1-17-1977

Inter- and Intra-Judge Reliability of Four Articulation Tests

Lydia Pearl Evans

THE UNIVERSITY OF NEW MEXICO
ALBUQUERQUE, NEW MEXICO 87106

Policy on Use of Theses and Dissertations

Unpublished theses and dissertations accepted for master's and doctor's degrees and deposited in the University of New Mexico Library are open to the public for inspection and reference work. *They are to be used only with due regard to the rights of the authors.* The work of other authors should always be given full credit. Avoid quoting in amounts, over and beyond scholarly needs, such as might impair or destroy the property rights and financial benefits of another author.

To afford reasonable safeguards to authors, and consistent with the above principles, anyone quoting from theses and dissertations must observe the following conditions:

1. Direct quotations during the first two years after completion may be made only with the written permission of the author.

2. After a lapse of two years, theses and dissertations may be quoted without specific prior permission in works of original scholarship provided appropriate credit is given in the case of each quotation.

3. Quotations that are complete units in themselves (e.g., complete chapters or sections) in whatever form they may be reproduced and quotations of whatever length presented as primary material for their own sake (as in anthologies or books of readings) ALWAYS require consent of the authors.

4. The quoting author is responsible for determining "fair use" of material he uses.

This thesis/dissertation by Lydia Pearl Evans has been used by the following persons whose signatures attest their acceptance of the above conditions. (A library which borrows this thesis/dissertation for use by its patrons is expected to secure the signature of each user.)

NAME AND ADDRESS	DATE

3/1/73—1M

This thesis, directed and approved by the candidate's committee, has been accepted by the Graduate Committee of The University of New Mexico in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

INTER- AND INTRA-JUDGE RELIABILITY OF FOUR ARTICULATION TESTS

Title

Lydia Pearl Evans

Candidate

Department of Communicative Disorders

Department

Bernard Spolsky

Dean

January 17, 1977

Date

Committee

Wayne E. Secrist, Ph.D.

Chairman

John D. Black

Mary L. Bolton

INTER- AND INTRA-JUDGE RELIABILITY OF FOUR ARTICULATION TESTS

By
Lydia Pearl Evans
B.A., University of New Mexico, 1975

THESIS

Submitted in Partial Fulfillment of the
Requirements for the Degree of
Master of Science

in the Graduate School of
The University of New Mexico
Albuquerque, New Mexico

December, 1976

ACKNOWLEDGMENTS

I wish to express my sincere thanks to Dr. Wayne E. Swisher who gave me tremendous support through his time, patience and enthusiasm while directing this thesis. I will always be grateful for the experience of learning from him and working with him throughout my graduate study.

I also wish to thank Mary Bolton and Dr. John Lybolt for serving on my committee. Their time and help is truly appreciated.

Sincere appreciation is also given to Dr. L.H. Koopmans for his assistance in the statistical procedures used in this study. His time and interest were invaluable to me.

A special thanks goes to my husband, Graham, and daughter, Susan, for their continued support throughout my graduate work and writing of this thesis. I am sincerely grateful for their patience, love and understanding.

And finally, I would like to thank the children who were the subjects for this study, and their parents for their cooperation during scheduled testing.

INTER- AND INTRA-JUDGE RELIABILITY OF FOUR ARTICULATION TESTS

By
Lydia Pearl Evans

ABSTRACT OF THESIS

Submitted in Partial Fulfillment of the
Requirements for the Degree of
Master of Science
in the Graduate School of
The University of New Mexico
Albuquerque, New Mexico

December, 1976

INTER- AND INTRA-JUDGE RELIABILITY OF FOUR ARTICULATION TESTS

Lydia Pearl Evans
Department of Communicative Disorders
University of New Mexico, 1976

Assessment of articulatory productions is usually accomplished with picture stimuli to elicit responses from young subjects. The levels of scoring phonemes on articulation tests are two-way (correct/incorrect), four-way (correct, distortion, substitution and omission) or by narrow phonetic transcription. The stimuli presented range from scoring one phoneme in a one word response (Templin Darley Test of Articulation, 1960) to scoring one phoneme in different contexts across word boundaries (McDonald Deep Test of Articulation, 1964) to scoring several phonemes in a one word response (Goldman-Fristoe Test of Articulation, 1969) to scoring several phonemes as the subject repeats a story to the examiner (Goldman-Fristoe Test of Articulation, 1969). There is little available evidence indicating whether or not these differences in levels of stimuli effect inter- and intra-judge reliability. It was expected that a one word, one phoneme task (Templin Darley Test of Articulation, 1960) would be the most reliable test for both inter- and intra-judge reliability and that scoring several phonemes in a story sentence test (Goldman-Fristoe Test of Articulation, 1969) would be the least reliable for both inter- and intra-judge reliability.

PROCEDURES

Articulation Tests

The following four articulation tests were selected for this study because each of them contains a different level of stimuli, and they are commonly used. They include:

1.) Templin Darley Test of Articulation (1960);

2.) McDonald Deep Test of Articulation (1964);

3.) Goldman-Fristoe Test of Articulation (1969) - Sounds-in-Words Subtest; and

4.) Goldman-Fristoe Test of Articulation (1969) - Sound-in-Sentence Subtest.

Subjects

Six children with inconsistent multiple functional articulation errors were selected for this study. They were administered the four tests individually. Their responses were recorded on a Ravox high fidelity type A 77 Audio recorder. The recordings were made at a speed of 7.5 inches per second.

Judges

Seven Supervisors from the Department of Communicative Disorders at the University of New Mexico served as Judges for the study. They listened to the responses of the six subjects on four tests by playback from the Ravox Audio Recorder through Phonic Mirror headphones. The Judges listened to each test twice in a randomized order to determine intra-judge reliability. The tests were scored on a four way basis of correct, distortion, substitution and omission.

Results

A two-way analysis of variance was computed to determine whether or not differences in the Judges' scores were due to a difference in the level of stimuli presented in the four given tests. A difference in test reliability was determined, giving information in regards to a need for training of individuals who administer the tests that require a more difficult listening task. Information is also presented regarding which tests are most reliable in the diagnosis of speech articulation disorders and in the measurement of therapy progress.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER I

INTRODUCTION

Articulation tests are one of the most widely used and important tools in the field of Speech Pathology. Assessment of articulatory productions is usually accomplished with picture stimuli to elicit responses from young subjects. The levels of scoring phonemes on articulation tests are two-way (correct/incorrect), four-way (correct, distortion, substitution and omission) and by narrow phonetic transcription. The stimuli presented range from scoring one phoneme in a one word response (Templin Darley Test of Articulation, 1960) to scoring one phoneme in different contexts across word boundaries (McDonald Deep Test of Articulation, 1964) to scoring several phonemes in a one word response (Goldman-Fristoe Test of Articulation, 1969) to scoring several phonemes as the subject repeats a story to the examiner (Goldman-Fristoe Test of Articulation, 1969). To date there is no research indicating whether or not these differences in levels of stimuli effect inter- and intra-judge reliability.

CHAPTER II

REVIEW OF LITERATURE

Research in the area of articulation testing indicates that there are several important variables to consider when determining articulation performance. At least four variables are: variability of subject; variability of examiner; variability of test instrument; and variability due to the interaction of subject and examiner (Winitz, 1969; Siegel, 1962; Jordan, 1960; Sommers, et.al., 1959).

1. VARIABILITY OF SUBJECT

Determination of degree of subject variability has been discussed in terms of test-retest reliability. High reliability should result when a test is given at least twice within a short period of time without any therapy between the tests. Templin (1947) found high test-retest coefficients when examining 57 subjects by either the pictorial or oral imitative method within an eight day period. The test used was a 50 item non-diagnostic articulation test. The subjects ranged from 2 to 5 years of age. Siegel (1962) had two experienced and two inexperienced examiners test 22 mentally retarded children between the ages of 8 and 12 on the 50 item Templin-Darley screening test. The examiners retested the children after a one week interval and found the test scores to be remarkably constant. Winitz (1963) studied temporal (subject) reliability of 100 kindergarten children. The subjects were

tested on two occasions within a period of 6 to 12 days. While the total number of correct sounds did not significantly change during this period, intra-subject variability for some given sounds was indicated.

