SPEECH-DISCRIMINATION SCORES MODELED AS A BINOMIAL VARIABLE

AARON R. THORNTON

University of Wisconsin, Madison

MICHAEL J. M. RAFFIN

Northwestern University, Evanston, Illinois

Many studies have reported variability data for tests of speech discrimination, and the disparate results of these studies have not been given a simple explanation. Arguments over the relative merits of 25- vs 50-word tests have ignored the basic mathematical properties inherent in the use of percentage scores. The present study models performance on clinical tests of speech discrimination as a binomial variable. A binomial model was developed, and some of its characteristics were tested against data from 4120 scores obtained on the CID Auditory Test W-22. A table for determining significant deviations between scores was generated and compared to observed differences in half-list scores for the W-22 tests. Good agreement was found between predicted and observed values. Implications of the binomial characteristics of speech-discrimination scores are discussed.

Speech-discrimination tests are used in clinical audiometry for several purposes, including diagnosis of ear disease, assessment of communicative impairment, and evaluation of hearing aid performance. To serve a useful purpose, a test must be able to place a subject in an appropriate category of subjects or differentiate his performance in a variety of listening situations. In the first of these cases, the clinician must be concerned with two sources of error (1) the relation between test performance and the parameter of interest (diagnostic category or extent of communicative impairment), and (2) consistency across alternate forms of the test. When different forms of a test are used to compare performance across listening conditions (for example, quiet vs noise), the variations in test-form difficulty are a limiting factor in the ability to measure differences among conditions, and the clinician cannot always determine whether differences in scores are a result of differences in test conditions or differences in test forms. Although test forms are constructed to be equally difficult, this equivalence is usually determined by mean performance on each of the test forms by a group of subjects. For an individual subject, however, the tests are seldom equally difficult, and performance can be expected to vary across forms. In clinical practice the differences in test-form difficulty must be considered when scores are evaluated.

Many previous studies have addressed the question of test-retest reliability,

particularly with the 25- and 50-word tests. In most cases the within-subject variability was confounded with between-subject differences by the comparisons of mean data across subjects or the use of correlation coefficients that are greatly dependent on the spread of scores in the sample of subjects being studied. An additional problem has been probable failure to recognize the special characteristics inherent in percentage scores. Egan (1948) suggested that variability of a test score is a function of the test score itself. He pointed out that variability is at a minimum near the extremes of the articulation scale (0 and 100%) and at a maximum in the middle of the range. He also recognized that variability is dependent on the length of a test list and that error is distributed normally only for mid-range scores. He did not, however, provide a theoretical framework for these observations.

The present paper proposes a simple model to describe variability (across forms) of speech-discrimination tests. Although the model is not original or new, it has largely been ignored in the past.

Examination of the construction of the majority of clinical speech-discrimination tests shows that variance among test forms can be estimated using probabilistic models. Test administration typically requires a subject to respond to each of a series of stimuli, and each response is categorized as correct or incorrect. The test score is reported as the percentage (proportion) of correct responses. Alternate forms of the test are composed of an equal number of similar stimuli. The most frequently used tests are 50-item lists of monosyllabic words. These will be used to discuss the prediction of variability across test forms—however, the discussion is applicable to any stimulus (for example, syllable, word, sentence) that is scored as having only two possible outcomes (such as, correct, incorrect).

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. Hagerman (1976) reached this conclusion by examining individual word difficulties. He noted that scores on theoretical lists of words having uniform difficulty could be described as a binomial variable. However, for lists of words having variable difficulty, he reported data to show that these lists could be equated to slightly longer lists of words having uniform difficulty, and variability then could be described by the binomial distribution corresponding to the longer list.

Hagerman's data also are consistent with a simpler binomial model based on sampling theory. Let us define a pool of stimuli to be used in a test of speech discrimination. For example, we can specify a pool of all common, monosyllabic English words selected to achieve phonetically balanced proportions, and spoken by a single person. For a given subject on a particular occasion each word of the original pool can be assigned to one of two categories—words that will be responded to correctly and words that will be responded to incorrectly. The proportion of words in the original pool to which the subject would respond correctly can be considered the subject's true score $\tilde{p}$ for the test. It is also the expected score for any randomly selected sample from the

pool. The distribution of scores obtained by repeated testing using random samples of equal length drawn from the original pool is described by a binomial distribution with $\tilde{p}$ = proportion of correct responses in the original pool and n = number of items in the (sample) test. This situation is not unlike that encountered with most speech-discrimination tests; each of the alternative test forms can be modeled as a random sample of stimuli drawn from a larger pool defined by the characteristics of the stimuli in all forms of the test combined.

