the first and second 25-word lists and the words in the third and fourth lists were combined. This protocol produced 4 scores at each presentation level for 25-word lists and 2 scores at each level for 50-word lists.

B. Confidence intervals

Confidence intervals were calculated for the four 25-word scores of each of the 26 listeners at each presentation level with the following formulas:

$$\text{Upper limit}\,(97.5\text{ percentile}) = \bar{x} + \left(1.96 * \sigma/\sqrt{n}\right), \quad (1)$$

$$\text{Lower limit}\,(2.5\text{ percentile}) = \bar{x} - \left(1.96 * \sigma/\sqrt{n}\right), \quad (2)$$

where $\bar{x}$ is the mean of the four scores, σ are the standard deviations of the four scores; and n is the number of scores (four).

These formulas assume that the data are normally distributed. An analysis was performed to determine if the repeated scores are accurately fit with normal distributions. The analysis revealed that there is no significant difference between the sets of repeated scores and best-fit normal distributions for scores ranging from 28% to 72%. This analysis is summarized in the Appendix.

The upper limit values and lower limit values calculated with Eqs. (1) and (2), respectively, were plotted and fit with best-fit second-degree polynomials. Paradoxically, the best-fit functions did not capture 95% of the scores. This is due to the fact that the best-fit functions reflect averages of the upper and lower limits of the four data points from each subject so that, by definition, 50% of the points are above the average and 50% are below the average. These functions were adjusted by adding constants to the third (constant) term of the polynomial equations to capture 95%, 90%, and 80% of the cases. Because the confidence intervals were determined empirically and not calculated by any mathematical model, they capture exactly the indicated proportion of cases.

To determine the confidence intervals for 50-word scores, all scores were plotted with the best-fit polynomials from the 25-word scores. The polynomial equations were then adjusted as described above to capture 95%, 90%, and 80% of the cases.

C. Confidence levels

The following procedure was used to determine the confidence levels associated with differences between an initial score and a second score for 25- and 50-word lists.

Step 1. For each listener, all of the list scores were arranged in groups based on the mean of the 4 scores (25-word lists) or 2 scores (50-word lists).

Step 2. The scores were grouped into categories corresponding to 10% ranges based on the mean scores. For 25-word lists, these categories were 6%–15%, 16%–25%,…, 86%–95%. Scores below 6% and above 95% were not included in this analysis because there were too few scores to determine the confidence levels.

Step 3. For each category, the following percentiles were calculated using Microsoft Excel (Build 15128.20224; Microsoft Corporation, Redmond, WA): 5, 10, 20, 30, 40, 50, 60, 70, 80, 90, 95.

Step 4. The percentile scores in each category were fit with second degree polynomials using an online polynomial regression tool.[1]

Step 5. For each percentile, the best-fit second degree polynomial was solved to produce percentile scores for each possible mean score (4%–100% in 4% increments for 25-word list scores and 2%–100% for 50-word list scores).

Step 6. Each score calculated from the best-fit polynomials was rounded to the nearest 4% (25-word list scores) or 2% (50-word list scores).

This procedure produced Tables I–IV, which provide a level of confidence that a second score is significantly higher or lower than a first score with three confidence levels (80%, 90%, and 95%).

III. RESULTS

A. 25-Word open-set scores

(1) Inter-list correlations. Bivariate plots of 25-word open-set scores for each pair of lists are shown in Fig. 1. Each panel shows 140 pairs of scores (10 normal listeners at 6 levels and 16 hearing-impaired listeners at 5 levels). The results indicate that list scores are highly repeatable with correlation coefficients that range from 0.88 to 0.91.

(2) Confidence intervals. Scores for all of the participants and all lists are plotted in Fig. 2 against the mean of the four scores. Individual data points include 240 scores for normal listeners and 320 scores for hearing-impaired listeners. Three sets of confidence intervals are shown (80%, 90%, and 95%) along with the Thornton and Raffin (1978) confidence intervals. It is evident that the variance of the data is substantially overestimated by the Thornton and Raffin confidence intervals. The 95% confidence intervals from the current study are provided in Table I.

(3) Standard deviations. Carney and Schlauch (2007) provide the following formula for estimating standard deviation from the binomial distribution:

$$\sigma = \sqrt{(p \times (1-p)/n)}, \tag{3}$$

where p is the word-recognition score expressed as a proportion and n is the number of test items. The standard deviations calculated by this formula and average standard deviations for the 4 test scores for each of the 26 listeners are shown in Fig. 3. Standard deviations are greatest for scores near 50% and smaller where the distribution is compressed near 0% and 100%. This dependence of the standard deviation on test score was first noted by Egan (1948) and is predicted by the binomial distribution (Hagerman, 1976; Thornton and Raffin, 1978; Carney and Schlauch, 2007). Like the confidence intervals predicted by the binomial distribution shown in Fig. 2, the predicted standard deviations substantially overestimate the variance of measured data.

B. 50-Word open-set scores

(1) Inter-list correlations. Bivariate plots of scores for the 50-word lists are shown in Fig. 4. Individual points are 140 pairs of scores (10 normal listeners at 6 levels and 16 hearing-impaired listeners at 5 levels). The results indicate that the list scores are highly repeatable with a correlation coefficient of 0.97.

TABLE I. The 95% confidence intervals for 25-word list scores and 50-word list scores. LL, lower limit (2.5 percentile); UL, upper limit (97.5 percentile).

	25-Word lists		50-Word lists	
Score	LL	UL	LL	UL
2			0	5
4	0	14	1	7
6			2	10
8	0	18	4	12
10			5	15
12	1	23	7	17
14			9	19
16	4	28	10	22
18			12	24
20	8	32	14	26
22			15	29
24	11	37	17	31
26			19	33
28	15	41	21	35
30			22	38
32	18	46	24	40
34			26	42
36	22	50	28	44
38			30	46
40	26	54	32	48
42			33	51
44	30	58	35	53
46			37	55
48	34	62	39	57
50			41	59
52	36	64	43	61
54			45	63
56	38	66	47	65
58			49	67
60	42	70	51	69
62			54	70
64	46	74	56	72
66			58	74
68	50	78	60	76
70			62	78
72	59	85	64	80
74			67	81
76	63	89	69	83
78			71	85
80	68	92	73	87
82			76	88
84	72	96	78	90
86			80	92
88	77	99	83	93
90			85	95
92	82	100	87	97
94			90	98
96	86	100	92	100
98			95	100
100	91	100	97	100

(2) Confidence intervals. Confidence intervals (80%, 90%, and 95%) for 50-word test scores are shown in Fig. 5. Individual data points are 120 scores for normal listeners and 160 scores for hearing-impaired listeners.

https://doi.org/10.1121/10.0013826

TABLE II. The confidence levels for pairs of 25-word list scores. To determine if a second score is significantly different from a first score, enter the table from the leftmost column at the value of the first score. Find the second score in the corresponding row. Three confidence levels are shown (95%, high; 90%, moderate; 80%, low). If the second score is in the "no difference" range, it is interpreted as not significantly different from the first score.