2. VARIABILITY OF EXAMINER

The reliability of an articulation test can also be measured in terms of inter- and intra-judge reliability. The effectiveness of a test score used for diagnostic purposes or therapy progress depends upon how well the examiner agrees with himself (intra) and how well he agrees with other examiners (inter).

a.) Intra-judge reliability

Henderson (1937) found 90% and 98% intra-reliability for two examiners when evaluating the same speech sample on two different occasions. Wright (1954) used three trained and experienced clinicians in the articulation testing for his study of reliability in a live testing situation. The results of intra-judge reliability indicated that the examiners had agreements of 74% and greater when judging on a discriminitive seven-point scale. Greatest reliability was indicated when tape recorded conditions were compared rather than when live and tape recorded situations were compared. Jordan (1960) reported that three examiners obtained contingency (C) coefficients of .804 for intra-judge reliability when listening to tape recorded responses of articulation testing on 5 children with several months elapsing between evaluations. The maximum C coefficient obtainable when using

four categories (correct, distortion, substitution and omission) is .866.

b.) Inter-judge reliability

Henderson (1937) found that in a live articulation testing situation three judges agreed in their scoring of two children on 80% of the test items. Wright (1954) also measured the consistency of each examiner with the other examiners in his study. Results indicated agreement between 77% and 87% from one examiner to another. Curtis and Hardy (1959) examined inter-judge reliability between two judges evaluating 195 responses from 30 children diagnosed as having an /r/ problem. The judges' reliability was 87% on a seven level classification system. Sommers, et.al. (1959) examined the reliability of six therapists who served as judges for their experiment in training parents. The percentage of agreement was determined when two live subjects were tested with one therapist presenting flash cards and all other therapists responding simultaneously. The mean percentage of agreement for all judges was 87.7%. Jordan (1960) reported that the inter-judge reliability obtained in his study was a C coefficient of .729 and .756 when one examiner's results were compared to the other two examiners' results. The maximum C coefficient obtainable was .866. Siegel (1962) found that reliability between experienced judges ranged from 95% to 99% but when experienced and inexperienced judges were compared, Siegel found significant differences in 16 of 18 comparisons.

3. VARIABILITY OF TEST INSTRUMENT

Test instrument variability is of most concern to this proposed study. While most articulation tests are designed to elicit spontaneous verbal responses to picture stimuli, some test pictures are often difficult or unrecognizable to subjects and are given imitatively. Several studies have investigated spontaneous vs. imitative presentations. Templin (1947) found that articulation scores did not differ for 100 children between the ages of two and six years when responses were elicited spontaneously to pictures, imitatively to examiner's verbal stimulus or to a combination of verbal and picture stimuli. However, studies by Snow and Milisen (1954) using the spontaneous method and imitative method found that subjects produced better responses to the imitative stimuli. Subjects for this study were 164 articulatory defective school children from grades 1, 2, 7 and 8. They were tested on a 65 item test with each response scored on a five point severity scale rather than the simple right versus wrong judgments made by Templin's examiners. Carter and Buck (1958) found that the 175 first grade children they tested demonstrated significant improvement of test results when responding to the imitation test. Siegel, Winitz and Conkey (1963) contrasted the imitative and spontaneous methods on a sample of 100 kindergarten children. The differences in correct responses of the two methods varied from 0% to 16.5%. The Chi-square analysis indicated that 8 of 40 sounds tested were produced correctly by more children in the imitative than in the spontaneous condition. These differences suggest that the imitative method

does significantly affect articulation responses of normal children. It would therefore appear that the most accurate means of surveying the speech of school age children is by the spontaneous (picture type) test. Smith and Ainsworth (1967) looked at conditions of pictures, imitative stimuli and to a combination of verbal and picture stimuli. They selected 40 first grade children demonstrating defective articulation and found more errors with picture stimuli and fewest errors with a combination of verbal and picture stimuli thus agreeing with the studies of Snow and Milisen (1954); Carter and Buck (1958) and Siegel, Winitz and Conkey (1963).

4. VARIABILITY DUE TO THE INTERACTION OF SUBJECT AND EXAMINER

Some of the most recent research in the area of articulation testing has developed over an interest in the interactions between the child and examiner. Winitz (1969) discusses the fact that a child who is unresponsive or taciturn may elicit different responses from an examiner than one who is quite responsive and talkative. A child with gross physical anomalies may encourage different behavior from the examiner than do more attractive children. Many other studies have been done which relate indirectly to the variabilities in responses due to subject and examiner interactions (Hartup, 1961; Stevenson, 1961; Winitz and Siegel, 1961; Siegel, 1962; and Winitz and Bellerose, 1963). A review of the literature revealed only one study directly concerned with this possible

variability. Shriberg (1971) investigated whether an adult's social behavior would influence children's articulation performance. The study was designed to look at Supportive (S) examiners versus Non-supportive (N) examiners by dividing 48 children into four groups having each group be administered the testing segment three different times. The four groups were given the tests as follows: Group 1 SSS; Group 2 NNN; Group 3 SNS; and Group 4 NSN sequences. The results indicate that the type of examiner behavior is important because Group 1, the only group not having a Non-supportive examiner, obtained the highest test-retest reliability score of 0.92 versus the scores of groups 2, 3, and 4 which were 0.67, 0.76 and 0.62 respectively. However, the children in Group 1 did not have a higher performance score than Group 2 which had total Non-supportive examiners. In fact, some children from the SNS and NSN groups may have actually scored higher in the Non-supportive conditions. This presents a perplexing problem. The need for further study of interpersonal variables was indicated in this study.

There are additional studies that can be discussed which have not been reviewed. Ruth Becky Irwin (1970), conducted an experiment to determine the consistency with which the Speech Pathologist makes his judgments of articulation productions of consonants. She had 65 students in a course concerned with clinical methods and practices in speech therapy at Ohio State University serve as judges for the study. The students were in two groups with both groups viewing the tape of six children

on the Ohio State University Audio-Visual Test (Form A) on three different occasions within a time period of 2 months elapsing for Group 1 and 16 months elapsing for Group 2. The test was scored on the basis of correct and incorrect. Comparison of test results of both groups on the three occasions revealed 86 to 87% agreement. Inconsistency of judgments was higher when misarticulations were being evaluated as opposed to evaluation of correct productions. The highest consistency of judgments occurred when articulatory productions of consonants were evaluated in words in which no misarticulations occurred. Irwin concluded her study by expressing a need for special attention to be given to the study of incorrect productions of sounds since the consistency of agreements was generally much poorer in identification of misarticulations than for correct productions of consonants. Shriberg and Swisher (1972) looked at the effect of several levels of stimuli on reliability of a group of judges. The important results of this study were that agreement figures were higher for two-way scoring than in four-way scoring and that single sound scoring generally obtained higher agreement than in multiple sound scoring.

The above studies indicate another variable that has not been directly evaluated. There are many commercially available articulation tests. Because these commercial tests contain different levels of stimuli to be presented and analyzed, there may well be a difference in the reliability of the test results when comparing these tests. A more intensive

study of inter- and intra-judge reliability on several commonly used articulation tests which encompass a variety of levels of stimuli would give information in regards to:

1.) Articulation tests that would be most reliable for less sophisticated examiners;

2.) effective articulation test training;

3.) possible differences between examiners in test results of the same client; and

4.) appropriate types of therapy programming in terms of level of stimuli presented in daily lesson plans to insure the most reliable assessment of therapy progress.

STATEMENT OF PROBLEM

It is the thesis of this study that differences in levels of stimuli in articulation tests will show a difference in terms of inter- and intra-judge reliability. On the basis of the results from the studies by Irwin (1970) and Shriberg and Swisher (1972), it is expected that a one word, one phoneme task (Templin Darley Test of Articulation, 1960) will be the most reliable test for both inter- and intra-judge reliability and that scoring several phonemes in a story sentence test (Goldman-Fristoe Test of Articulation, 1969) will be the least reliable for both inter- and intra-judge reliability.

CHAPTER III

PROCEDURES

Articulation Tests

The following four articulation tests were selected for this study because each of them contains a different level of stimuli. These are published tests all commonly used. They include:

1.) Templin Darley Test of Articulation (1960). The Diagnostic Test portion of the Templin Darley Test will be used for this study. It contains a complete battery of 141 items to obtain a detailed description and evaluation of a child's articulation. This test is designed to score one phoneme per picture stimulus.

2.) McDonald Deep Test of Articulation (1964). This test is designed to look at one of thirteen given phonemes in a variety of context and in different syllable roles including the arresting and releasing positions.

3.) Goldman-Fristoe Test of Articulation (1969). The Sounds-in-Words Subtest of the Goldman-Fristoe Test will be used for this study. The chief efforts in this subtest were directed toward selecting stimulus items that would elicit the desired response spontaneously and toward reducing the time required for the test by examining more than one phoneme per word.