To the extent that speech-discrimination test scores can be described by a binomial distribution, the clinician and researcher should be familiar with the unique characteristics of proportions, particularly with their variability, confidence intervals, and statistical tests of significance. A binomial distribution is completely specified when we know $\tilde{p}$, the proportion of successes in a population (in this case true score or proportion of correct responses in the pool of stimuli), and n, the number of cases drawn from the population (in this case number of words in a list). All other characteristics of the population are irrelevant with respect to variability across test forms. Confidence intervals for estimating a subject's true score and critical differences for determining when two test scores deviate significantly may be computed from the obtained test scores without regard to type of stimulus, subject, or listening conditions as long as these remain constant across test administrations.

METHOD AND RESULTS

Characteristics of the binomial model were compared to the performance of hearing-impaired listeners on a widely used clinical test of speech discrimination. Records of 4120 administrations of the Central Institute for the Deaf (CID) Auditory Test W-22 (Hirsch et al., 1952) were drawn from patient files in the Department of Speech Pathology and Audiology at the Veterans Administration Hospital in Iowa City. The clinic case load typically consisted of patients in their late 50s to early 60s, but they ranged in age from approximately 20 to 80 years. The recorded, 50-item, monosyllabic word lists were presented at 40 dB re SRT whenever possible. Each of the four alternate forms of the test (Lists 1-4) were used with 1030 ears, and six standard randomizations (A–F) of word order for each list were represented as shown in Table 1. Table 2 shows the distribution of scores for each of the lists. Table 3 shows the distribution of item difficulty within each list.

The standard deviation of a binomial distribution depends on both the probability of a success and the number of cases drawn. In percentage,

$$SD = 100 \sqrt{\left[\frac{(\tilde{p})(1-\tilde{p})}{n}\right]}. \qquad (1)$$

A comparison was made between standard deviations computed for binomial variables and standard deviations of scores on subsets of the 50-word tests.

TABLE 1. Number of cases sampled for each randomization of the four 50-word lists of the CID Auditory Test W-22.

Randomization	List			
	1	2	3	4
A	131	126	175	181
B	63	57	177	182
C	174	182	197	178
D	248	248	174	174
E	158	156	62	65
F	256	261	245	250
Total	1030	1030	1030	1030

Each 50-word list (1-4) was divided into two 25-word lists and five 10-word lists by a sequential division of randomization A. Scores on these shorter lists were then computed from each of the 4120 50-word tests. Tests were grouped by 50-word scores, which were used as the best estimate of $\tilde{p}$ (true score), and an analysis of variance was used to compute variability at 25 levels of $\tilde{p}$ (50-word scores of 50-96%) for the 25-word scores, and 17 levels of $\tilde{p}$ (50-word scores of 44-98%) for the 10-word scores. The 44-56, 58-64, and 66-70 percentage levels were pooled as three broader groupings in the 10-word analysis. Standard deviations were computed as

$$SD = \sqrt{\left(\frac{SS \text{ total} - SS \text{ subjects}}{df \text{ total} - df \text{ subjects}}\right)} \text{ or } \sqrt{\left(\frac{SS \text{ within-subjects}}{df \text{ within-subjects}}\right)}. \quad (2)$$

For example, Table 2 shows that 80 subjects were estimated to have a true score p of 72% based on their 50-word test scores. The tests were rescored as 160 25-word tests and 400 10-word tests. The standard deviations of these part-list scores were 8.22% and 13.51% respectively. The correspondence between the empirically measured standard deviations at each level and theoretical binomial standard deviations is shown in Figure 1. Test variability appears to be dependent on a subject's true score and the number of words in the test.