25-Word list scores							
	Second score lower than first				Second score higher than first		
	Confidence %				Confidence %		
1st score	95	90	80	No difference	80	90	95
4				0–4	8	12	≥16
8			0	4–8	12	16	≥20
12		0	4	8–12	16	20	≥24
16	0	4	8	12–20	24	28	≥28
20	≤4	8	12	16–24	28	32	≥36
24	≤8	12	16	20–28	32	32	≥36
28	≤12	16	20	24–32	36	40	≥44
32	≤16	20	24	28–36	40	44	≥48
36	≤20	24	28	32–40	44	48	≥52
40	≤24	28	32	36–44	48	52	≥56
44	≤28	32	36	40–48	52	56	≥60
48	≤32	36	40	44–52	56	60	≥64
52	≤36	40	44	48–56	60	64	≥68
56	≤40	44	48	52–60	64	68	≥72
60	≤44	48	52	56–64	68	72	≥76
64	≤48	52	56	60–68	72	76	≥80
68	≤52	56	60	64–72	76	80	≥84
72	≤56	60	64	68–76	80	84	≥88
76	≤60	64	68	72–80	84	88	≥92
80	≤68	72	76	80–84	88	92	≥96
84	≤72	76	80	84	88	92	≥96
88	≤76	80	84	88	92	96	100
92	≤80	84	88	92	96	100	
96	≤84	88	92	96	100		
Confidence	High	Mod	Low	No difference	Low	Mod	High

TABLE III. The confidence levels for pairs of 50-word list scores. To determine if a second score is significantly different from a first score, enter the table from the leftmost column at the value of the first score. Find the second score in the corresponding row. Three confidence levels are shown (95%, high; 90%, moderate; 80%, low). If the second score is in the "no difference" range, it is interpreted as not significantly different from the first score.

50-Word lists							
	Second score lower than first				Second score higher than first		
	Confidence %				Confidence %		
First score	95	90	80	No difference	80	90	95
2			0	2–6	8	10	≥12
4			0	2–8	10	12	≥14
6			2	4–10	12	14	≥16
8	0	2	4	6–12	14	16	≥18
10	0	2	6	8–14	16	18	≥20
12	≤2	4	8	10–16	18	20	≥22
14	≤4	6	10	12–18	20	22	≥24
16	≤6	8	12	14–20	22	24	≥26
18	≤8	10	14	16–22	24	26	≥28
20	≤10	12	16	18–24	26	28	≥30
22	≤12	14	18	20–26	28	30	≥32
24	≤14	16	20	22–28	28	32	≥34
26	≤16	18	22	22–30	32	34	≥36
28	≤18	20	24	26–32	34	36	≥38

TABLE III. (*Continued*)

	50-Word lists						
	Second score lower than first				Second score higher than first		
	Confidence %				Confidence %		
First score	95	90	80	No difference	80	90	95
30	≤20	22	26	28–34	36	38	≥40
32	≤22	24	28	30–36	38	40	≥42
34	≤24	26	30	32–36	38	42	≥44
36	≤26	28	32	34–38	40	44	≥46
38	≤28	30	34	36–40	42	46	≥48
40	≤30	32	36	38–42	44	48	≥50
42	≤32	34	38	40–44	46	50	≥52
44	≤34	36	40	42–46	48	52	≥54
46	≤36	38	42	44–48	50	54	≥56
48	≤38	40	44	46–50	52	56	≥58
50	≤40	42	46	48–52	54	58	≥60
52	≤42	44	48	50–54	56	60	≥62
54	≤44	46	50	52–56	58	62	≥64
56	≤46	48	52	54–58	60	64	≥66
58	≤48	50	54	56–60	62	66	≥68
60	≤50	52	56	58-62	64	68	≥70
62	≤52	54	58	60–64	66	70	≥72
64	≤54	56	60	62–66	68	72	≥74
66	≤56	58	62	64–68	70	74	≥76
68	≤58	60	64	66–70	72	76	≥78
70	≤62	64	66	68–72	74	78	≥80
72	≤64	66	68	70–74	76	80	≥82
74	≤66	68	70	72–76	78	82	≥84
76	≤68	70	72	74–78	80	84	≥86
78	≤70	72	74	76–80	82	86	≥88
80	≤72	74	76	78–82	84	88	≥90
82	≤74	76	80	82–84	86	90	≥92
84	≤78	80	82	84–86	88	92	≥94
86	≤80	82	84	86–90	90	94	≥96
88	≤82	84	86	88–90	92	96	≥98
90	≤84	86	88	90–92	94	98	100
92	≤86	88	90	92–94	96	98	
94	≤88	90	92	94–96	98	100	
96	≤92	94	94	96–98	100		
98	≤94	96	96	98–100			
100	≤96	98	100	100			
Confidence	High	Mod	Low	No difference	Low	Mod	High

Confidence intervals are narrower for 50-word scores than for 25-word scores. Similar to the 25-word open-set results, the Thornton and Raffin (1978) confidence intervals for 50-word scores substantially overestimate the variance of measured scores. The 95% confidence intervals are provided in Table I.

C. 25-Word closed-set scores

(1) Inter-list correlations. Bivariate plots of 25-word closed-set scores for each pair of lists are shown in Fig. 6. Each panel shows 140 pairs of scores (10 normal listeners at 6 levels and 16 hearing-impaired listeners at 5 levels). The results indicate that list scores are highly repeatable with correlation coefficients that range from 0.80 to 0.88. The lower correlation coefficients relative to the open-set data result from the narrower range of scores in the closed-set condition. As the range of the data is restricted, the correlation coefficient decreases (Holms *et al.*, 2021).

(2) Confidence intervals. Confidence intervals (80%, 90%, and 95%) determined for the closed-set data are shown in the top panel of Fig. 7. Individual data points are 240 scores for normal listeners and 320 scores for hearing-impaired listeners. Note that the range of the closed-set data is restricted because chance performance (25%) is the theoretical minimum score. The 95% confidence intervals for open- and closed-set conditions are shown in the bottom panel of Fig. 7. The open- and closed-set

https://doi.org/10.1121/10.0013826

TABLE IV. The correlation coefficients (r) for the 25-word open-set data in Fig. 1, the 50-word open-set data in Fig. 4, and the 25-word and the closed-set data in Fig. 7 are listed along with the r/error calculated with the equation, $\text{Error} = 0.67 \times ((1 - r^2)/n^{0.5})$. n is the number of subjects × the number of presentation levels.