4.) Goldman-Fristoe Test of Articulation (1969). The

Goldman-Fristoe Sound-in-Sentence Subtest was developed to provide a systematic means of assessing speech sound production in a complex context similiar to that found in conversational speech. As the subject repeats a story, the examiner scores several phonemes per sentence.

Subjects

Six children with inconsistent multiple functional articulation errors will be selected for this study from a group of children diagnosed as having inconsistent multiple functional articulation errors from the University of New Mexico Speech and Hearing Clinic. The children selected will be male or female, preferably between the ages of 6 and 8 who have hearing within normal limits and are monolingual English speaking.

Presentation of Tests to Subjects

The four tests will be administered by this investigator to each subject individually. The subjects will speak into a Shure Brothers Model 545 Unidym III microphone. Recording will be made on a Ravox high fidelity type A 77 Audio Recorder with automatic gain control to insure constant sound control on Scotch high grade, low noise audio recording tape. The recordings will be made at a speed of 7.5 inches per second. A total of 1452 responses will be accumulated by having the four subjects respond to the above mentioned articulation tests.

Judges

Six Speech Pathology Supervisors at the University of

New Mexico will be the judges for this study. They were chosen because their training and experience has placed them in a position requiring that they make articulation judgments on a regular basis in a training clinic. Each judge will be required to pass a pure-tone audiometric screening test at 20 dB for the following frequencies in order to participate in the study: 250, 500, 1000, 2000, 4000 and 8000.

Presentation of Tests to Judges

The judges will listen to the responses of six subjects on four tests by playback from the Ravox Audio Recorder through Phonic Mirror headphones with adjustable air filled ear cuffs at a loudness level comfortable to all listeners at a given time. The judges will listen to each test twice to establish an intra-judge reliability score. The order of test will be randomized to control for memorizing a given voice and errors to that voice. Before the listening task begins, each judge will be given the following list of procedures concerning the listening task:

1. There are four Master Tapes of the 48 tests you will be scoring. The tests will be scored in order from 1 to 48. The sequence number of each test is given in the top right hand corner of each response sheet.
2. Do Not repeat any portion of any test.
3. Score every sound produced. If a particular sound was not elicited, leave it blank and go on to the next response.
4. Scoring will be done in the following manner:
   a. Any correct response should be indicated by a check (✓) or plus (+) sign;
   b. If a sound is substituted such as "tup" for "cup",

write a /t/ on the response sheet. Anything other than one complete phoneme substitution for another is a distortion, such as a lateral /s/ production for the word "soup".

c. Write "OM" on the response sheet for any sound that is omitted entirely.

d. Write "Dis" on the response sheet when a given sound is distorted.

e. Indicate an addition by writing the additional sound plus the correct sound on the response sheet.

5. Listening should be done in a maximum of one hour per session with at least a 15 minute break between the 1 hour sessions.

Scoring of Tests

Scoring of the tests will be on a four-way basis of correct, distortion, substitution and omission. The judges will listen to the tests for a maximum of one hour per session. The total listening time will be approximately six hours, requiring six listening sessions per judge.

TREATMENT OF DATA

Intra-judge reliability

The judgments for each response of each test by each judge will be transferred to a scoring matrix (Form 1, Appendix A). Intra-judge reliability will be calculated by dividing the number of times that each judge agreed with himself by the total number of times he could have agreed with himself and transferred to a scoring matrix (Form 2, Appendix A). For example, if after scoring the Templin Darley twice the differences for Judge 1 were 6, then 141 (total number of agreements possible) would be divided into 135 (total number of times

Judge 1 agreed with himself) getting a 96% intra-judge reliability score. A mean reliability score for each judge will be determined by adding the four intra-judge reliability scores and dividing by four.

Inter-judge reliability

The number of times each judge agreed with the majority decision represented by each judge on listening trial 1 will be transferred to a scoring matrix (Form 1, Appendix A). Inter-judge reliability will be calculated by dividing the number of times that each judge agreed with every other judge by the total number of times that they could have agreed. For example, if there was a difference of 8 agreements between Judge 1 and the majority decision, then 141 (total number of agreements possible on the Templin Darley) would be divided into 133 (total number of times Judge 1 agreed with the majority decision on the Templin Darley) getting a reliability score of 92%. This percentage will be determined by comparing each judge to the majority decision on all four tests for each subject and transferred to a scoring matrix (Form 2, Appendix A). The mean inter-judge reliability score for every judge will be determined by adding all six inter-judge reliability scores and dividing by six.

Two-way Anova

After calculating the mean intra- and inter-judge reliability scores for each judge, the Two-way Analysis of Variance will be used to determine whether there are any

statistically significant differences among the judges between the tests and the subjects (Two-way Analysis of Variance, Hays, 1963).

CHAPTER IV

RESULTS

The results of this study are presented in terms of inter-judge and intra-judge reliability data. The inter-judge reliabilities were based on the number of agreements with the majority decision of all judges divided by the number of possible agreements on each of four articulation tests. The intra-judge reliabilities were based on the number of times each judge agreed with himself on all responses divided by the number of times he could have agreed with himself on all four tests. The four articulation tests used for the study will be represented in the Tables and Figures in this chapter as follows: 1 - Goldman-Fristoe Test of Articulation Sounds in Words Subtest, 2 - Goldman-Fristoe Test of Articulation Sounds in Sentences Subtest, 3 - McDonald Deep Test, and 4 - Templin Darley Test of Articulation.

Inter- and intra-judge reliabilities will be discussed by presenting the mean inter- and intra-judge reliabilities, averaged over subjects, for all judges and all tests; the mean reliabilities, averaged over judges, for all subjects and all tests; and the standard deviations of these means.

The SPSS (statistical package for the social sciences) subprogram ANOVA was computed at the University of New Mexico Computer Center to determine the significance of the mean scores discussed. The ANOVA program permitted performance, jointly, of three Two-Way Analyses of Variance which examined Test/Subject interactions, Test/Judge interactions and Subject/Judge inter-

actions. These results will also be discussed in both inter- and intra-judge reliability sections.

Inter-judge reliability

Table 1 presents the mean reliabilities, averaged over all subjects' responses and standard deviations for four tests and all six judges and the standard deviation of each mean. The reliabilities range from 88% for test 1 to 71% for test 3. The standard deviations for tests 1, 2, and 4 are .03, .05 and .04, respectively, and are much lower than the standard deviation of test 3 which is .19. Test 3 had the lowest reliability and the highest standard deviation.

JUDGE	TEST			
	1	2	3	4
1	.81	.72	.48	.77
2	.88	.76	.51	.85
3	.90	.85	.70	.86
4	.89	.80	.85	.78
5	.89	.80	.97	.87
6	.90	.74	.74	.83
MEAN	.88	.78	.71	.83
SD	.03	.05	.19	.04

Table 1. Mean inter-judge reliabilities, averaged over all subjects' responses and standard deviations for four tests and all six judges.

Figure 1 illustrates the variations of the mean scores for all subjects as determined by each judge for all four tests. As can be seen, the widest excursion is found for test 3 with the mean scores ranging from 48 to 98 percent, while tests 1, 4 and 2 mean scores ranged from 81 to 90 percent, 77 to 87 percent and 72 to 85 percent, respectively.

Figure 2 illustrates the mean inter-judge reliability scores of all judges on each of the four tests. Again the wide range of mean percentages can be seen for test 3.