When the form of a distribution is known, an inference about the population mean can be made from a sample mean. This frequently takes the form of a confidence interval, a range of scores about the sample mean that has a specified probability of encompassing the population mean. The range is usually positioned about the sample mean such that the probabilities of the population mean falling outside either end of the range are equal. Tables and charts of confidence intervals for proportions (for example, Steel and Torrie, 1960; Pearson and Hartley, 1966) may be supplemented by use of a Z-table and an equation explained by Hays and Winkler (1970). The size of the confidence interval and its symmetry about the sample score are dependent on both the sample score and the number of events in the sample. For example, when a subject scores 92% on a 50-word test, the 95% confidence interval of the true score is from 81% (92 − 11) to 98% (92 + 6). For a score of 68% it is 54%

TABLE 2. Distribution of scores for each of the four 50-word lists of the CID Auditory Test W-22.

	List				
Score	*1*	*2*	*3*	*4*	*Total*
0	2	6	5	10	23
2	0	1	1	2	4
4	0	1	0	2	3
6	0	2	1	2	5
8	0	0	2	2	4
10	2	1	3	5	11
12	2	4	3	3	12
14	0	2	2	2	6
16	0	2	2	5	9
18	2	0	1	3	6
20	2	4	3	2	11
22	0	3	0	2	5
24	1	4	4	2	11
26	2	4	2	5	13
28	4	2	4	5	15
30	2	3	2	5	12
32	1	4	4	3	12
34	1	6	4	4	15
36	1	6	5	4	16
38	3	0	3	7	13
40	5	5	6	6	22
42	7	5	3	5	20
44	3	4	10	4	21
46	2	5	2	4	13
48	1	4	5	6	16
50	1	8	4	7	20
52	6	5	2	6	19
54	9	4	9	8	30
56	6	4	8	10	28
58	5	8	11	16	40
60	6	6	6	13	31
62	7	10	10	19	46
64	7	13	12	9	41
66	11	13	10	17	51
68	13	9	4	9	35
70	10	12	19	20	61
72	19	16	19	26	80
74	25	19	23	21	88
76	19	23	22	29	93
78	25	27	29	25	106
80	36	41	41	45	163
82	56	50	45	46	197
84	47	44	52	36	179
86	59	45	56	51	211
88	59	44	66	77	246
90	67	71	77	68	283
92	73	84	83	73	313
94	106	96	107	88	397
96	127	141	121	99	488
98	139	123	96	80	438
100	49	36	21	32	138
Total	1030	1030	1030	1030	4120

TABLE 3. Percentage of incorrect responses made to each word of the CID Auditory Test W-22. Each list was measured on an independent sample of 1030 ears.

	List 1		List 2		List 3		List 4	
Item	Word	% Incorrect	Word	% Incorrect	Word	% Incorrect	Word	% Incorrect
1	up	2.82	now	5.24	out	3.79	why	5.53
2	none	2.91	well	5.63	when	3.98	men	6.41
3	what	2.91	one	5.73	on	4.27	ought	8.06
4	yard	3.40	eat	5.83	book	4.56	in	8.64
5	him	3.59	that	6.31	no	5.24	jump	8.84
6	us	3.88	odd	6.80	are	6.02	cook	9.03
7	you	3.98	yore	6.89	oil	6.51	my	9.32
8	it	4.18	by	7.28	ate	7.28	wood	9.32
9	hunt	4.37	air	7.38	done	7.57	who	9.61
10	dad	4.76	tree	7.38	this	7.67	toy	9.81
11	there	4.76	star	8.25	he	8.45	pale	9.90
12	me	5.34	own	9.03	pie	8.74	aid	10.10
13	poor	5.53	die	9.13	jar	8.84	at	10.19
14	or	5.92	flat	9.42	add	10.00	bread	10.19
15	wet	6.60	young	9.42	may	10.00	our	10.97
16	low	6.70	hurt	9.52	glove	10.58	of	11.26
17	could	6.80	too	9.52	have	10.49	they	12.62
18	not	6.89	oak	9.71	if	10.58	shoe	13.01
19	as	7.57	and	9.81	raw	10.87	yet	13.88
20	isle	7.86	smart	10.29	shove	11.17	leave	13.98
21	give	8.06	live	10.58	bill	11.46	bee	15.24
22	day	8.25	off	11.07	cute	12.14	will	15.44
23	law	8.35	does	11.26	do	12.82	clothes	16.02
24	true	8.74	way	11.36	lie	13.20	through	16.21
25	felt	9.03	dumb	11.85	end	13.40	yes	16.70
26	toe	9.03	then	12.33	farm	13.40	am	16.89
27	ran	9.22	jaw	13.40	smooth	13.59	where	17.48
28	skin	9.61	bin	14.47	tie	13.59	his	18.16
29	wire	10.10	see	14.56	hand	13.69	so	20.78
30	earn	10.49	new	16.12	is	18.06	arm	20.87
31	she	11.07	hit	17.28	three	18.54	go	20.78
32	high	11.94	ham	17.77	ten	18.74	few	22.52
33	them	12.82	ill	18.35	chair	20.29	eyes	23.20
34	stove	15.34	show	18.35	though	20.49	all	23.30
35	twins	16.12	cars	20.49	we	22.04	hang	23.79
36	see	19.03	thin	20.49	use	23.50	ear	26.12
37	owl	20.10	tare	21.36	say	24.08	chin	26.60
38	carve	20.19	with	22.43	king	24.56	than	28.64
39	jam	20.29	chest	22.52	wool	25.15	save	28.93
40	thing	23.01	ease	23.20	camp	27.09	can	32.91
41	east	24.27	gave	23.98	year	28.64	near	33.11
42	an	25.56	move	24.37	aim	29.17	darn	34.27
43	bells	27.48	cap	25.24	start	30.78	tin	34.95
44	chew	28.84	else	25.73	dull	33.40	stiff	36.41
45	ace	33.79	ail	26.41	owes	36.89	art	38.84
46	bathe	35.44	key	34.76	ears	37.57	tea	39.42
47	ache	37.77	pew	33.98	tan	38.64	net	39.81
48	knees	38.35	rooms	46.99	west	39.90	nuts	48.64
49	deaf	52.23	send	48.74	knit	40.19	dust	48.74
50	mew	58.93	knee	74.85	nest	49.81	dolls	52.43
Mean		14.48		17.06		17.63		20.76
SD		12.86		12.85		11.38		12.07