Mod	Number of words	List	n	r	r^2	Error	r/Error
Open	25	1 vs 2	146	0.91	0.82	0.010	89.2
Open	25	1 vs 3	146	0.90	0.80	0.011	80.0
Open	25	1 vs 4	146	0.89	0.79	0.012	75.0
Open	25	2 vs 3	146	0.88	0.78	0.013	70.0
Open	25	2 vs 4	146	0.91	0.82	0.010	89.5
Open	25	3 vs 4	146	0.91	0.83	0.009	97.7
Open	50	1 + 2 vs 3 + 4	146	0.97	0.94	0.003	292.1
Closed	25	1 vs 2	146	0.88	0.77	0.013	67.3
Closed	25	1 vs 3	146	0.84	0.70	0.017	50.1
Closed	25	1 vs 4	146	0.86	0.75	0.014	60.2
Closed	25	2 vs 3	146	0.84	0.70	0.017	48.8
Closed	25	2 vs 4	146	0.83	0.68	0.018	46.0
Closed	25	3 vs 4	146	0.80	0.64	0.020	39.6

confidence intervals are nearly identical, indicating that the same confidence intervals can be applied to the interpretation of open- and closed-set scores.

D. Confidence levels

Tables II and III provide confidence levels to determine if a second score is significantly higher or lower than a first score. For each possible first score, second scores corresponding to 80%, 90%, and 95% confidence are provided, indicating the level of confidence that the second score is lower than or higher than the first score.

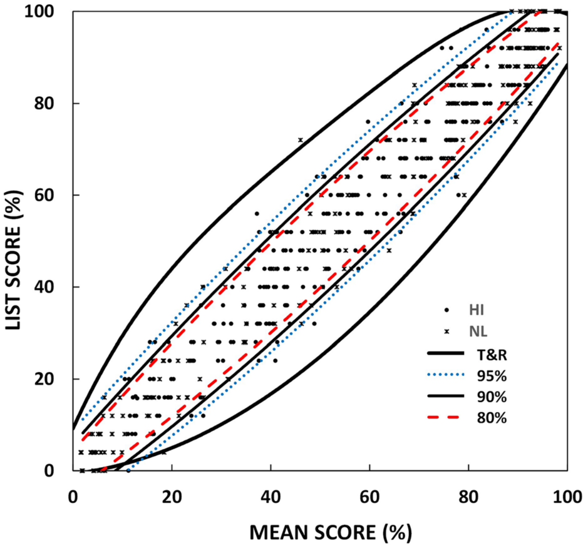

FIG. 2. (Color online) Individual word-recognition scores from 10 listeners with normal hearing (NL; circles) and 16 listeners with hearing impairment (HI; black circles) and the 80%, 90%, and 95% confidence intervals derived from the test scores and the binomial distribution (Thornton and Raffin, 1978; T&R, bold lines]). The data points for each subject are in sets of 4 scores from different 25-word lists. For each individual, the four scores are plotted against the mean of the four scores. A total of 560 scores is shown.

IV. DISCUSSION

A. Measured and predicted variances of word-recognition scores

The results shown in Figs. 2, 5, and 7 indicate that confidence intervals predicted by the binomial distribution

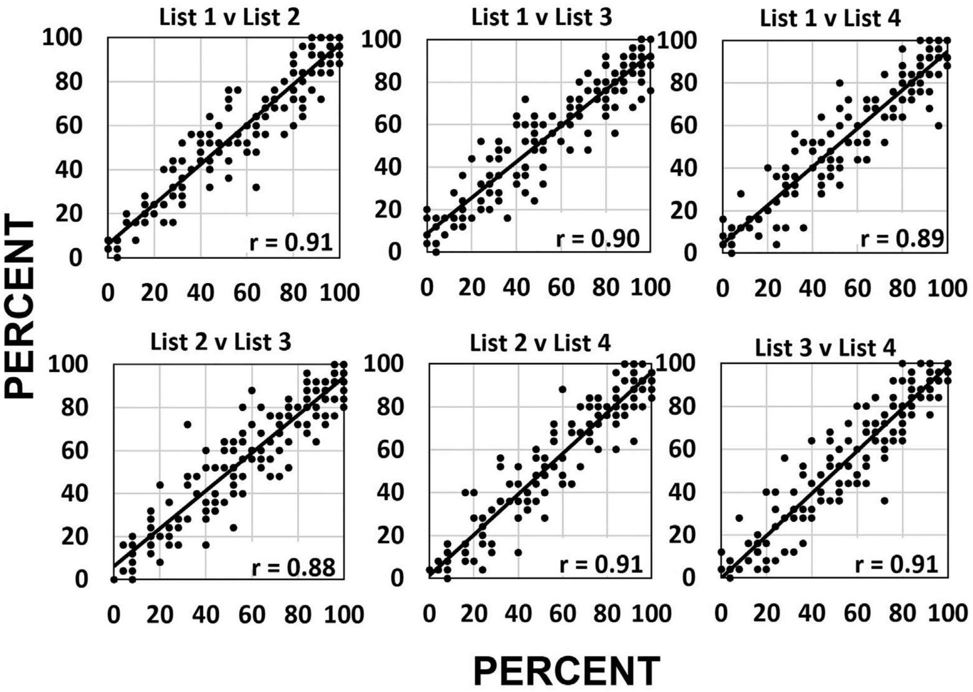

FIG. 1. The inter-list correlations of word-recognition scores are illustrated for 4, 25-word lists presented in a quiet, open-set paradigm to 10 listeners with normal hearing and 16 listeners with sensorineural hearing loss at 6 and 5 presentation levels, respectively. Each panel displays 140 pairs of scores. The materials were from the female recording of NU-6 (Department of Veterans Affairs, 2006).

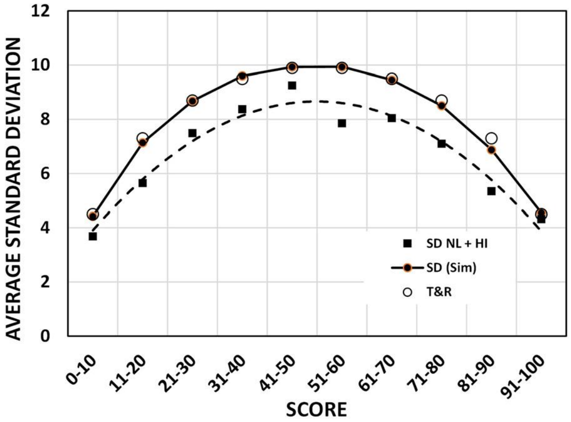

FIG. 3. (Color online) The average standard deviations of the four 25-word-list standard deviations by each of the 10 listeners with normal hearing and 16 listeners with hearing impairment predicted by the binomial distribution, SD (Sim, filled circles), calculated from each set of 4 individual scores from each of the 26 listeners (SD NL + HI, squares), and calculated from pairs of scores by Thornton and Raffin (1978; T&R, open circles). The average standard deviations are shown for 140 scores stratified by the mean of 4 repeat scores. The dashed line is the best-fit, second-degree polynomial for the SD NL + HL data.

