Reliability of Judgments of Articulation of Cleft Palate Speakers

BETTY JANE W. PHILIPS, Ed.D.
KENNETH R. BZOCH, Ph.D.
Miami, Florida

Articulation test data are usually considered to have a satisfactory degree of reliability. This has been demonstrated by investigators employing correlation techniques (*4*, *7*, *8*) and by those reporting percentage of agreement among examiners (*1–3*). Siegel (*5*), however, questioned the interpretation of reliability as indicated by a high correlation because he also found significant differences in the absolute test scores determined by the examiners for each subject.

Some investigators, such as Sommers and others (*6*), have found that the level of reliability can be improved by having judges train together in evaluating articulation responses. However, under conditions such as those imposed by collaborative research at widely scattered clinical centers or in longitudinal projects, such training may not be possible.

Therefore, it was the purpose of this study to determine the reliability of judgments of tape-recorded speech samples of cleft palate subjects when evaluated by speech pathologists using written instructions and definitions of criteria for evaluation.

The following specific questions were investigated: a) What is the reliability of intra- and interjudge evaluations of the intelligibility of connected speech, the identification of articulation errors, and the classification of types of articulation errors? b) Does the reliability of articulation judgments vary with the phonetic classification of the sounds, the position of the sounds, or the error descriptions? c) Are there wide variations in the specific articulation scores despite a satisfactory level of reliability as determined by percentage of interjudge agreement? d) If marked variation exists in the specific articulation scores, does it impair the clinical research value of articulation evaluations?

Procedure

SUBJECTS. Tape recordings were made of the speech of 50 cleft palate subjects. The only requirement for inclusion in the study was that the

Dr. Philips is Consultant in Speech Pathology, University of Miami School of Medicine and Dr. Bzoch is Chairman, Department of Communicative Disorders College of Health Related Professions, University of Florida.

This project was supported in part by PHS Research Grant, DE-01836-01, National Institute of Dental Research.

subjects must be six years of age or older. Type or severity of the congenital cleft was not a consideration for inclusion. The subjects exhibited a range of articulation skills varying from normal to the severe dyslalia sometimes noted in cleft palate speech.

SPEECH SAMPLE. Two tape recordings were made for each patient; one was a recording of single word articulation test responses and the other recording was a sample of connected speech. The evaluation of the articulation test required judgments of 100 phonemic elements: 67 single consonants and 33 consonant blends. A picture stimulus elicited the subject's response for each of the 100 test words. Connected speech was elicited by having each subject count from one to ten and recite the nursery rhyme, "Baa, Baa, Black Sheep . . . " This particular rhyme was chosen because of its familiarity to children and because of the broad range of consonant sounds it contains.

TAPE RECORDING. Recordings were made at 3¾ inches per second using an Ampex Recorder, Model 9-10. Each subject was seated in a chair with a headrest to standardize body position. The microphone, mounted on a floorstand, was placed ten inches from the mouth of each subject. A consistent intensity level was maintained throughout the recording by monitoring with a VU meter outside the audiometric booth where the recordings were made.

The subject's response to each picture was elicited three times for each word. An interval of approximately four seconds separated each item on the edited recordings.

The fifty tapes of connected speech and the fifty tapes of articulation test responses were edited to eliminate any extraneous remarks. Duplicate tapes were made by a professional recording company so that every judge would have a complete set of tapes. In addition, the tape recordings of five subjects were duplicated and assigned different numbers so that each of the judges, unknowingly, made two separate evaluations for each of these subjects.

JUDGES. Ten speech pathologists from different centers throughout the country, each of whom had extensive clinical experience with cleft palate subjects, evaluated the tapes. Each judge was required to make 5,500 specific articulation judgments and to rate 55 connected speech samples.

JUDGING PROCEDURE. The judges evaluated only four recordings in a single day and in the following order: a recording of the articulation test, a connected speech sample, and after a brief pause, a second articulation test, and a second connected speech sample. No two of the four recordings judged in a day were taken from the same subject.

All of the judges evaluated all the tapes in the same order. They were permitted to stop the tape or to re-play any section of it, if they so desired. Written instructions and definitions were provided to assist the judges in classification of the speech sound errors. The judges were instructed to evaluate only the most severe error heard in the three responses to each item. The production could be evaluated as a) normal production, b) distortion by nasal emission only, c) an indistinct production, d) a simple substitution, e) a gross substitution (that is, a glottal stop or pharyngeal fricative), or f) an omission.