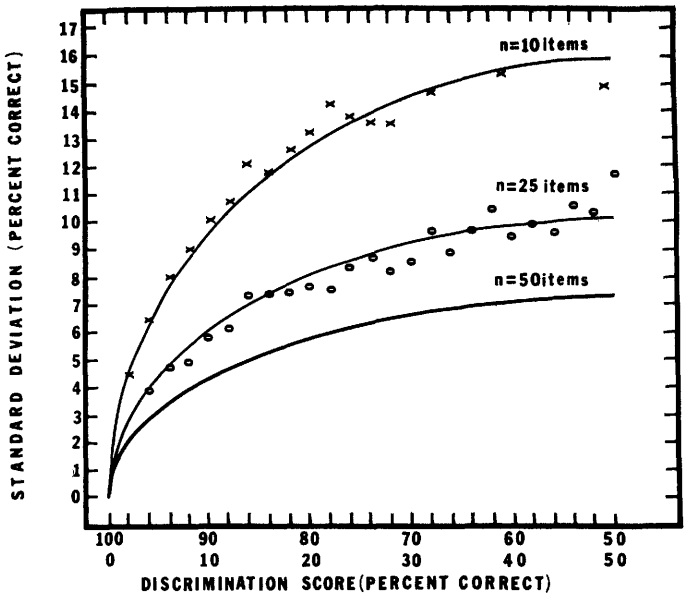

FIGURE 1. Within-subject standard deviations for 10- and 25-word tests grouped by estimated true scores (50-words). Solid lines show standard deviations of binomial distributions as a function of p (in percentage) for $n = 10$, $n = 25$, and $n = 50$. Measured standard deviations for $n = 10$ and $n = 25$ are shown by X and 0 respectively.

(68 − 14) to 80% (68 + 12). When the test is shortened to 25 words, scores of 92% and 68% have confidence intervals of 74% to 99% (92 − 18, 92 + 7) and 46% to 85% (68 − 22, 68 + 17). As can be seen, a simple rule cannot be generated to describe all cases.

Although the binomial confidence intervals may be useful in estimating a range of uncertainty about the location of a subject's true score, they cannot be used to solve the more common problem of determining when two scores are significantly different. To test an obtained difference in scores against a hypothesis of no difference, the theoretical distribution of difference scores must be defined, but the characteristics of the distribution of differences between binomial variables are not available.