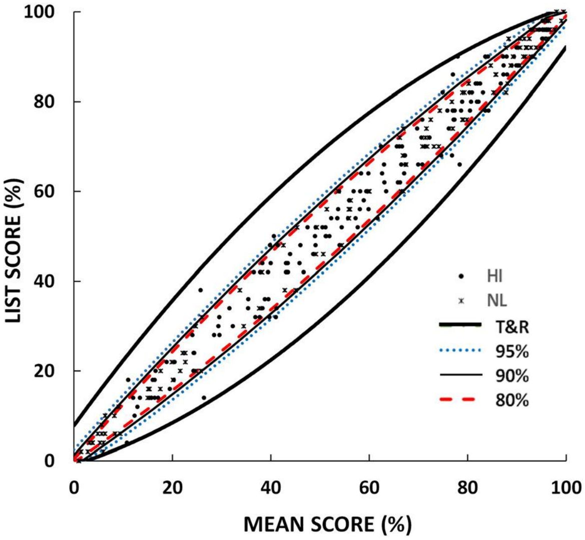

FIG. 5. (Color online) Individual word-recognition scores from 10 listeners with normal hearing (NL; ×) and 16 listeners with hearing impairment (HI; •) and confidence intervals derived from the test scores (95%, 90%, and 80%) and from the binomial distribution (Thornton and Raffin, 1978; T&R, bold lines). The data points for each subject are in sets of 2 scores from different 50-word lists. The individual scores are plotted against the mean of the two scores. A total of 280 scores is shown, jittered to reveal overlapping points.

substantially overestimate the variability of repeated word-recognition test scores. This overestimation is evident in Fig. 8, which shows the confidence interval range (upper limit minus lower limit) for a first score of 50% for 80%, 90%, and 95% confidence intervals from the present study, along with the 95% confidence intervals predicted by the binomial distribution in the reports of Hagerman (1976), Thornton and Raffin (1978), Raffin and Thornton (1980), and Carney and Schlauch (2007). Beattie *et al.* (1978), commenting on a prepublication version of the study by Thornton and Raffin (1978), noted that NU-6 word-recognition scores obtained with the Auditec recordings (male speaker; Auditec, Inc., St. Louis, MO) from 212 ears of 163 listeners were substantially less variable than the confidence intervals of Thornton and Raffin (1978) would predict. Raffin and Schafer (1980) questioned the variance reported by Beattie *et al.* (1978) based on possible methodological errors. The lower variance from Beattie *et al.* (1978) is consistent with the results of this study.

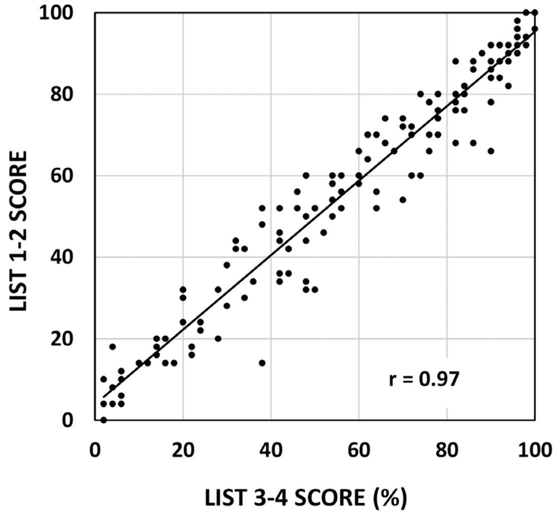

FIG. 4. The repeated test scores based on two 50-word lists. Pairs of 25-word lists (lists 1 and 2 and lists 3 and 4) were combined to form the two 50-word lists presented in quiet to 10 listeners with normal hearing and 16 listeners with sensorineural hearing loss at 6 and 5 presentation levels, respectively. The results for 140 pairs of scores are shown. The materials were from the female recording of NU-6 (Department of Veterans Affairs, 2006).

The closed-set data in Fig. 7 are included in this analysis for two reasons. First, because the variability predicted by the binomial distribution is identical for open- and closed-set test scores, the closed-set data offered another comparison of predicted and measured variability with a different data set. The binomial distribution predicts the variance of a test score when (1) responses to individual test items have two and only two possible values, e.g., *yes/no*, *true/false*, *correct/incorrect*, (2) the proportion of each response in a population is known, and (3) the number of test items is known (Thornton and Raffin, 1978, p. 509). Accordingly, the variances predicted for open- and closed-set scores are identical. Second, because the closed-set paradigm used in this study is in clinical use, we wished to evaluate differences in repeated scores when the materials are presented in a closed-set format.

Two important features of the confidence intervals in Fig. 8 deserve comment. First, confidence intervals derived from the binomial distribution are substantially broader than those obtained from repeated scores. Second, there are substantial differences in confidence intervals derived from the binomial distribution in the several studies cited despite the