In evaluating the passage of connected speech, the judges circled a number on a five-point scale which best defined the degree of intelligibility. The following scale values were used: *one*, clearly intelligible, no difficulty in understanding the speech, no articulation errors noticed; *two*, intelligible, no difficulty in understanding the speech, occasional articulation errors noticed; *three*, usually intelligible, speech usually understandable, but consistent articulation errors may cause some confusions for the listener; *four*, partially unintelligible, some difficulty in understanding the speech due to many articulation errors noticed; and *five*, unintelligible, extreme difficulty in understanding the speech, many articulation errors noticed.

Findings

INTELLIGIBILITY RATINGS OF CONNECTED SPEECH. To provide an indication of the degree of defectiveness of the subjects, a tabulation was made of the number of times each judge assigned each of the intelligibility ratings. More speakers were assigned ratings of *one*, *two*, and *three*, than ratings of *four* and *five*. Although, as a group, the majority of the subjects were not seriously defective in intelligibility, the sample did cover the entire range of the intelligibility scale.

Intrajudge agreement was determined by having each evaluator, unknown to him, judge five of the subjects twice. A Pearson product-moment correlation coefficient between the first and second set of ratings was determined for each judge. Although the correlations ranged in value from .25 to 1.00, the average correlation was .79, indicating a high positive relationship between the scores obtained on the first and second rating.

In Table 1 the levels of agreement within and between judges are compared. In the intrajudge comparison perfect agreement about a specific rating occurred 54 % of the time and was within one point on the scale an additional 40% of the time. To determine interjudge reliability each evaluator was paired with every other evaluator and the agreement between these 45 pairs of evaluations was determined for each of the 50 subjects. The findings, similar to those for intrajudge evaluations, indicate close agreement. Only 5 % of the time did agreement on rating of intelligibility vary more than one point.

ARTICULATION TEST, TOTAL ERROR SCORES. The number of errors marked by each of the ten evaluators was averaged for each of the 50 subjects. The scores ranged from 1 to 94, out of a possible 100. There was a greater frequency of lower error scores than of higher scores which indicates, as did the ratings of intelligibility, that the majority of the subjects were not severely defective in articulation.

TABLE 1. Intra- and interjudge agreement on ratings of intelligibility.

Level of agreement	*Percentage of agreement*	
	Intrajudge	*Interjudge*
Perfect agreement	54%	49%
Within one scale value	40%	46%
Within two scale values	6%	4%
Within three scale values	0%	1%

To determine intrajudge reliability the same test-retest method was used as for the ratings of intelligibility. A Pearson r was determined for each evaluator's test-retest of five subjects; correlations ranged from .43 to .99, and averaged .93.

Percentage levels of agreement as to presence or absence of articulation errors were also computed. Intrajudge agreement was found by counting the number of agreements on the evaluation of each phonemic element for each patient. If on both the first and second evaluations the evaluators indicated that a production was normal, or if on both they indicated that it was defective, the evaluators were considered to be in agreement regardless of the classification given the perceived error. Since the total number of possible agreements per subject was 100, the count of the number of agreements yielded a percentage. The same method was used for determining interjudge agreement. Each evaluator was paired with every other evaluator. Table 2 shows that intrajudge agreement ranged from 71% to 92% and averaged 85%. Interjudge agreement ranged from 68% to 79% and averaged 74%. Both intra- and interjudge agreements are well above the level of chance.

TABLE 2. Intra- and interjudge agreement on presence or absence of articulation errors.

Judges	*Percentage intrajudge agreement*	*Percentage of interjudge agreement*								
		B	*C*	*D*	*E*	*F*	*G*	*H*	*I*	*J*
		(All entries are percentages)								
A	92	77	74	70	76	79	75	77	72	74
B	87		74	76	76	78	73	76	71	75
C	84			77	76	75	71	75	70	79
D	85				76	74	71	74	71	77
E	91					79	71	78	68	78
F	87						72	78	75	77
G	88							74	73	71
H	85								70	77
I	71									69
J	87									

VALIDITY OF ARTICULATION TEST SCORES. The intervals on the rating scale for intelligibility of connected speech samples were defined in terms of both levels of understandability and level of noticeability of articulation errors in connected speech. As a measure of the validity of the articulation scores, a Pearson *r* was used to estimate the strength of the relationship between the articulation error scores and the intelligibility ratings, for each judge. The average for these correlation coefficients was .44.