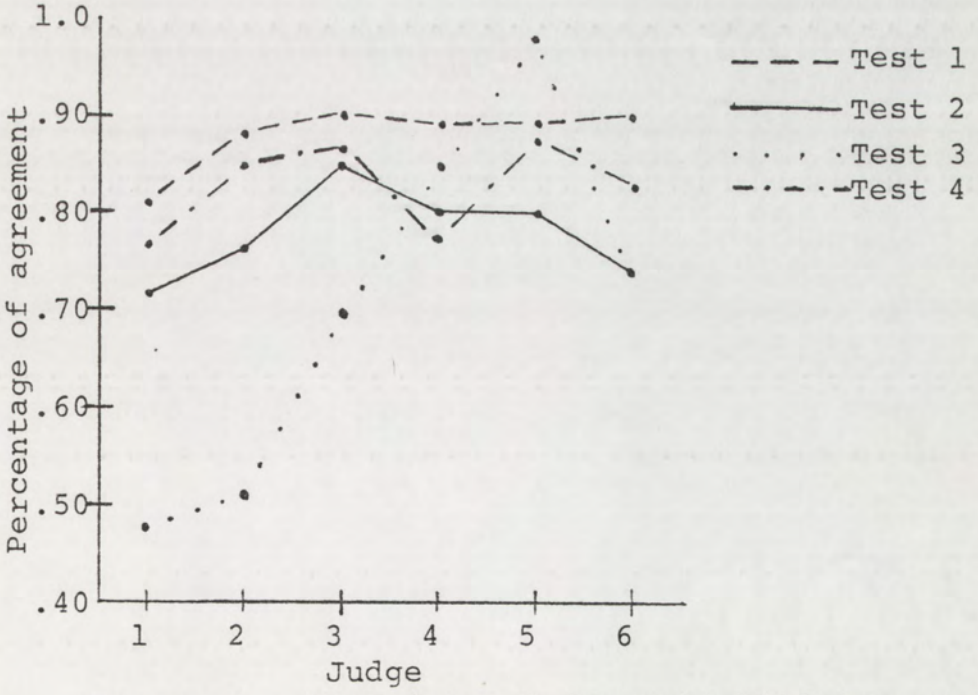

Figure 1. Mean inter-judge reliabilities on four articulation tests as determined for six judges.

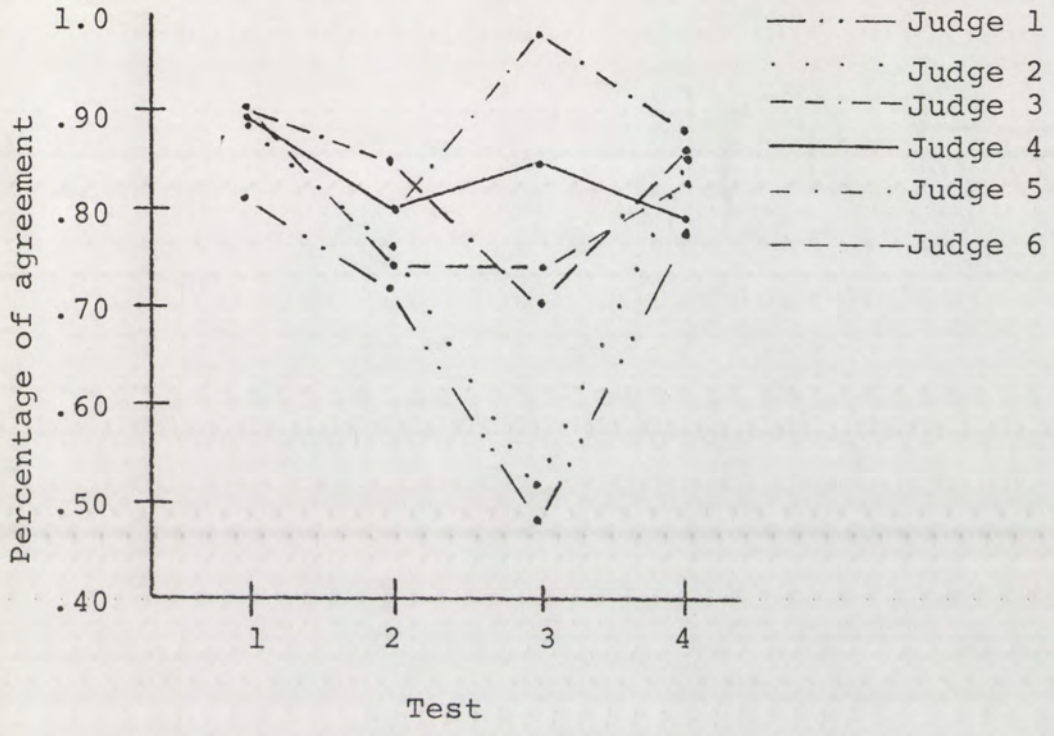

Figure 2. Mean inter-judge reliabilities as determined by six judges on four articulation tests.

Table 2 presents the mean inter-judge reliability scores averaged over judges for each subject on the four articulation tests and the standard deviations of all means. The range of reliabilities is from 88 percent for test 1 to 72 percent for test 3. The standard deviations for tests 1, 2 and 4 are .05, .08 and .04, respectively. The standard deviation of test 3 is .26 which is much higher than the other three standard deviations. Again test 3 has the lowest mean percentage of all tests.

The graphical representation of Table 2 can be seen in Figure 3 which illustrates the mean inter-judge reliabilities for all judges' agreements for the six subjects on the four

articulation tests. Test 3 shows a range of reliabilities from 33 percent to 96 percent while tests 1, 2 and 4 show ranges of 81 to 94 percent, 70 to 90 percent and 70 to 89 percent, respectively.

SUB-JECT	TEST			
	1	2	3	4
1	.91	.84	.68	.82
2	.85	.78	.87	.78
3	.94	.90	.96	.89
4	.81	.70	.51	.85
5	.90	.74	.94	.84
6	.85	.70	.33	.80
MEAN	.88	.78	.72	.83
SD	.05	.08	.26	.04

Table 2. Mean inter-judge reliabilities and standard deviations for each of four tests for all six subjects.

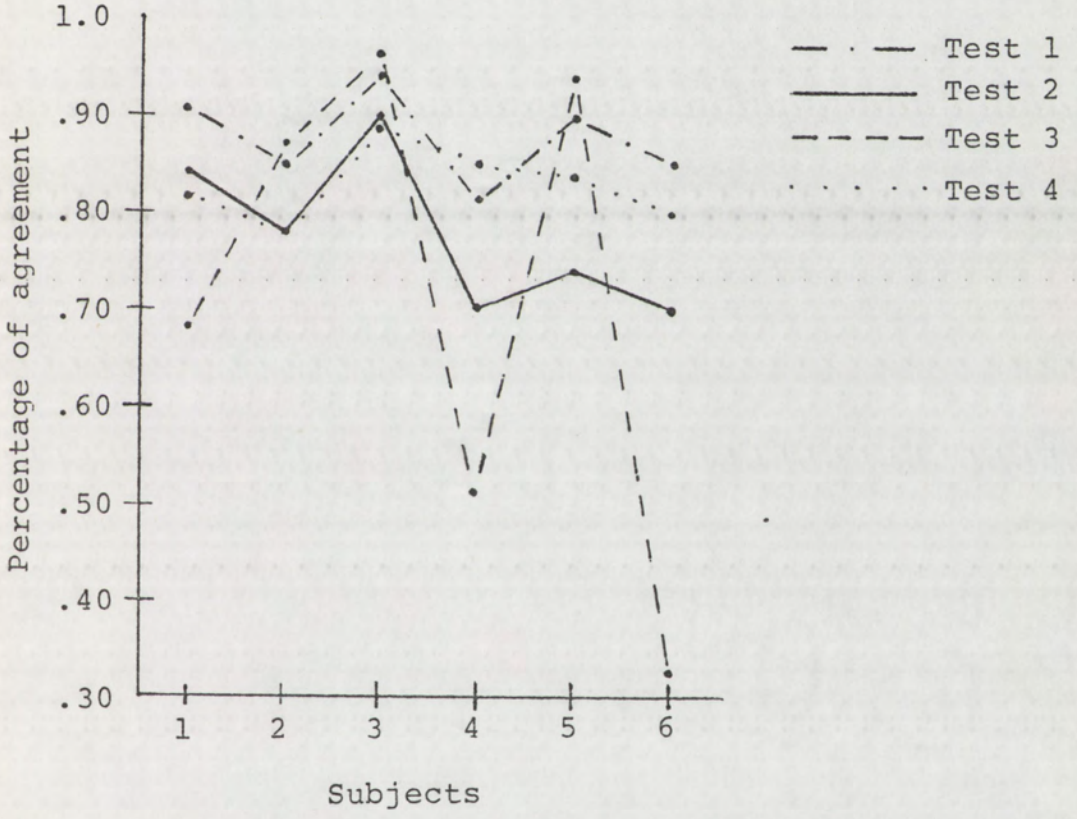

Figure 3. Mean inter-judge reliabilities for six subjects on four articulation tests.

Figure 4 presents the mean scores of Table 2 graphically by giving the mean inter-judge reliabilities for the six subjects' on the four articulation tests. Again the wide range of reliabilities for test 3 can be seen.