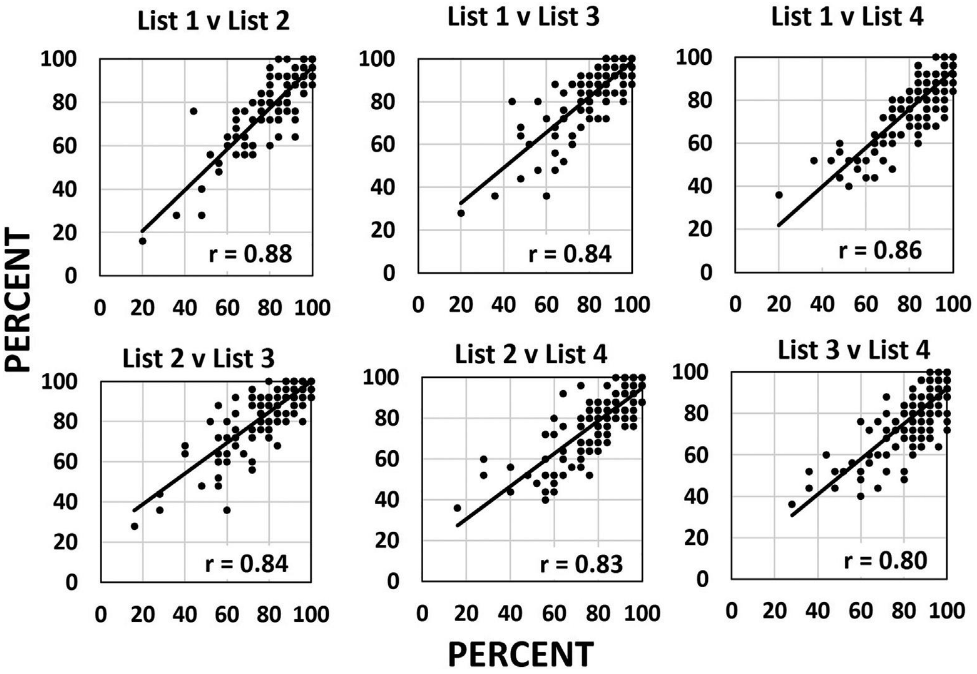

FIG. 6. The inter-list correlations of word-recognition scores based on four 25-word lists presented in quiet in a closed-set paradigm to 10 listeners with normal hearing and 16 listeners with sensorineural hearing loss at 6 and 5 presentation levels, respectively. Each panel displays 140 pairs of scores. The materials were from the female recording of NU-6 (Department of Veterans Affairs, 2006).

Raffin and Thornton (1980, p. 18) assertion that "Discrepancies from our previous tables are due to differences in the number of significant digits carried by the two computers." The distribution of scores estimated from the binomial distribution should only depend on the score and the number of test items (Thornton and Raffin, 1978, p. 509). The reasons for the differences among the reports are not clear.

There have been two previous attempts to validate the confidence intervals derived from the binomial distribution. Thornton and Raffin (1978) analyzed test scores from 4120 patients obtained from patient records in a U.S. Department of Veterans Affairs (VA) clinic. Fifty-word scores were divided into two 25-word scores to obtain repeated measures. They reported that the 90% and 95% confidence intervals predicted by the binomial distribution accurately identified the percentage of scores that fell outside of the confidence intervals (7.9% for the 90% confidence interval and 5.4% for the 95% confidence interval). The scores obtained from clinic records are usually highly skewed toward high scores because tests are typically conducted at high presentation levels where most patients score between 90% and 100%. This is evident in Fig. 9, which shows the distribution of 4150 scores from Thornton and Raffin (1978, Table 5) and a large clinical database (22 088 scores) analyzed by Margolis and Saly (2008). A large proportion of scores in both datasets are greater than 88% (58% of the sample in the data from Thornton and Raffin (1978) and 72% of the sample in the database from Margolis and Saly, 2008). For those scores, it is not possible to evaluate whether they fall above the upper limit of the confidence interval because of the ceiling of 100%. The distribution of scores does not permit an accurate assessment of the confidence intervals. Raffin and Schafer (1980) compared scores obtained from a clinic database and a cohort of volunteers. They did not report the actual scores but concluded that the 95% confidence intervals accurately captured 95% of the scores. It is likely that the scores that were analyzed are similar in their distribution characteristics to those shown in Fig. 9. There is probably a large proportion of scores above 88% for which the proportion that fall above the 95% confidence interval cannot be determined.

The measured and predicted standard deviations appearing in Fig. 3 represent additional evidence that the variance of repeated scores is overestimated by the binomial distribution. Dillon (1982) reported that standard deviations predicted by the binomial distribution were in close agreement with those reported by Thornton and Raffin (1978) for repeated measures of a 25-word test. The simulated standard deviations and those from Thornton and Raffin (1978), which are shown in Fig. 3, are consistent with the observation by Dillon (1982), and substantially larger than those observed in the present study. At least two factors could contribute to the differences between the standard deviations from Thornton and Raffin (1978) and those from this study. First, because standard deviations decrease with sample size (Holms *et al.*, 2021), the Thornton and Raffin standard deviations, based on two measures for each subject, may overestimate the true standard deviations. The measured standard deviations in Fig. 3 are based on four repeated measurements and could also overestimate the true standard deviation, which would increase the disparity between the

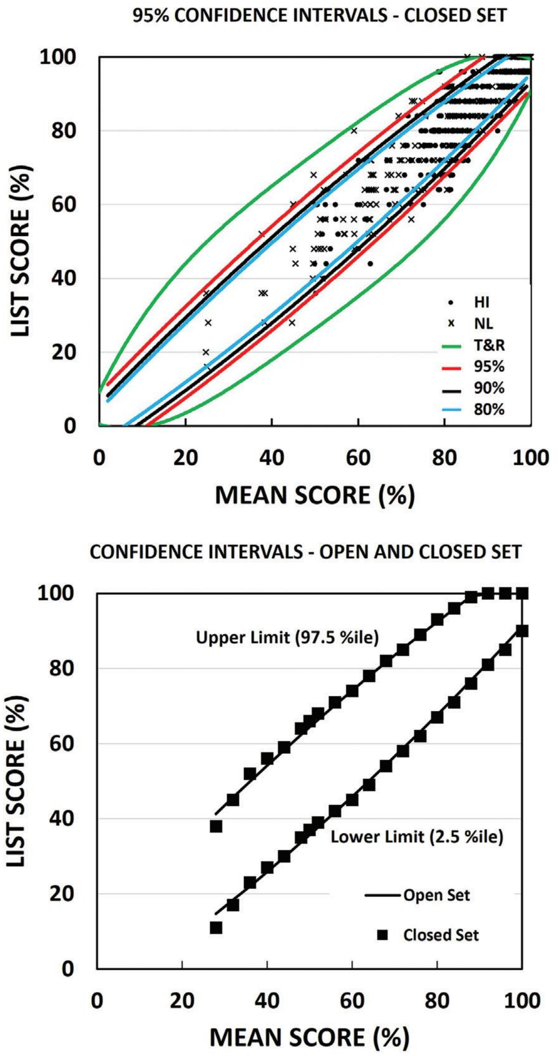

FIG. 7. (Top) Individual word-recognition scores from 10 listeners with normal hearing (NL; ×) and 16 listeners with hearing impairment (HI; •) in a closed-set paradigm with two 50-word lists along with derived 95%, 90%, and 80% confidence intervals. The individual scores are plotted against the mean of the two scores. For comparison the 95% confidence interval with the binomial distribution from Thornton and Raffin (1978; T&R, bold lines) is shown. A total of 560 scores is shown. (Bottom) The upper and lower limits of 95% confidence intervals from scores obtained with the 25-word open- and closed-set procedures. The solid lines are best-fit polynomials for the open-set upper limits (97.5 percentile) and lower limits (2.5 percentile). The squares are average closed-set upper limits (97.5 percentile) and lower limits (2.5 percentile).

measured and predicted variances. Second, the two scores obtained from each listener by Thornton and Raffin (1978) were acquired by determining a score from the first and second halves of a 50-word list. This does not produce lists of equal difficulty. The differences in list difficulty are cited by Dillon (1982) as a contributor to word-recognition score variability. The scores reported from this study were based on 25-word lists that were designed to be equivalent based on item difficulties from a large sample of normal and hearing-impaired listeners.