By ranking the means of the articulation scores and the means of the intelligibility ratings for each of the subjects, a rank-difference correlation of .87 was obtained. This rank-difference correlation coefficient was higher than the Pearson *r*, probably because of the increased number of scale values obtained when mean intelligibility ratings were used (counteracting the problem of usage of a five-point scale in the Pearson *r*). Further, reliability of the articulation scores was increased by use of the averaged or mean articulation scores.

VARIABILITY OF ARTICULATION ERROR JUDGMENTS. Data regarding the test-retest conditions are presented in Table 3. For each judge, test scores are reported for each of the five test-retest conditions, the range of difference between test-retest, and the average difference. Difference in intrajudge scores ranged from 0–51 points and averaged 8.5. The range of interjudge scores for each of the 50 subjects is presented in Table 4. These differences in scores ranged from 17 to 76 points and averaged a range of 41.8 points. These ranges show that, although individual judges were fairly consistent, there was considerable variability among judges.

FACTORS AFFECTING INTERJUDGE AGREEMENT. The wide variability of test scores led the investigators to study the data to determine where the evaluators demonstrated the greatest differences in judgment. The percentages of agreement among the pairs of examiners were determined according to position of sound in the word (initial, medial, final) and phonemic classification. The percentages of agreement are given in Table 5. For both the position of sounds in words and phonemic classification the levels of agreement showed some variation. Fricative, glides and blends, and sounds in final positions account for a greater proportion of the disagreements.

To determine whether or not the severity of the speech disorder, as indicated by the intelligibility ratings, caused the evaluators to be more or less discrepant in their judgments, a rank-order correlation was determined for the range of the articulation test scores and the mean intelligibility ratings for each of the fifty subjects. A rank-order correlation of .19 was found which would indicate that there is little relationship between the variability in the test scores and the defectiveness of the speech of individual cases in the sample.

VARIABILITY IN CLASSIFICATION OF ERRORS. Another area of interest is the level of agreement on specific classification of errors. That is, when an error was judged to be present, did the judges agree on the type of error

TABLE 3. Variability of intrajudge articulation error scores. For each judge, test scores (number of errors) are given for each of five test-retest conditions, range of difference between test-retest, and average difference.

Judge	*Subjects*					*Range of difference*	*Average difference*
	1	2	3	4	5		
A							
test	2	25	79	21	25	0–7	3.4
retest	2	29	74	14	24		
B							
test	22	61	67	37	45	1–19	7.2
retest	17	42	57	38	44		
C							
test	6	26	58	37	55	8–25	12.8
retest	14	39	68	62	47		
D							
test	16	47	67	44	50	2–13	7.4
retest	29	49	77	52	46		
E							
test	11	71	36	32	28	2–19	11.2
retest	13	53	55	37	40		
F							
test	2	20	43	23	41	0–13	5.6
retest	2	18	56	24	29		
G							
test	29	68	90	42	39	1–11	3.8
retest	26	70	91	53	41		
H							
test	8	45	44	24	36	3–11	5.8
retest	5	40	55	29	31		
I							
test	1	37	86	53	61	1–51	23.6
retest	52	77	85	61	43		
J							
test	13	45	61	41	39	3–8	4.2
retest	21	42	58	45	42		

which occurred? As shown in Table 6, intrajudge agreement ranged from 50% to 91%. Interjudge agreement ranged from only 6% to 19% and was uniformly low.

A breakdown of interjudge percentage of agreement for each category

TABLE 4. Variability of interjudge articulation error scores. For each of the 50 subjects, the lowest score, the highest score, the difference (or range), and the mean of the scores.

Subject	*Lowest score*	*Highest score*	*Mean*	*Range*
1	2	59	16	57
2	8	82	35	74
3	6	68	25	62
4	6	36	19	30
5	1	18	7	17
6	13	89	34	76
7	3	47	21	44
8	25	44	33	19
9	2	20	11	18
10	2	31	9	29
11	36	90	63	54
12	3	52	17	49
13	6	36	20	30
14	51	80	59	29
15	10	33	22	23
16	7	82	30	75
17	9	61	24	52
18	1	29	11	28
19	3	29	14	26
20	2	25	12	23
21	4	40	20	36
22	2	31	12	29
23	2	20	9	24
24	14	82	34	68
25	54	82	67	28
26	53	89	67	36
27	8	48	26	40
28	14	62	37	48
29	14	86	32	72
30	32	94	57	62
31	1	28	7	27
32	4	24	11	20
33	20	71	44	51
34	34	81	59	47
35	4	71	21	67
36	21	53	35	32
37	26	69	41	43
38	8	83	35	75
39	6	40	24	34
40	25	61	41	36
41	27	65	42	38
42	22	74	36	52
43	13	89	41	76
44	3	45	25	42
45	39	85	58	46
46	3	26	14	23
47	38	92	69	54
48	15	70	34	55
49	7	45	25	38
50	73	92	83	19