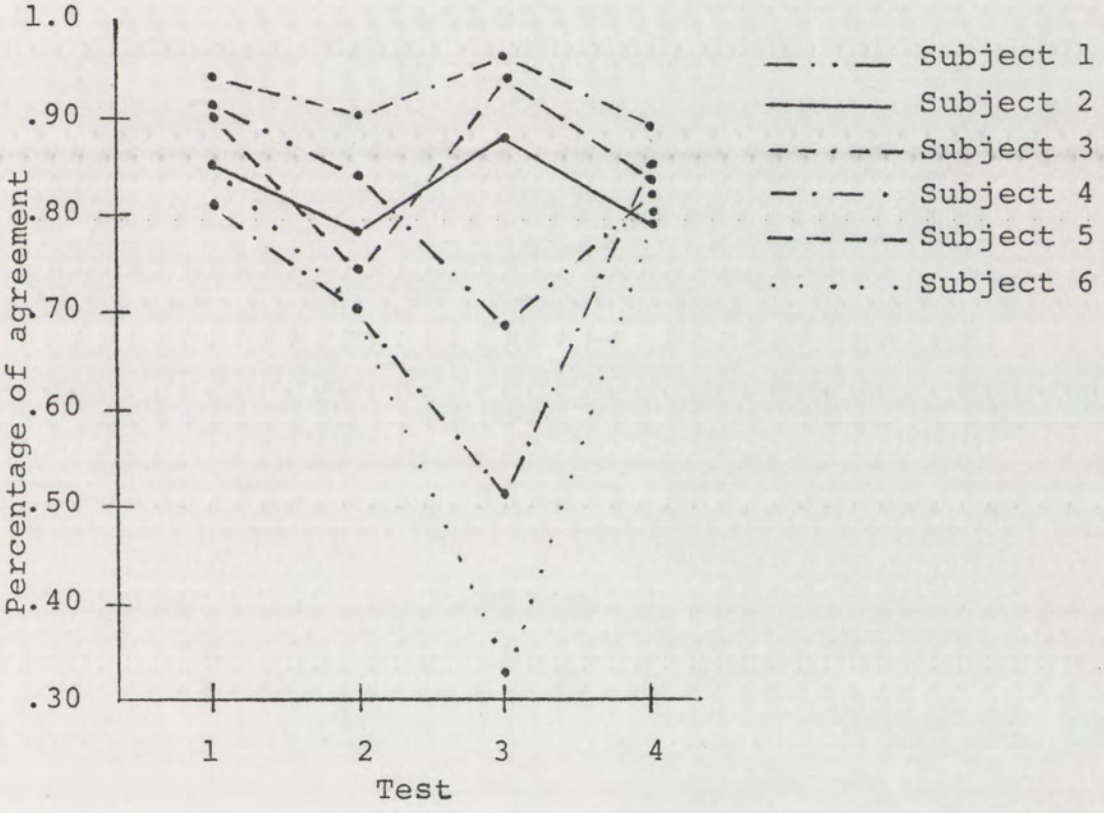

Figure 4. Mean inter-judge reliabilities for the six subjects on four articulation tests.

In order to determine whether the interactions of the mean scores illustrated in the above-mentioned Figures were significant, an SPSS program providing, jointly, the three necessary Two-Way Analyses of Variance was performed. The results are given in Table 3. Inspection of this table reveals

significant test/subject interactions at the .001 level of significance and test/judge interactions at the .001 level of significance. These significances indicate that the tests did affect the subjects' responses as judged by the listeners. However, the non-significance of the subject/judge interactions for the ANOVA indicates that differences in reliability from one judge to another are about the same from subject to subject.

Another Analysis of Variance was computed at the Statistics Department at the University of New Mexico using data transformations designed to validate the usual Analysis of Variance hypotheses. These results were compared to those presented in Table 3. It was found that the same F-values were obtained in both analyses.

These analyses indicate that inter-judge reliability test differences were obtained and that the judges' responses to the subjects' responses were affected by the tests. However, the most reliable and least reliable tests were not the Templin Darley Test of Articulation and the Goldman-Fristoe Test of Articulation Sounds in Sentences Subtest as predicted. Because of the relative closeness of the two most reliable tests of this study (Goldman-Fristoe Sounds in Words and Templin Darley) and the least reliable tests of the study (Goldman-Fristoe Sounds in Sentences and McDonald Deep), a meaningful rank ordering cannot be given. A similiar study may indeed produce results that would reverse the two most reliable tests and the two least reliable tests.

Source of Variation	Sum of Squares	DF	Mean Square	F	Signif of F
Main Effects	4.018	13	0.309	7.778	0.001
Test	1.169	3	0.390	9.803	0.001
Subject	1.876	5	0.375	9.441	0.001
Judge	0.974	5	0.195	4.902	0.001
2-Way Interactions	6.118	55	0.111	2.799	0.001
Test/Subject	3.960	15	0.264	6.644	0.001
Test/Judge	1.587	15	0.106	2.663	0.001
Subject/Judge	0.570	25	0.023	0.574	0.999
Explained	10.136	68	0.149	3.751	0.001
Residual	8.703	219	0.040		
Total	18.839	287	0.066		

Table 3. Results of the Analysis of Variance performed on inter-judge reliabilities.

Intra-judge reliability

Table 4 presents the mean percentages of all subjects' responses on the four articulation tests as determined by all six judges and the standard deviation of each mean. The reliabilities ranged from 88 percent for test 1 to 80 percent for test 2. The standard deviations for tests 1, 2 and 4 are .04, .04 and .06, respectively. The standard deviation of test 3 is .10. Although the reliability mean of test 3 is 83 percent, the standard deviation indicates that the judges' means varied from this mean to a greater extent than the means for tests 1, 2 and 4.

Figure 5 illustrates by judge the variations between the mean scores presented in Table 4. The mean variations of tests 1, 2 and 4 were 81 to 91 percent, 74 to 85 percent, and 73 to 89 percent respectively. The range of means for test 3 was 66 to 95 percent.

Figure 6 illustrates the same results by subject. The range of mean scores presented in Table 4 is demonstrated more clearly.

JUDGE	TEST			
	1	2	3	4
1	.81	.74	.66	.73
2	.85	.76	.79	.81
3	.90	.83	.87	.89
4	.89	.78	.83	.84
5	.89	.85	.86	.85
6	.91	.83	.95	.89
MEAN	.88	.80	.83	.84
SD	.04	.04	.10	.06

Table 4. Mean intra-judge reliabilities, averaged over subjects, and standard deviations for all tests and all judges.

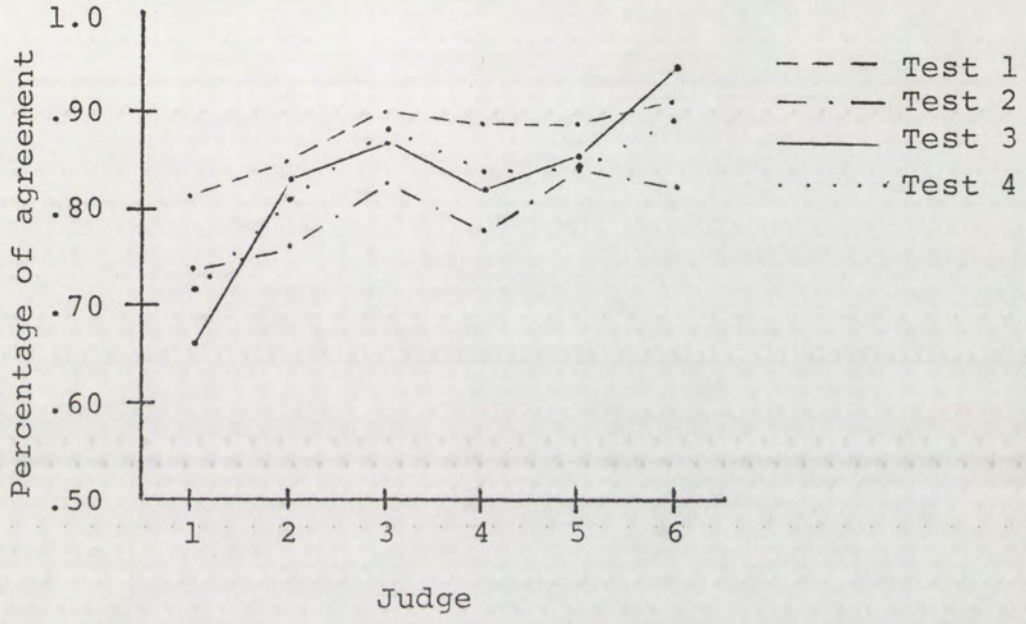

Figure 5. Mean intra-judge reliabilities on four articulation tests for all six judges.