Elpern (1961, Table II, p. 34) examined differences in scores for 25-word lists created by dividing the 50-word CID W-22 (Central Institute for the Deaf, St. Louis, MO) recorded 50-word lists into first-half and second-half lists. The mean differences between first-half and second-half scores for eight 50-word lists ranged from 0% to 5% (mean = 1.4%). The differences between the 25-word list scores obtained from this study and the lists that were constructed to be equivalent based on item-difficulty data ranged from 0.4% to 1.8% (mean = 1.0%). This small difference may represent a small contribution to the greater differences obtained with split-half 25-word lists that are not controlled for average difficulty.

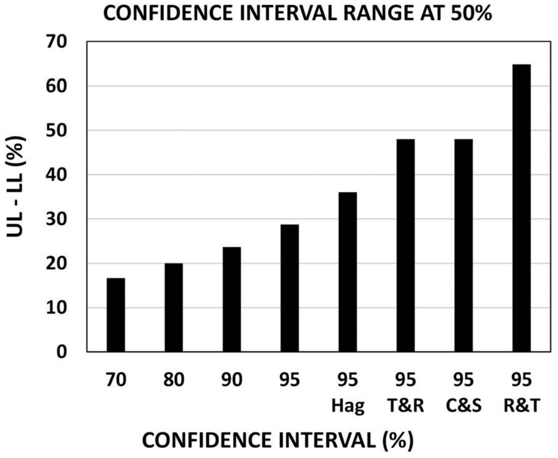

FIG. 8. The confidence interval ranges [upper limit (UL) minus lower limit (LL)] for 50% scores for (1) 70%, 80%, 90%, and 95% confidence intervals derived from repeated measures and (2) 95% confidence intervals estimated with the binomial distributions by Hagerman (1976; Hag), Thornton and Raffin (1978; T&R), Carney and Schlauch (2007; C&S), and Raffin and Thornton (1980; R&T).

Dubno *et al.* (1995) reported an analysis of the intersubject variability of word-recognition scores for both ears of 212 subjects with sensorineural hearing loss. They reported that the binomial distribution underestimates the variance of

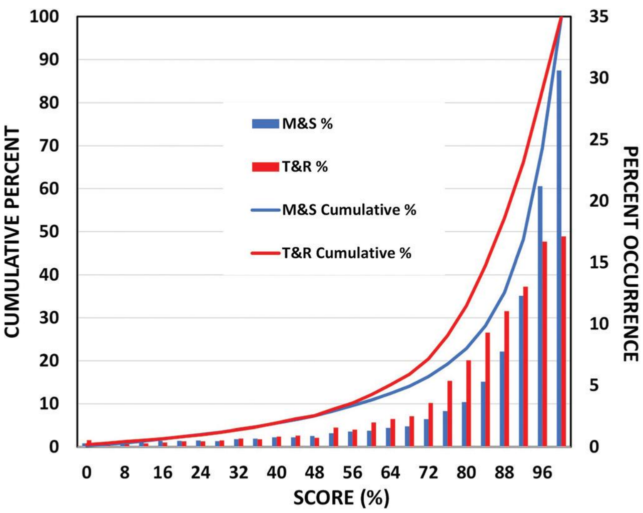

FIG. 9. The cumulative percentage of scores (left ordinate) and percentage of scores (right ordinate) for two datasets: Thornton and Raffin (1978; $n = 4120$; T&R) and Margolis and Saly (2008; $n = 22\,088$; M&S).

scores across subjects. Dillon (1982) discussed the various contributions to intersubject variability of word-recognition scores. It is possible that these sources of variability affected the Dubno *et al*. data in a way that is not accounted for by the binomial distribution. The same principles of independence discussed above apply to the use of the binomial distribution to predict intersubject variance. Just as the assumption of independence is violated for the within subject variance analyzed in this study, the binomial distribution should not be expected to accurately characterize intersubject variance of speech-recognition scores.

B. Independence

The likely source of the disparities between confidence intervals based on repeated measures and those predicted by the binomial distribution is the assumption of independence of test scores that is required by the binomial distribution method. Repeated word-recognition scores violate the assumption of independence in two ways: (1) individual responses from a single subject are not independent and (2) word-recognition scores from repeated tests are not independent.

Two events are independent if one cannot be predicted by the other. Witte and Witte (2007, p. 431) point out that "a violation of independence occurs whenever a single subject contributes more than one observation." Similarly, Cohen *et al*. (2003, p. 532) point out that "… events exhibited by one individual tend to be correlated," and Bijma *et al*. (2017, p. 134) cautioned that "In the case of repeated measures on the same object or person, the dependence of the measurements is unavoidable."

In a word-recognition test, each response from an individual subject is processed by the same normal or impaired auditory system, interpreted by the same central auditory processes, and subjected to the same biases. Bertsekas and Tsitsiklis (2008, p. 34) point out that "… if the occurrence of two events is governed by distinct and noninteracting physical processes, such events will turn out to be independent." An individual's responses to test items are not governed by distinct and noninteracting processes. If the responses were not subject to the processing characteristics of the subject, it would not test that person's word-recognition ability. This interdependence of responses, however, is not accounted for in predictions derived from the binomial distribution.

The interdependence of individual responses results in the highly consistent behavior that is evident in the high correlations shown in Figs. 1, 4, and 6. These high correlations indicate that repeated test scores are not independent. Bertsekas and Tsitsiklis (2008, p. 218) point out that "If X and Y are independent, they are also uncorrelated." High correlations for repeated test scores were shown 60 years ago by Resnick (1962). The high correlations indicate that word-recognition scores, as for the individual responses of which they are composed, are highly interdependent. This interdependence decreases the variability relative to predictions based on an assumption of independence. This accounts for the failure of the binomial distribution to accurately predict the variance of repeated measures.

A test of independence based on correlation coefficients was developed by Bowley (1907, p. 320) and discussed by Mari and Kotz (2001, p. 29). The error associated with a correlation coefficient is given by

$$\text{Error} = \left(0.67\frac{1-r^2}{\sqrt{n}}\right), \tag{4}$$

where r is the Pearson product moment correlation coefficient and n is the number of pairs of measurements. When the ratio of r to the error exceeds a value of 6.0, the variables are not independent. Table IV shows the calculations of the r/error ratios for open- and closed-set scores based on 25 and 50 test items. These ratios substantially exceed the 6.0 criterion, indicating that pairs of scores are not independent.