TABLE 5. Levels of interjudge agreement for articulation test scores according to position of sound in test word and phonetic classification.

Position	*% of agreement*	*Phonetic classification*	*% of agreement*
initial	79.97%	plosive	75.58%
medial	77.59%	fricative	67.42%
final	67.04%	affricative	75.22%
		aspirate	92.18%
		nasal	78.78%
		glide	67.27%
		blend	70.89%

TABLE 6. Intra- and interjudge agreement on specific error classification.

Judge	*% intrajudge agreement*	*% of interjudge agreement*								
		B	C	D	E	F	G	H	I	J
		(All entries are percentages)								
A	84	8	7	7	8	6	13	11	15	9
B	75		9	9	9	8	12	10	11	8
C	78			15	7	8	8	11	9	15
D	68				8	10	8	10	9	14
E	50					7	12	8	11	7
F	78						8	8	8	8
G	91							11	19	8
H	71								11	11
I	64									9
J	82									

TABLE 7. Interjudge agreement on classification of type of errors.

Error classification	*Lowest % of agreement*	*Highest % of agreement*
Emission	0%	27%
Indistinct	1%	18%
Substitution	2%	38%
Gross substitution	0%	33%
Omission	0%	26%

was made to see if one or more categories of error classification accounted for the low levels of agreement. These data, presented in Table 7, demonstrate that the judges had low levels of agreement on each of the five categories of error classification.

Discussion

RELIABILITY OF ARTICULATION ERROR JUDGMENTS. The findings demonstrate satisfactory reliability for evaluations of intelligibility of speech. Both intra- and interjudge percentages of agreement on articulation errors were found to be indicative of reliability. However, observation of the variability of the test scores creates a serious question concerning this assumption of reliability for interjudge evaluations.

Intrajudge variability was negligible compared to interjudge variability, for which considerable disparity was found (as indicated in Table 4). For example, Subject 2 was judged by one evaluator to have only 8 errors and by another to have 82 errors. The overall interjudge variability, which averaged 41.8 points, indicates that the scores obtained by two different evaluators should not be used comparatively. These findings seem to restrict the usefulness of articulation test scores and error descriptions between speech pathologists. Further, any comparisons with normative data on articulation errors must take into account the variability possible between examiners. Perhaps some of the discrepancies in the literature between studies describing the articulation characteristics of speech defective subjects may be accounted for by the factor of variability.

The averaging of the articulation scores for a number of judges provides a partial compensation for the wide variability. The greater validity of the averaged data was indicated by the high positive correlation of the mean articulation scores and the mean intelligibility ratings. The averaging of the judgments of several evaluators increased the reliability of articulation test scores, that is, the average values show better approximation to a "true" score, as defined here.

POSSIBLE CAUSES OF VARIABILITY. In many clinical and research settings it is not practical to obtain mean articulation error scores by having groups of judges make evaluations. Therefore, it is important to consider the possible causes of the variability demonstrated in this study to determine methods for increasing the reliability of speech data obtained from perceptual judgments of tape-recorded speech samples. The potential for higher agreement seems to be demonstrated by the high percentages of intrajudge agreement (see Table 2).

As shown in Table 5, the levels of interjudge agreement according to position in word and phonemic classification suggest that fricatives, glides and blends, and final sounds, all demonstrated a level of agreement lower than the average 74%. Consideration of this finding in redefinition of the criteria for judgment of errors might help to improve interjudge agreement without direct training of the judges.

Extremely low interjudge agreements were found for all categories used to classify errors. Percentages of interjudge agreement on specific error classifications, presented in Table 6, ranged from 6% to 19%. The higher level of agreement on the interjudge basis suggests that each of the judges employed, with some consistency, his own interpretation of the defined criteria for error classification. It is suggested that redefinition of criteria for classification of errors should give consideration to the interpretation utilized by the individual judges. Further, the effect of reduction of the number of categories of error classification should be evaluated.