Table 5 presents the mean intra-judge reliability scores, averaged over judges, for each subject on the four articulation tests and the standard deviation of each mean. The range of reliabilities is from 87 percent for test 1 to 80 percent for test 2. The standard deviations for tests 1, 2 and 4 are .05, .07 and .03, respectively. The standard deviation of test 3 is again the highest at .15. This indicates that the major variation of reliability among judges again occurs for test 3. Although 83 percent reliability of test 3 is higher than 80 percent reliability of test 2, there is much greater variability of judges' reliabilities for test 3 than test 2.

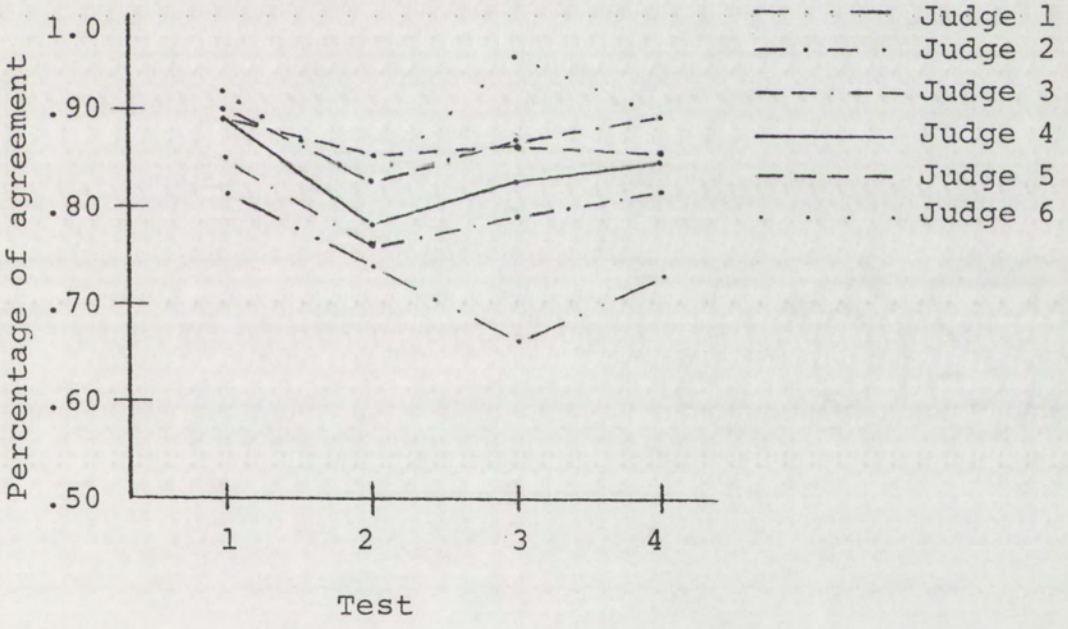

Figure 6. Mean intra-judge reliabilities on four articulation tests for each of six judges.

One graphical representation of Table 5 can be seen in Figure 7 which illustrates the mean intra-judge reliabilities for all judges' agreements for the six subjects on the four articulation tests. Test 3 presents a range of reliabilities from 61 to 98 percent while tests 1, 2 and 4 show ranges from 82 to 94 percent, 70 to 88 percent and 81 to 90 percent, respectively.

Figure 8 presents the mean scores of Table 5 graphically by giving the mean intra-judge reliabilities on the four articulation tests for each subject. Again, the widest range of reliabilities can be seen for test 3.

SUB-JECT	TEST			
	1	2	3	4
1	.89	.85	.70	.83
2	.86	.79	.92	.83
3	.94	.88	.98	.90
4	.82	.75	.61	.82
5	.87	.70	.94	.81
6	.86	.81	.81	.82
MEAN	.87	.80	.83	.84
SD	.05	.07	.15	.03

Table 5. Mean intra-judge reliabilities and standard deviations of all judges' agreements for each of four tests for all six subjects.

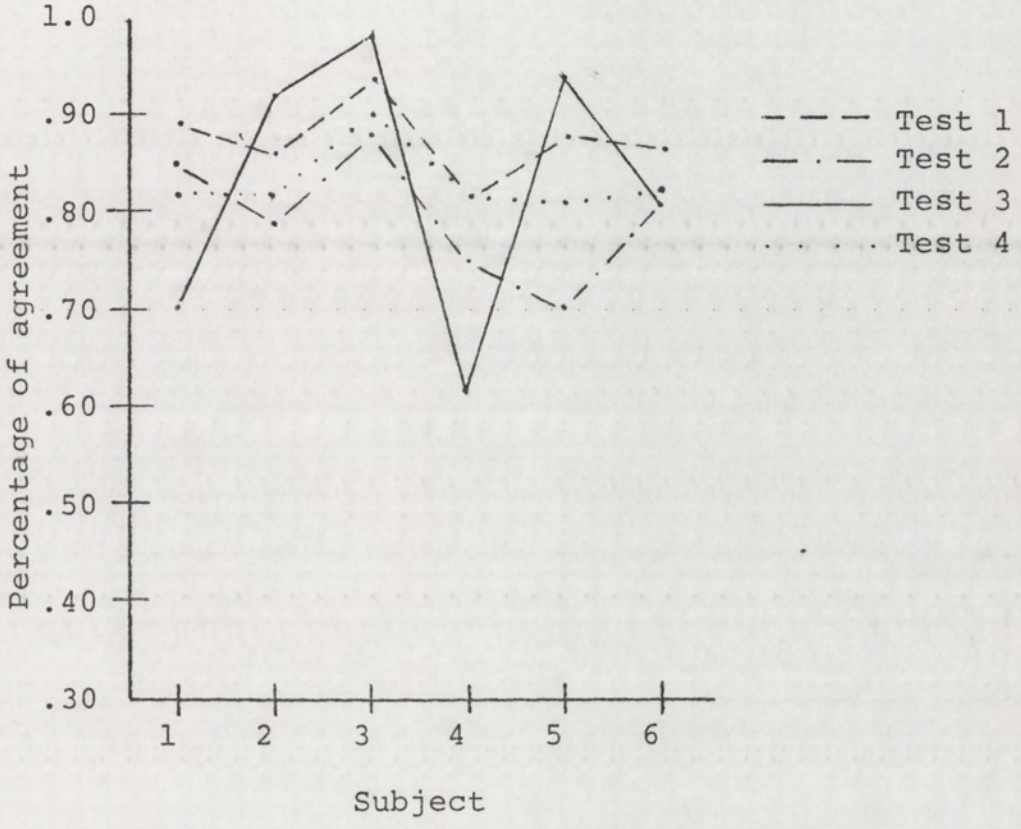

Figure 7. Mean intra-judge reliabilities by subjects for four articulation tests.

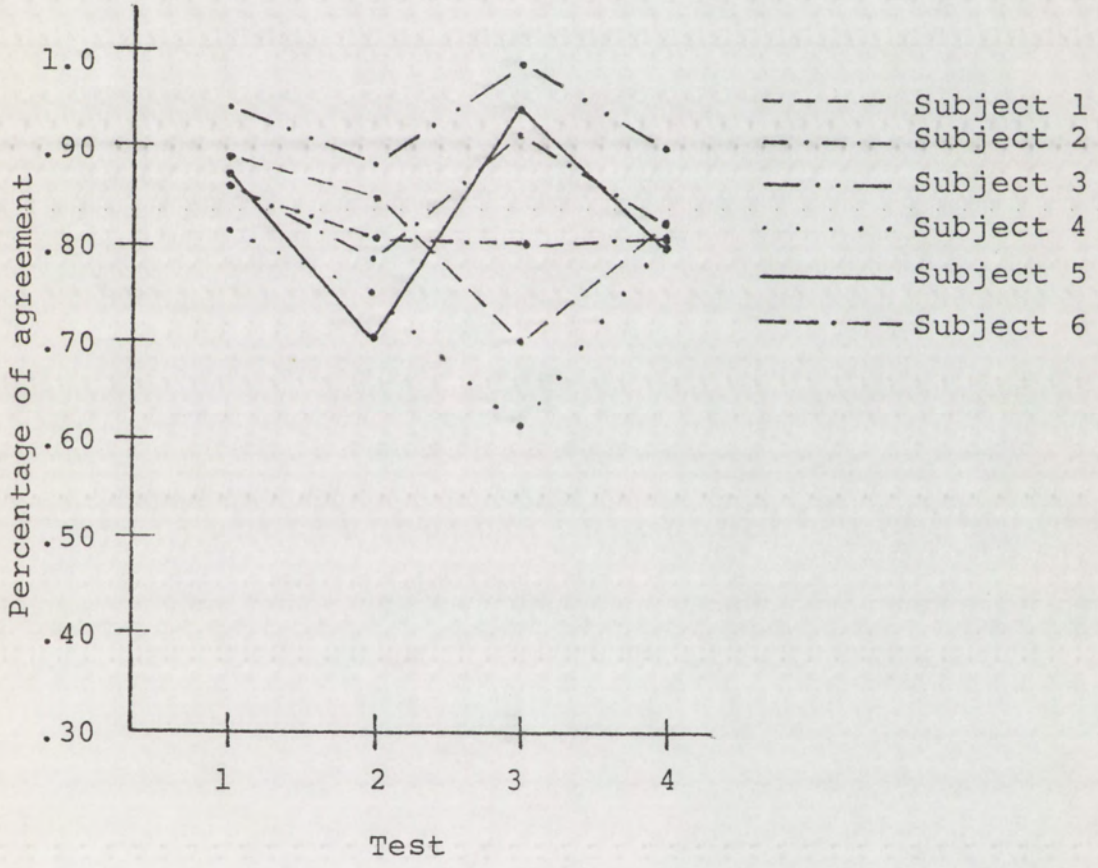

Figure 8. Mean intra-judge reliabilities on four articulation tests for each subject.