An approximate solution to determining the significance of observed differences in scores may be made by first transforming the scores to a variable that has uniform variance, then calculating the variance of a difference between transformed scores and, finally, estimating the probability of an observed difference occurring by chance by using a Z-table. The Freeman and Tukey (1950) averaged angular transformation for stabilization of variance in binomial data was used as the basis for generating tables of critical differences. For a word list of a specific length, each possible score was transformed to an angle by

$$\theta = \sin^{-1}\sqrt{\left(\frac{x}{n+1}\right)} + \sin^{-1}\sqrt{\left(\frac{x+1}{n+1}\right)}, \qquad (3)$$

where n = number of words in the list and x = number of correct responses. The estimated variance of this angle was adjusted for sample size as suggested by Mosteller and Youtz (1961):

$$\text{For } n \geq 50\text{: } \sigma_\theta^2 = \frac{1}{n + \frac{1}{2}}, \tag{4}$$

$$\text{for } 10 < n < 50\text{: } \sigma_\theta^2 = \frac{1}{n + 1}. \tag{5}$$

The estimated variance of a difference between two independent angles would simply be twice the above values when n's are equal. Using this variance and a Z-table, 95% confidence intervals about a hypothesized angular difference of zero were computed for n = 10, 25, 50, and 100. For n = 25 a 90% confidence interval was also computed.

The angular confidence intervals were used to generate tables of critical differences that might be used clinically (see Table 4). The term critical difference is used to avoid confusion with the confidence interval for predicting a true score. With respect to any test score, the critical difference is specified by upper and lower limits, which are the largest and smallest test scores whose transformed θ-values fall within the angular confidence interval (for a hypothesized true difference of 0) (see Appendix). Because binomial test-score distributions change in steps, the critical differences will usually have a confidence level less than that specified for the angular confidence interval. Also, the angular transformation does not work well as the tails of the distribution, and this error is carried to the critical differences.

The critical differences (shown in Table 4) show an asymmetry and dependence on sample score and n similar to that described for the confidence intervals of the true scores. For example, if a subject scored 92% on one form of a 50-word test, there is a 95% probability that his score on another form of the test would fall within the range 78-98% (−14%, +6%). For a score of 68% the range is 50%–84% (−18%, +16%), which is greater but also more symmetrical. For the 25-word tests the critical differences for scores of 92% and 68% increase to 72%–100% and 44%–88% respectively.

The 95% critical differences (shown in Table 4) and 90% critical differences (not shown) were compared to observed differences between 25-word scores for each of the 4120 50-word tests. Each of the four lists was divided into two 25-word tests by a random assignment of the 50 words different from randomizations A–F. The number of differences in half-list scores that exceeded the theoretical critical differences were tallied and are shown in Table 5. Good agreement is seen between predicted and obtained results. The total percentage of scores falling outside the suggested critical differences was 5.4% for the 95% limits and 7.9% for the 90% limits. The imbalance between upper and lower limit errors for high scores is consistent with the corresponding skewness of the binomial distribution at these points and with the discrete characteristics of the distribution. The total proportion of differences falling beyond the

TABLE 4. Lower and upper limits of the 95% critical differences for percentage scores. Values within the range shown are not significantly different from the value shown in the percentage Score columns ($p > 0.05$).

% Score	n = *50*	n = *25*	n = *10*	*% Score*	n = *100**
0	0-4	0-8	0-20	50	37-63
2	0-10			51	38-64
4	0-14	0-20		52	39-65
6	2-18			53	40-66
8	2-22	0-28		54	41-67
10	2-24		0-50	55	42-68
12	4-26	4-32		56	43-69
14	4-30			57	44-70
16	6-32	4-40		58	45-71
18	6-34			59	46-72
20	8-36	4-44	0-60	60	47-73
22	8-40			61	48-74
24	10-42	8-48		62	49-74
26	12-44			63	50-75
28	14-46	8-52		64	51-76
30	14-48		10-70	65	52-77
32	16-50	12-56		66	53-78
34	18-52			67	54-79
36	20-54	16-60		68	55-80
38	22-56			69	56-81
40	22-58	16-64	10-80	70	57-81
42	24-60			71	58-82
44	26-62	20-68		72	59-83
46	28-64			73	60-84
48	30-66	24-72		74	61-85
50	32-68		10-90	75	63-86
52	34-70	28-76		76	64-86
54	36-72			77	65-87
56	38-74	32-80		78	66-88
58	40-76			79	67-89
60	42-78	36-84	20-90	80	68-89
62	44-78			81	69-90
64	46-80	40-84		82	71-91
66	48-82			83	72-92
68	50-84	44-88		84	73-92
70	52-86		30-90	85	74-93
72	54-86	48-92		86	75-94
74	56-88			87	77-94
76	58-90	52-92		88	78-95
78	60-92			89	79-96
80	64-92	56-96	40-100	90	81-96
82	66-94			91	82-97
84	68-94	60-96		92	83-98
86	70-96			93	85-98
88	74-96	68-96		94	86-99
90	76-98		50-100	95	88-99
92	78-98	72-100		96	89-99
94	82-98			97	91-100
96	86-100	80-100		98	92-100
98	90-100			99	94-100
100	96-100	92-100	80-100	100	97-100