The hypothesis that the severity of the articulation problem might relate to the degree of divergence among the judges was tested using the mean intelligibility rating as an indication of the severity of the disorder. The rank order correlation between mean intelligibility ratings and range of articulation test scores was only .19. This indicates that there is little if any relationship between intelligibility of the subject and variability of the articulation test scores assigned by the judges.

This study was conducted under specific conditions which must be considered in evaluating the findings: a) cleft palate subject, b) tape-recorded speech samples, c) written criteria for making judgments of articulation errors, and d) virtually no communication among judges. Of particular concern is whether tape recordings permit sufficiently audible cues for decisions about nasal emission, glottal stops and pharyngeal fricatives, and omissions. Under the conditions used in this study, the use of articulation test data seems to be limited for both clinical and research purposes except when grouped data are used as an index of speech deviations. In some clinical and research settings, this problem has been overcome by having judges trained to a given level of agreement on classification of articulation errors before data are obtained. Such training is not always possible and, when provided, probably introduces a bias to which subsequent evaluators must be trained if they are to obtain comparable information. Of importance to the speech clinician is the identification of factors necessary for obtaining greater reliability of articulation judgments. Standardization of criteria for evaluation of articulation errors by use of operant definitions and clear instructions for the identification and classification of errors would contribute to the exchange of more reliable information.

Summary

Ten speech pathologists evaluated tape-recorded articulation test responses and connected speech samples of fifty cleft palate subjects. The judges did not train together to reach a predetermined level of agreement. They followed written directions concerning error judgments. Their evaluations were studied to determine levels in intra- and interjudge agreement. The findings indicate satisfactory levels of agreement for judgments of intelligibility of connected speech samples. As determined by percentage of agreement, identification of articulation errors also appears to have satisfactory intra- and interjudge reliability. However, under the conditions of this study, the degree of variability which exists among examiners in identification and classification of articulation errors seriously impairs the reliability of this type of test score. Agreement on classification of error types was below the level of chance. The findings appear to limit the clinical and

research interpretations of articulation test scores. These data indicate that only the use of averaged or mean articulation scores would provide a satisfactory degree of reliability. However, the high degree of reliability for intrajudge evaluations indicates that reliability of interjudge agreements on identification and classifications of articulation errors might be improved by redefining and standardizing the criteria for each evaluation.

reprints: *Dr. Kenneth R. Bzoch*
Dept. of Communicative Disorders
College of Health Related Professions
University of Florida
Gainesville, Florida 32601

Acknowledgements. In addition to the second-named author, the following served as judges for the study: A. Berlin, D. Dickson, E. Hahn, J. Irwin, B. J. McWilliams, H. Morris, L. Schendel, C. Starr, and R. Van Hattum. The assistance of these judges, and the assistance of George J. Mouly, Ph.D., and C. Lee Philips, Ph.D., as research consultants, is gratefully acknowledged.

References

1. Brungard, Maude, Effect of consistency of /r/ and /s/ on gains made with and without therapy. Unpublished Ph.D. dissertation, The Pennsylvania State University, 1961.
2. Dorsey, H. A., The relationship between performance of kindergarten children on a 3 position test and a deep test of articulation. Unpublished Ph.D. dissertation, The Pennsylvania State University, 1959.
3. Henderson, Florence, Objectivity and constancy in articulation testing. *J. educ. Res.*, *31*, 348–356, 1937.
4. Jordan, E. P., Articulation test measures and listener ratings of articulation defectiveness. *J. speech hearing Res.*, *3*, 303–319, 1960.
5. Siegel, G. M., Experienced and inexperienced articulation examiners. *J. speech hearing Dis.*, *27*, 28–35, 1962.
6. Sommers, R. K., Florence G. Copetas, Delores C. Bowser, G. R. Fichter, Ann K. Furlong, F. E. Rhodes, and Z. G. Saunders, Effects of various durations of speech improvement upon articulation and reading. *J. speech hearing Dis.*, *27*, 54–61, 1962.
7. Winitz, H., Language skills of male and female kindergarten children. *J. speech hearing Res.*, *2*, 377–386, 1959.
8. Wright, H. N., Reliability of evaluations during basic articulation and stimulation testing. *J. speech hearing Dis.*, Monogr. Suppl. *4*, 19–27, 1954.