Table 2 presents the mean inter-judge reliability scores averaged over judges for each subject on the four articulation tests and the standard deviations of all means. The range of relia-reliabilities is from 88 percent for test 1 to 72 percent for test 3. The standard deviations for tests 1, 2 and 4 are .05, .08 and .04, respectively. The standard deviation of test 3 is .26 which is much higher than the other three standard deviations. for the six subjects on the four articulation tests. Test 3 shows a range of reliabilities from 33 percent to 96 percent while tests 1, 2 and 4 show ranges of 81 to 94 percent, 70 to 90 percent and 70 to 89 percent, respectively.

These analysis indicate that intra-judge reliability test differences were obtained and that the judges' responses to the subjects' responses were affected by the tests. However, the most reliable and least reliable tests were not the Templin Darley Test of Articulation and the Goldman-Fristoe Test of Articulation Sounds in Sentences Subtest as predicted. Because of the relative closeness of the two most reliable tests of this study (Goldman-Fristoe Sounds in Words and Templin Darley) and the least reliable tests of the study (Goldman-Fristoe Sounds in Sentences and McDonald Deep), a meaningful rank ordering cannot be given. A similar study may indeed produce results that would reverse the two most reliable tests and the two least reliable tests.

Source of Variation	Sum of Squares	DF	Mean Square	F	Signif Of F
Main Effects	3.460	13	0.266	7.513	0.001
Test	0.577	3	0.192	5.432	0.001
Subject	1.247	5	0.249	7.039	0.001
Judge	1.636	5	0.327	9.236	0.001
2-Way Interactions	2.383	55	0.043	1.223	0.158
Test/Subject	1.339	15	0.089	2.519	0.002
Test/Judge	0.240	15	0.016	0.452	0.999
Subject/Judge	0.804	25	0.032	0.908	0.999
Explained	5.843	68	0.086	2.425	0.001
Residual	7.758	219	0.035		
Total	13.601	287	0.047		

Table 6. Results of the Analysis of Variance performed on intra-judge reliabilities.

CHAPTER V

DISCUSSION

The purpose of this study was to examine the inter- and intra-judge reliability of six listeners as they responded to six children with multiple, functional articulation errors on four articulation tests. The four tests used for the study were the following: 1) Goldman-Fristoe Test of Articulation Sounds in Words Subtest, 2) Goldman-Fristoe Test of Articulation Sounds in Sentences Subtest, 3) McDonald Deep Test, and 4) Templin Darley Test of Articulation. Each of these tests contains a different level of stimuli. These stimuli range from scoring one phoneme in a one word response (Templin Darley Test of Articulation) to scoring one phoneme in different contexts across word boundaries (McDonald Deep Test of Articulation) to scoring several phonemes in a one word response (Goldman-Fristoe Test of Articulation) to scoring several phonemes as the subject repeats a story to the examiner (Goldman-Fristoe Test of Articulation).

It was hypothesized that a one word, one phoneme task (Templin Darley Test of Articulation) would be more reliable in terms of inter- and intra-judge reliability than the other three tests. It was also hypothesized that the least reliable test would be the Sounds in Sentences Subtest of the Goldman-Fristoe Test of Articulation because it requires the listener to judge sounds in words given in random order from subject to subject.

This chapter will discuss the results found for

inter-judge reliability and intra-judge reliability. In addition, the examiner will discuss the implications drawn from the results of this study in terms of articulation test training necessary and suggestions for future research.

Inter-judge reliability

The highest inter-judge reliability mean of all subjects' responses on four articulation tests as determined by all six judges was 88 percent for the Goldman-Fristoe Test of Articulation Sound in Words Subtest. This test also had the smallest standard deviation, indicating that all of the judges' scores for every subject were very close to this mean. In comparison to the other tests, the Templin Darley Test of Articulation had a very close mean reliability of 83 percent; the Sounds in Sentences Subtest portion of the Goldman-Fristoe Test of Articulation was lower with a reliability percentage of 78; and the least reliable was the McDonald Deep Test with a 71 percent reliability score.

The Two-Way Analysis of Variance data indicated that the tests did affect the judges' scoring of the subjects' responses. Depending on which test a subject was given, a different analysis of the articulation skills was determined for that individual. Inter-judge reliability is extremely important when considering the fact that a given articulation client may have more than one clinician.

One reason that the Goldman-Fristoe Test Sounds in Words Subtest and the Templin Darley Test of Articulation were the most reliable may be that the judge has only to listen to one

word at a time and the rate of deliverance is governed by the judge. This would agree with the study reported by Irwin (1970) in which two separate groups of speech pathology students were more consistent in making judgments of misarticulated sounds in words than in phrases or trios. The length of the Templin Darley in comparison to the Goldman-Fristoe may be a factor responsible for the differences in their reliability scores. During the listening portion of this study, several judges commented on the length of the Templin Darley and said that they usually do not present all 141 items of the test, but rather selected items. This attitude of the judges may have affected their scores for the test. There is a 4:1 ratio of Templin Darley response items to the Goldman-Fristoe response items. The length of a test in terms of response items may be an important variable to consider in future research.

Although the least reliable test in the study was not the Sounds in Sentences Subtest of the Goldman-Fristoe Test of Articulation as predicted, its reliability was 78 percent, which was only 7 percent higher than the least reliable test. Administration of the Sound in Sentences Subtest of the Goldman-Fristoe Test of Articulation requires that the listener be able to hear a story repeated by the subject and score the key word responses. The ordering of key words in the story may vary from subject to subject, therefore, the listener must know all key words and the sound or sounds to be judged in that word. In this test situation the subject governs the order of key words given and the rate at which these words will be delivered.

The Sounds in Sentences Subtest of the Goldman-Fristoe Test of Articulation was designed to provide a conversational speech sample. The results of studies by Irwin (1970) and Shriberg and Swisher (1972) indicated that as a stimulus item increases with difficulty, such as from single sound scoring to multiple sound scoring, the reliability decreases. In administering the Sounds in Sentences Subtest of the Goldman-Fristoe Test of Articulation, the listener is not only judging multiple sounds in sentences given by the subject, but is also expected to judge two sounds within some words.

The results indicated that the McDonald Deep Test was the least reliable test in the study with a 71 percent inter-judge reliability mean. The standard deviation of this 71 percent reliability mean was the largest of all tests at .19, indicating that there was a wide range of scores as given by the judges. Listening to the same sound in different contexts across word boundaries is apparently a more difficult task than listening to just one sound per word. Upon examination of actual test responses of each subject by each judge, it was revealed that the higher inter-judge reliabilities came when a subject produced the same sound substitution for all different contexts. According to the judges' responses, subjects 2, 3 and 5 produced substitutions of /f/θ/, /w/r/ and /w/h/, respectively. The other three subjects produced inconsistent misarticulation of the target phoneme and the reliabilities were much lower. For example, the judges' responses to subject 6 ranged from correct responses to substituted responses, to distorted

responses to omitted responses for different contexts with the same target phoneme on the McDonald Deep Test. The test was designed to find the contexts in which the error phoneme is correctly articulated. However, the results of this study indicated that when a subject does change production of the target phoneme in different contexts the reliability decreases.