The dependence of test scores is an important feature of a test. If test and retest scores were not highly correlated (not interdependent), the test would not measure a stable characteristic of the listeners' hearing. This fact alone disqualifies the binomial theory method of estimating confidence intervals of repeated word-recognition test scores. The reduced variance resulting from the dependence of scores is evident in the smaller standard deviations of repeated data relative to those predicted by the binomial distribution (Fig. 3).

The Thornton and Raffin (1978) confidence intervals have been widely interpreted to argue that word-recognition scores based on 25-word lists are too variable to be clinically useful (e.g., Wiley *et al*., 1995). However, many years after the publication by Thornton and Raffin (1978), the majority of clinicians continued to employ 25-word lists (Martin and Morris, 1989; Martin *et al*., 1994), most likely because of the required time necessary to administer 50-word lists. We believe that the wide use of 25-word lists by highly trained and experienced clinicians suggests that the results contribute to the clinical evaluation of their patients. The analysis presented here suggests that word-recognition scores from 25-word lists are highly repeatable and their confidence intervals have been overestimated by untested predictions based on the binomial distribution. The findings that the binomial distribution overestimates the intra-subject variance (present study) and underestimates the inter-subject variance (Dubno *et al*., 1995) strongly suggests that variance and confidence intervals should be based on empirical data rather than mathematical models.

C. Clinical applications of confidence intervals and confidence levels

There are two questions that clinicians seek to answer with speech-recognition testing, question 1 asks, Is a test score significantly different from another score obtained on a different date or with different listening conditions? Different listening conditions include listening with and without hearing aids and changes due to aging, disease state, and treatment of disease. The confidence intervals and confidence levels presented here are offered as a method for interpreting differences between repeat scores. Question 2 asks, Is a test score significantly better or worse than scores of patients with similar hearing losses? The confidence intervals derived by Dubno *et al.* (1995) were offered to evaluate scores in this manner. An examination of the Dubno *et al.* (1995) confidence intervals will be reported in a subsequent article. Future development of speech-recognition test materials should include considerations of these two clinical applications. The variance of repeated measures obtained from test data (not mathematical models) should be determined to address question 1. The distribution of test scores for groups of patients with various degrees of hearing loss should be determined to address question 2.

Table I provides 95% confidence intervals for each possible score for 25- and 50-word lists. A second score, above the upper limit (UL) or below the lower limit (LL), is outside of the 95% confidence interval. which is conventionally used as a criterion for determining if two scores are from the same distribution. For the interpretation of clinical scores, other confidence intervals may be useful.

The confidence levels provided in Tables II and III provide more information than the conventional 95% confidence interval. Scores corresponding to three confidence levels (80%, 90%, and 95%) are provided. To determine if a second score is difference from a first score, one would enter Tables II and III in the first column at the row corresponding to the first score, and find the second score in that row. The level of confidence associated with the difference between scores is characterized as high (95%), moderate (90%), or low (80%). The scores associated with confidence levels less than 80% are interpreted as "no difference." Because scores based on 50-word lists are less variable than those based on 25-word lists, the ranges associated with no difference for 50-word scores (Table III) are narrower than those for 25-word scores (Table II).

D. Limitation of the study

The confidence intervals derived from word-recognition scores presented in this report are based on one set of speech materials [Northwestern University auditory test number 6 (NU-6); Tillman and Carhart, 1966] spoken by a female talker (Causey *et al.*, 1983; Department of Veterans Affairs, 2006). It is possible that scores obtained with other materials or other speakers will show different variability. The study should be replicated with other materials that are in clinical use.

V. SUMMARY AND CONCLUSION

Repeated measures of word-recognition scores using open- and closed-set lists of monosyllabic words were used to generate confidence intervals for differences in pairs of test scores. The confidence intervals are substantially narrower than those predicted by the binomial distribution. The differences in confidence intervals derived by the two methods result from the lack of independence of repeated test scores, which violates a critical assumption of the binomial distribution method of predicting variance. Test scores based on 25-word lists, which have been assumed to be highly variable based on binomial distribution predictions, are highly repeatable and constitute a useful clinical test of word-recognition ability. Tests that employ more items produce narrower confidence intervals in the measured and predicted data. Open- and closed-set scores are characterized by identical confidence intervals. The confidence intervals and confidence limits are provided to assist in the interpretation of differences in repeated test scores.

APPENDIX: NORMALITY OF REPEATED WORD-RECOGNITION SCORES

Normality of repeated word-recognition scores was tested by the following method. For each set of four 25-word open-set scores for normal-hearing and hearing-impaired listeners, the scores were transformed to normalize the mean to equal 50%. The transformed score is defined by

$$S_T = S_M + (S_M - 50),$$

where S_T is the transformed score and S_M is the mean of the four scores.

This process aligns the four scores to a common mean so that the variance around the mean can be determined. For the purpose of plotting the distribution of transformed scores, each score was assigned to one of ten bins ranging from 28–32 to 68–72. The scores below 28% and above 72% were omitted from the analysis because scores close to 0% and 100% have compressed variabilities relative to scores in the middle of the range (Thornton and Raffin, 1978). Figure 10 shows the distribution of transformed scores and best-fit normal distribution (solid line).

The Shapiro-Wilk test of normality (Shapiro and Wilk, 1965) was applied to the distributions of transformed scores for normal-hearing and hearing-impaired listeners. The

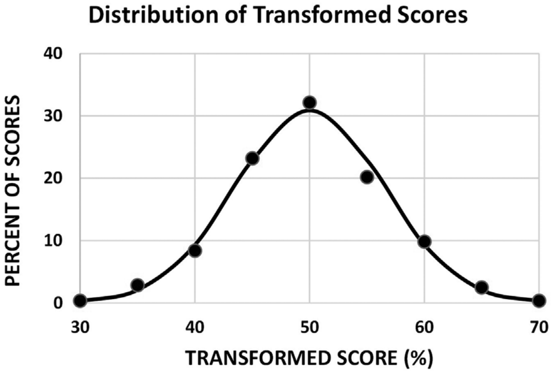

FIG. 10. The transformed 25-word recognition scores. The transformation normalized the mean of each set of four scores to 50%. The solid line is the best-fit, normal distribution.

https://doi.org/10.1121/10.0013826

results indicate that for each group, there is no difference between the transformed scores and normal distribution ($p = 0.07$ for normal-hearing listeners and $p = 0.22$ for hearing-impaired listeners).

[1] See Stats Blue, "Polynomial regression calculator," available at https://stats.blue/Stats_Suite/polynomial_regression_calculator.html (February 7, 2022).

