1973

The Development of a Severity Rating Scale for Evaluating Harsh Voice Quality

Sandra Kay Slater
Eastern Illinois University

PAPER CERTIFICATE #2

TO: Graduate Degree Candidates who have written formal theses.

SUBJECT: Permission to reproduce theses.

The University Library is receiving a number of requests from other institutions asking permission to reproduce dissertations for inclusion in their library holdings. Although no copyright laws are involved, we feel that professional courtesy demands that permission be obtained from the author before we allow theses to be copied.

Please sign one of the following statements:

Booth Library of Eastern Illinois University has my permission to lend my thesis to a reputable college or university for the purpose of copying it for inclusion in that institution's library or research holdings.

July 30, 1973
Date

I respectfully request Booth Library of Eastern Illinois University not allow my thesis be reproduced because _____

_____

_____

_____ _____

Date Author

pdm

THE DEVELOPMENT OF A SEVERITY RATING SCALE FOR EVALUATING HARSH VOICE QUALITY

(TITLE)

BY

SANDRA KAY SLATER

THESIS

SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF

Master of Science

IN THE GRADUATE SCHOOL, EASTERN ILLINOIS UNIVERSITY
CHARLESTON, ILLINOIS

1973
YEAR

I HEREBY RECOMMEND THIS THESIS BE ACCEPTED AS FULFILLING
THIS PART OF THE GRADUATE DEGREE CITED ABOVE

July 30, 1973
DATE

July 30, 1973
DATE

ACKNOWLEDGEMENTS

I wish to express my appreciation to the following individuals who have contributed their efforts toward the preparation of the following study.

To Dr. Lynn Miner, for serving as thesis advisor and supervising the collection of data, the statistical procedures, and the writing of the paper. Acknowledgement is given to Dr. Wayne Thurman, Dr. Jerry Griffith, and Mrs. Mary Beth Armstrong for serving as members of the graduate committee.

I wish to extend my gratitude to the East Central Illinois public school speech therapists who supplied subjects for this study; to the university professors who allowed me to use their classes; and to the students who sacrificed their time to perform the scaling tasks.

I express a special "thank you" to my parents for their abundant help and encouragement throughout this study.

TABLE OF CONTENTS

LIST OF TABLES

Table

CHAPTER I

STATEMENT OF THE PROBLEM

Everyone seems to be in agreement that not all voices sound alike. Some voices are more pleasant to listen to than others. However, there seems to be little agreement as to what constitutes a deviant voice quality, and more precisely, what constitutes "harsh" voice quality. The term "harsh" voice quality is difficult to identify, at least, with much consistency between clinicians. Everyone seems to have his own acoustical image of "harsh" voice quality. These personal systems seem to be "somewhat" meaningful to their possessors, yet appear to communicate little with others. Therefore, inter- and intra-clinician communication in regard to "harsh" voice quality is minimal. Even more pronounced is the lack of client-clinician communication. Clients are often told that their voices sound a little "rough" or "strained." Then, periodically, during a therapy situation, they are told that their voice quality sounds a little less strained or somewhat smoother. Actually, what has the client been told? Not much! For, what exactly is strained voice quality, a little less strained voice quality, or somewhat smoother voice quality?

There is a growing awareness of the proposed relation-

ship between the use of some deviant voice qualities, such as "harsh", and injuries to the vocal folds. During the last several years, statements such as "continual glottal fry, like hypertense phonation, can be injurious to the vocal folds" (Fisher, 1966), have become common entries in literature relating to speech pathology. Consequently, this proposes a definite need for one to provide speech clinicians with a quick, reliable and valid estimate and evaluation of the degree of a child's "harsh" voice quality. If existing differences in degree of harsh voice quality are not differentially diagnosed, then they may really have no significant meaning.

Little effort has been directed toward the development of a consistent standard by which to identify what is meant by terms such as "strained", "rough", or "somewhat smoother" voice quality. It is pointed out by Lafon and Guichard (1971) that one needs to obtain quantified values when collecting data. They feel that "objective verification of the results of therapy is possible" and that clinicians should direct their efforts toward obtaining objective evaluations.

The purpose of this study was to compile a master tape of the severity of harsh voice quality on a seven point equal appearing interval scale, and to then determine if specifically trained clinicians could reliably use the tape. Thus, provide speech clinicians with a means for objective evaluation of harsh voice quality.

A scale, as proposed in the present study would represent a "meaningful parameter" of speech, only if judges could reliably and validly classify acoustical stimuli. Young (1969) states that severity rating of voice quality is a "perceptual event" and that "to depend on observers for measurement is to recognize that classifying speech as deviant requires the judgment of an observer." Scale values can represent "meaningful parameters of speakers" (1969) if judges can reliably and validly classify acoustical stimuli.

The initial intent was to: 1) collect a sample of the voice quality which public school clinicians label as "harsh", (2) and construct a master tape, composed of a range of degrees of the "harsh" voice quality previously identified, (3) then utilize this defined range of harsh voice quality in voice analysis. The tape was constructed, primarily, for use as a tool for objective evaluation of a client's progress in therapy.