This decrease in reliability may be due to the fact that judges form biases or expectations for hearing a certain production of the target sound and when this production changes, it is difficult to determine the response given. For example, when a subject with a /w/r/ substitution produces a response that is not a /w/ in a given context, the judge may have difficulty judging that response. This would explain the variety of correct, distorted, substituted and omitted responses given during this study when subjects did change their production of a given phoneme.

Intra-judge reliability

The results of intra-judge reliability indicates that the Goldman-Fristoe Test of Articulation Sounds in Words Subtest was the most reliable at 88 percent, which is the mean score of all judges' responses for every subject. In comparision, all of the mean reliabilities were close to test 1, with tests 2, 3 and 4 having reliabilities of 80, 83 and 84 percent respectively. The standard deviations of all means were .06 and above for every test except the McDonald Deep Test which had a standard deviation of .10, indicating that the range of the judges' responses

were greater for the McDonald Deep Test than the other three tests. The actual test responses for each subject by each judge revealed that the range of judges' responses were due to subjects 1, 4 and 6. For all three subjects, the judges were inconsistent with themselves from listening trial 1 to listening trial 2. These inconsistencies were omission versus substitution or correct responses for subject 1, correct versus distortion or substitution for subjects 4 and 6. For subjects 2, 3 and 5, the mean percentages of all judges' agreements for the McDonald Deep Test were 92, 98 and 94 percent, respectively. Actual test responses of each subject by each judge revealed that the judges agreed with themselves consistently with substitution of /f/θ/ for subject 2, /w/r/ substitution for subject 3 and /w/h/ substitution for subject 5. The McDonald Deep Test was designed to find the contexts in which the error phoneme is correctly articulated. However, the results of this study indicate that when a subject is inconsistent in production of the target phoneme in different contexts, the reliability decreases.

The Two-Way Analysis of Variance data revealed that the tests did affect the subjects' responses as determined by the judges. Depending on which test a subject was given, a different analysis of the articulation skills was determined. Articulation tests are often used to assess speech therapy progress and the determination of progress is dependent upon whether or not the judge is reliable with himself.

The results of this study indicate that the McDonald

Deep Test does affect the judges' recordings of a subject's response to a greater extent than the other three tests. Although the reliability of 83 percent was not the lowest reliability for all the tests, the standard deviation indicated that there was more variation around the mean than the other tests, making it less reliable when a judges' responses were compared for two listening trials.

The results of this study present important consideration for speech pathologists in assessing articulation. The results indicate that reliability training is warranted for every test examined for both inter- and intra-judge reliability. Mere presentation of test manuals and discussion of precedures for administration is not enough for the clinician that will be administering articulation tests. Test training should involve practice trials until skills are proficient at a determined level of reliability. The importance of test training was also expressed in the study by Shriberg and Swisher, 1972. For this study, six University Supervisors were the judges. Their training and experience placed them in the position requiring that they make articulation judgments on a regular basis in a training clinic. When considering that many students in speech pathology and many less experienced judges are administering these and other tests, it is assumed that their reliabilities would be even lower than those given in this study without reliability training.

The results of this study indicate that there are differences in inter- and intra-judge reliabilities among the four

tests administered. Future research examining the reliabilities of these and other articulation tests is warranted. The results of a similiar study with a group of judges trained on all tests and a group of untrained judges would provide more information about the importance of training programs. Another study that is indicated by the results of the present study is comparison of children with multiple, functional articulation errors and children who have only one functionally misarticulated sound on the McDonald Deep Test. Perhaps other misarticulated sounds within the context are affecting the reliability of judgments of a given target sound production. A third research question indicated by the results of this study is whether or not the length of articulation tests, such as the 141 responses on the Templin Darley Test of Articulation is a variable affecting reliability of that test.

Inter- and intra-judge reliability studies not only indicate training necessary for given tests, but also provide information in regards to the most reliable articulation tests available today.

According to the data presented in this study, the most reliable test to consider in assessing articulation skills is the Goldman-Fristoe Test of Articulation Sounds in Words Subtest. However, the mean reliability of all four tests strongly indicates that formal training in the administration of all four tests is necessary.

APPENDIX

Dear Mr. and Mrs. Doe:

I am doing an articulation study at the University of New Mexico under the direction of Dr. Wayne Swisher. I would like to use your son, John, in this study. I will be giving four articulation tests which will require two thirty minute sessions with your son. The names and results of the tests will be kept confidential. The study is designed to give us more understanding of reliability of articulation tests. Your participation would be greatly appreciated. If you have no objections, please sign below and return in the enclosed envelope as soon as possible.

_____
Mr. and Mrs. J.H. Doe

Sincerely,

_____
Lydia P. Evans
Graduate Student
University of New Mexico

_____
Wayne E. Swisher, Ph.D.
Supervisor

Majority decision	Test ___ Subject Res. #	JUDGE 1 A	1 B	2 A	2 B	3 A	3 B	4 A	4 B	5 A	5 B	6 A	6 B
	1												
	2												
	3												
	4												
	5												
	6												
	7												
	8												
	9												
	10												
	11												
	12												
	13												
	14												
	15												
	16												
	17												
	18												
	19												
	20												
	21												
	22												
	23												
	24												
	25												
	26												
	27												
	28												
	29												
	30												
	31												
	32												

Form 1: Judgments of each response for each test by each judge and the majority decision of listening trial A.

Test ___

SUBJECT

Judge	1	2	3	4	5	6
1						
2						
3						
4						
5						
6						

Form 2: Intra- and Inter-judge percentages for each judge on test ___ for all subjects and the mean percentage of all judges.

BIBLIOGRAPHY

1. Carter, E.T., and Buck, M.W. Prognostic testing for functional articulation disorders among children in the first grade. JSHD, 1958, 23, 124-133.

2. Curtis, J.F. and Hardy, J.C. A phonetic study of misarticulation of /r/. JSHR, 1959, 2, 244-257.

3. Goldman, R. and Fristoe, M. Goldman-Fristoe Test of Articulation. American Guidance Service, Inc. Publishers' Building, Circle Pines, Minnesota, 1969.

4. Hartup, W.W. Sex and Social reinforcement effects with children. Paper presented at the meeting of the American Psychological Association, New York, September, 1961.

5. Henderson, F.M. Accuracy in testing the articulation of speech sounds. Journal of Educational Research, 1938, 31, 348-356.

6. Irwin, R.B. Consistency of judgments of articulation productions. JSHR, 1970, 13, 548-555.

7. Jordan, E.P. Articulation test measures and listener ratings of articulation defectiveness. JSHR, 1960, 3,

8. McDonald, E.T. The McDonald Deep Test of Articulation. Stanwix House, Inc., 3020 Chartiers Ave. Pittsburgh, Pennsylvania, 1964.

9. Shriberg, L.D. The effect of examiner social behavior on children's articulation test performance. JSHR, 1971, 14, 449-462.

10. Shriberg, L.D. and Swisher, W.E. Development of an articulation scoring training program. Paper presented at the American Speech and Hearing Association National Convention, San Francisco, 1972.

11. Siegel, G.M. Experienced and inexperienced articulation examiners. JSHD, 1962, 27, 28-35.

12. Siegel, G.M., Winitz H., and Conkey, H. The influence of testing instruments on articulatory responses of children. JDHD, 1963, 28, 67-76.

13. Smith, M.W., and Ainsworth, S. The effects of three types of stimulation on articulatory responses of speech defective children. JSHR, 1967, 10, 348-353.

14. Snow, K., and Milisen, R. The influence of oral versus pictorial presentation upon articulation testing results. JSHD, Monograph Supplement, 1954, 4, 29-36.

15. Sommers, R.K., Shilling, S.P., Paul, C.D., Copetas, F., Bowser, D.C. and McClinton, C.J. Training parents of children with functional misarticulation. JSHR, 1959, 2, 258-265.

16. Stevenson, H.W. Social reinforcement with children as a function of CA, sex of E, and sex of S. Journal of Abnormal Social Psychology, 1961, 63, 147-154.

17. Templin, M.C. A non-diagnostic articulation test. JSHD, 1947, 12, 392-396.

18. Templin, M.C. and Darley, F.L. The Templin-Darley Test of Articulation. Iowa City, Iowa: Bureau of Educational Research and Service, Extension Division, State University of Iowa, 1960.

19. Winitz, H. Temporal reliability in articulation testing, JSHD, 1963, 28, 247-251.

20. Winitz, H. and Siegel, G.M. Intratest variability in articulation testing. Unpublished study, 1961.

21. Winitz, H. and Bellerose, B. Phoneme generalization as a function of phoneme similarity and verbal unit of test and training stimuli. JSHR, 1963, 6, 379-392.