ANSI (**2018**). S3.6-2018, *American National Standard Specification for Audiometers* (American National Standards Institute, New York).

Beattie, R. C., Svihovec, D. A., and Edgerton, B. J. (**1978**). "Comparison of speech detection and spondee thresholds and half- versus full-list intelligibility scores with MLV and taped presentations of NU-6," J. Am. Aud. Soc. **3**(6), 267–272.

Bertsekas, D. P., and Tsitsiklis, J. N. (**2008**). *Introduction to Probability*, second ed. (Athena Scientific, Belmont, MA), pp. 34, 218.

Bess, F. (**1983**). "Clinical assessment of speech recognition," in *Principles of Speech Audiometry*, edited by D. Konkle and W. Rintelmann (Academic, Baltimore, MD), pp. 127–201.

Bijma, F., Jonker, M., and van der Vaart, A. (**2017**). *An Introduction to Mathematical Statistics* (Amsterdam University Press, Amsterdam, Netherlands), p. 134.

Bowley, A. L. (**1907**). *Elements of Statistics*, 3rd ed. (Scribner's, New York), p. 320.

Carney, E., and Schlauch, R. S. (**2007**). "Critical difference table for word recognition testing derived using computer simulation," J. Speech. Lang. Hear. Res. **50**(5), 1203–1209.

Causey, G. D., Hermanson, C. L., Hood, L. J., and Bowling, L. S. (**1983**). "A comparative evaluation of the Maryland NU 6 auditory test," J. Speech Hear. Disord. **48**(1), 62–69.

Cohen, J., Cohen, P., West, S. G., and Aiken, L. S. (**2003**). *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences* (Lawrence Erlbaum Associates, Mahwah, NJ), p. 532.

Department of Veterans Affairs. (**2006**). *Speech Recognition and Identification Materials. Disc 4.0* (VA Medical Center, Mountain Home, TN).

Dillon, H. (**1982**). "A quantitative examination of the sources of speech discrimination test score variability," Ear Hear. **3**(2), 51–58.

Dubno, J. R., Lee, F. S., Klein, A. J., Matthews, L. J., and Lam, C. F. (**1995**). "Confidence limits for maximum word-recognition scores," J. Speech. Lang. Hear. Res. **38**(2), 490–502.

Egan, J. P. (**1948**). "Articulation testing methods," Laryngoscope **58**(9), 955–991.

Elpern, B. (**1961**). "The relative stability of half list and full list discrimination tests," Laryngoscope **71**(1), 30–36.

Gelfand, S. A. (**2018**). *Hearing: An Introduction to Psychological and Physiological Acoustics*, 6th ed. (CRC Press, Boca Raton, FL).

Hagerman, B. (**1976**). "Reliability in the determination of speech discrimination," Scand. Audiol. **5**(4), 219–228.

Hirsh, I. J., Davis, H., Silverman, S. R., Reynolds, E. G., Eldert, E., and Benson, R. W. (**1952**). "Development of materials for speech audiometry," J. Speech Hear. Disord. **17**(3), 321–337.

Holms, A., Illowsky, B., and Dean, S. (**2021**). "Using the central limit theorem," University of Oklahoma and De Anza College, available at https://stats.libretexts.org/@go/page/4584 (Last viewed 13 February 2022).

Kramer, S. J., and Brown, D. K. (**2019**). *Audiology—Science to Practice*, third ed. (Plural, San Diego, CA), p. 168.

Margolis, R. H., and Saly, G. L. (**2008**). "Distribution of hearing loss characteristics in a clinical population," Ear Hear. **29**(4), 524–532.

Margolis, R. H., Wilson, R. H., Saly, G. L., Gregoire, H. M., and Madsen, B. M. (**2021**). "Automated forced-choice tests of speech recognition," J. Am. Acad. Audiol. **32**(9), 606–615.

Mari, D. D., and Kotz, S. (**2001**). *Correlation and Dependence* (Imperial College Press, London, UK), p. 29.

Martin, F. N., Armstrong, T. W., and Champlin, C. A. (**1994**). "A survey of audiological practices in the United States," Am. J. Audiol. **3**(2), 20–26.

Martin, F. N., and Clark, J. G. (**2019**). *Introduction to Audiology*, 13th ed. (Pearson, London, UK), p 120.

Martin, F. N., and Morris, L. J. (**1989**). "Current audiologic practices in the United States," Hear. J. **42**(4), 25–44.

Martin, F. N., and Sides, D. G. (**1985**). "Survey of current audiometric practices," ASHA **2**, 29–36.

Penrod, J. (**1994**). "Speech threshold and word recognition/discrimination testing," in *Handbook of Clinical Audiology*, edited by J. Katz (Williams and Wilkins, Baltimore, MD), pp. 147–164.

Raffin, M. J., and Schafer, D. (**1980**). "Application of a probability model based on the binomial distribution to speech-discrimination scores," J. Speech. Lang. Hear. Res. **23**(3), 570–575.

Raffin, M. J. M., and Thornton, A. R. (**1980**). "Confidence levels for differences between speech discrimination scores: A research note," J. Speech. Lang. Hear. Res. **23**(1), 5–18.

Resnick, D. M. (**1962**). "Reliability of the twenty-five word phonetically balanced lists," J. Aud. Res. **2**(1), 5–12.

Shapiro, S. S., and Wilk, M. B. (**1965**). "An analysis of variance test for normality (complete samples)," Biometrika **52**(3/4), 591–611.

Thibodeau, L. M. (**2000**). "Speech audiometry," in *Audiology Diagnosis*, edited by R. J. Roeser, M. Valente, and H. Hosford-Dunn (Thieme Medical, New York), pp. 281–309.

Thornton, A. R., and Raffin, M. J. M. (**1978**). "Speech discrimination scores modeled as a binomial variable," J. Speech Hear. Res. **21**(3), 507–518.

Tillman, T. W., and Carhart, R. (**1966**). "An expanded test for speech discrimination utilizing CNC monosyllabic words," Northwestern University Auditory Test No. 6, USAF School of Aerospace Medicine Technical Report (Brooks Air Force Base, San Antonio, TX).

Wiley, T. L., Stoppenbach, D. T., Feldhake, L. J., Moss, K. A., and Thordardottir, E. T. (**1995**). "Audiologic practices: What is popular versus what is supported by evidence," Am. J. Audiol. **4**(1), 26–34.

Wilson, R. H., and McArdle, R. (**2015**). "The homogeneity with respect to intelligibility of recorded word-recognition materials," J. Am. Acad. Audiol. **26**(4), 331–345.

Witte, R. S., and Witte, J. S. (**2007**). *Statistics*, 8th ed. (Wiley, Hoboken, NJ), p. 431.