*If score is less than 50%, find % Score = 100-observed score and subtract each critical difference limit from 100.

TABLE 5. Number of scores on second 25-words that are less than the lower limit (< LL) or greater than the upper limit (> UL) of the 90% and 95% critical differences shown in Table 4.

Score on First 25	n	*90%*		*95%*	
		< LL	*> UL*	*< LL*	*> UL*
0	23	0	0	0	0
4	12	0	3	0	0
8	17	0	1	0	1
12	12	1	1	1	1
16	15	1	0	1	0
20	19	2	0	0	0
24	19	1	0	1	0
28	22	0	1	0	0
32	28	3	1	0	0
36	26	0	0	0	0
40	35	3	3	0	0
44	38	1	2	0	1
48	31	3	2	1	2
52	64	2	2	1	2
56	58	4	1	3	1
60	81	7	6	4	1
64	93	1	8	0	8
68	102	2	9	1	6
72	147	5	10	3	1
76	221	1	9	1	9
80	290	4	25	3	5
84	383	6	16	4	16
88	455	14	35	8	35
92	537	11	0	5	0
96	687	35	0	12	0
100	705	85	0	85	0
Totals	4120	192	135	134	89
		327		223	
		7.9%		5.4%	

limits, however, is consistent with the specified confidence levels. A change in the upper limits would only result in greater discrepancies. For example, if the upper limit of the 95% critical difference for a score of 92% were lowered one step from 100% to 96%, then 19.7% (106) of the 537 observed differences would fall beyond the upper limit.

CONCLUSIONS

The binomial characteristics of speech-discrimination tests make variability among test forms dependent on both the number of items in the test and the subject's true score for the class of items used. The optimal number of items for a test must be determined from an estimate of the true scores of the subjects to be tested and a recognition of the trade-off between administration time and variability. For some clinical purposes, 25 words may be sufficient, whereas 100 may not be enough for others. In selecting the number of stimuli, the clinician needs to know how accurately he must estimate a true score or, alter-

natively, how small a difference between test scores that he must be able to measure with certainty. For example, if a clinician determines a subject's score to be 100% on 25 words, then he can estimate the true score of the subject to be within the range of 86% to 100% (95% confidence interval; see Steele and Torrie, 1960, p. 454). For some clinical screening purposes this may be sufficient precision. However, a 25-word score of 48% would place the range of uncertainty for the true score at 28%–69%, which permits only a gross classification. If the person having the 100% score were retested at a later date and the 25-word score dropped to 88%, the 12% difference in scores would be strong evidence that the true score was different on the two occasions (see Table 4). However, a similar (12%) decrease of the 48% score to 36% would be only weak evidence of a change in true score because the second score lies well within the 95% critical difference. When judging and comparing the performance of hearing aids, it is important to use as many words as may be needed to measure changes in true scores associated with differences in the hearing aids. These shifts in true score cannot be determined when they do not exceed chance variation. Testing in the presence of noise often lowers the true scores and consequently increases the variability and error. Increased differences among the scores obtained with various hearing aids tested in noise may in many cases be explained by the increased variability of the measuring instrument as opposed to real differences in the true scores for the different hearing aids.

We recommend that clinicians always indicate the number of items used in a test, or report the confidence interval in addition to the obtained score. This would accurately reflect the uncertainty of the location of a subject's true score. For comparison of scores, the critical differences should be used to determine significance. The skewness of the critical differences and their dependency on the number of items and subject's score make it difficult to judge the significance of differences in scores without the help of a table. This is particularly true for the clinician in training.

We wish to emphasize that the binomial characteristics of speech-discrimination tests are relatively independent of subject characteristics, listening conditions, and type of stimulus. These factors will, however, influence the true score of a subject. Although it is tempting to try to devise a test with fewer items that will have the same variability across forms as a larger test, it is unlikely that this can be done without substantial changes in construction and scoring to eliminate the binomial characteristics. As shown by our sample, most subjects score high on currently used clinical tests. From Figure 1 we can see that the variance changes rapidly as 100% is approached, and this feature makes it imperative to use variance stabilization transformations on test scores before performing statistical tests that assume that variance is independent of the score. Finally, application of the binomial model might be extended to other tests used in the field of communicative disorders when these tests meet the basic assumptions required for the model. This application is particularly important when the test scores are used to classify subjects or measure progress during treatment.

ACKNOWLEDGMENT

Appreciation is extended to Herbert Jordan for making his clinical records available for this study. The majority of the work on this paper was completed while both authors were at the University of Iowa, and portions of the work have been presented at the 1976 Annual Convention of the American Speech and Hearing Association in Houston, Texas and the May 1977 Meeting of the Canadian Speech and Hearing Association in Victoria, British Columbia. Requests for reprints should be addressed to the first author at the Department of Communicative Disorders, University of Wisconsin, 1975 Willow Drive, Madison, Wisconsin 53706.

REFERENCES

EGAN, J. P., Articulation testing methods. *Laryngoscope,* **58**, 955-991 (1948).

FREEMAN, M. F., and TUKEY, J. W., Transformations related to the angular and the square root. *Ann. math. Statist.,* **21**, 607-611 (1950).

HAGERMAN, B., Reliability in the determination of speech discrimination. *Scand. Audiol.,* **5**, 219-228 (1976).

HAYS, W. L., and WINKLER, R. L., *Statistics: Probability, Inference and Decision.* (Vol. 1) New York: Holt, Rinehart and Winston, 332 (1970).

HIRSH, I. J., DAVIS, H., SILVERMAN, S. R., REYNOLDS, E. G., ELDERT, E., and BENSON, R. W., Development of materials for speech audiometry. *J. Speech Hearing Res.,* **17**, 321-337 (1952).

MOSTELLER, F., and YOUTZ, C., Tables of the Freeman-Tukey transformations for the binominal and Poisson distributions. *Biometrika,* **48**, 433-440 (1961).

PEARSON, E. S., and HARTLEY, H. O., *Biometrika Tables for Statisticians.* (Vol. 1) Cambridge (Eng.): Cambridge Univ. Press (1966).

STEELE, R. G. D., and TORRIE, J. H., *Principles and Procedures of Statistics.* New York: McGraw-Hill (1960).

Received September 5, 1977.
Accepted January 3, 1978.

APPENDIX

The procedure for computing critical differences will be illustrated by the following example. Consider the problem of determining the 95% critical differences for a score of 90% on a 50-word test (45 correct).

By Equation (3) $\theta_{90\%} = 2.473$ radians

By Equation (4) $\sigma_\theta^2 = 0.0199$ radians

$2\sigma_\theta^2 = 0.0398$ radians

For a difference between two angles $\theta_1 - \theta_2$: $\sigma_{\theta_1 - \theta_2} = 0.1996 \text{ radians } \sqrt{(2\sigma_\theta^2)}$

The 95% confidence interval for an angular difference will be bounded by upper (UL) and lower limits (LL).

$$\theta_{LL} = \theta - 1.96\,\sigma_{\theta_1 - \theta_2} \qquad \theta_{UL} = \theta + 1.96\,\sigma_{\theta_1 - \theta_2}$$

For $\theta_{90\%}$: $\theta_{LL} = 2.082$ radians $\theta_{UL} = 2.864$ radians

By a process of iteration with Equation (3), one can determine minimum and maximum scores that have θs within the computed limits.

$\theta_{76\%} = 2.106$ radians $\theta_{98\%} = 2.697$ radians (within the limits)

$\theta_{74\%} = 2.061$ radians $\theta_{100\%} = 3.001$ radians (exceed the limits)

Critical differences for a score of 90% are 76% and 98%.

Speech-Discrimination Scores Modeled as a Binomial Variable

Aaron R. Thornton, and Michael J. M. Raffin
J Speech Hear Res 1978;21;507-518