Specifically, the following questions were posed at the onset of this study.

1. Can untrained observers reliably use a seven point equal appearing interval scale to rate the severity of harsh voice quality?

2. Can specifically trained clinicians reliably use the tape to rank order the severity of harsh voice quality?

3. Specifically what does a clinician do to become trained? In other words, what training procedures does a clinician use to train herself to use the master tape of harsh voice quality?

Stated as a research hypothesis:

> A master tape can be compiled to represent reliably the severity of harsh voice quality on a seven point equal appearing interval scale, and specifically trained clinicians can use the tape to rank order reliably degrees of harsh voice quality.

The construction of this master tape of the severity of harsh voice quality was an extension of a previous study (Dudley, 1970), which concluded that the psychological scaling method of equal appearing intervals serves as a reliable and practical measurement tool for "quantifying the perceptual impact of voice quality deviations."

estimate. However, a test-retest reliability estimate of 51 percent was obtained by re-presenting the voice samples for re-classification, to a second group of judges. The reliability estimate involved the classification of six types of voice quality; hoarse, harsh, breathy, nasal, strident, and thin. By the use of a sonographic analysis technique, he found that in harsh voices the first formant tended to be lower than normal. (Table 1) The median was selected as a measure of central tendency "so that extreme variations of frequency location would have less influence and the measures would, then be more representative of the group of formants in any one vowel in the quality group." (Thurman, 1953)

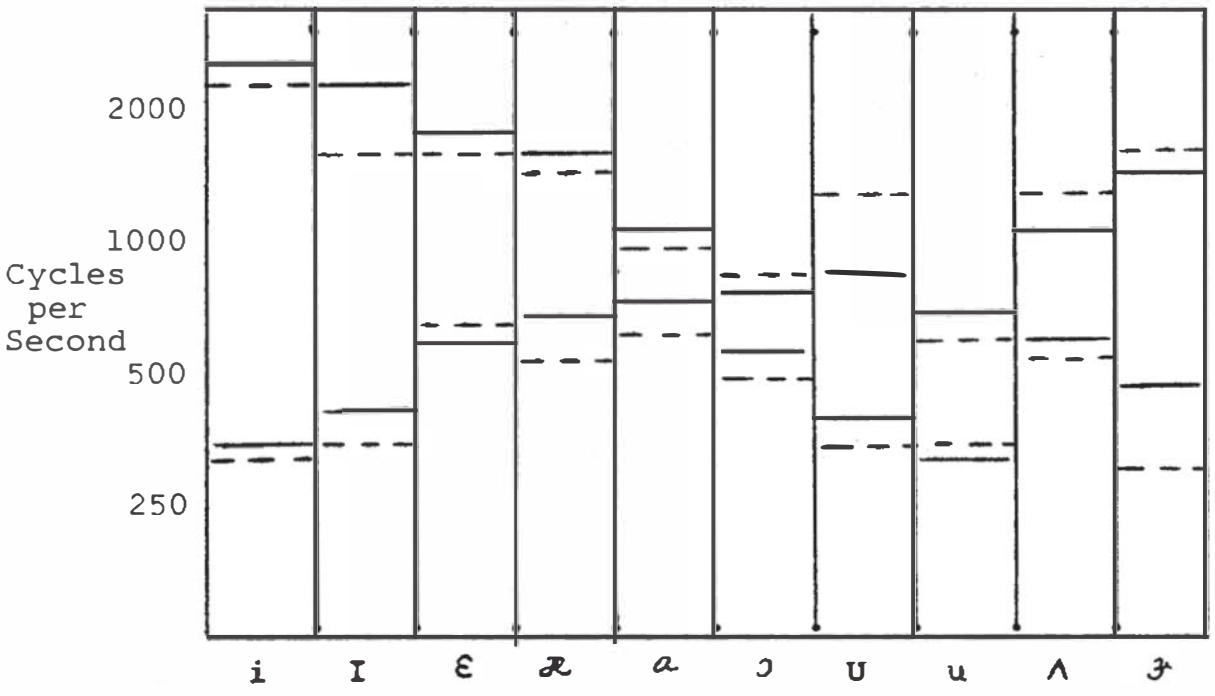


Table 1

Median first and second vowel formant frequency locations for the harsh group (dotted lines)

in ten vowels as compared to normal formant locations given by Fletcher[1] (solid lines).

This proposes a satisfactory means for identifying harsh voice quality, not degrees of harsh voice quality, unless one is specifically trained in graphic analysis. This could prove to be an effective yet involved technique to employ for the purpose of identifying degrees of harsh voice quality.

The Jewish Hospital Voice Profile (Wilson, 1972) presents another option. This method of voice evaluation requires the evaluator to make subjective judgments involving a speaker's pitch, laryngeal opening, resonance, and vocal range. Therefore, the evaluator has the very demanding task of "tuning into" and evaluating several features of voice quality at once. This method also requires the evaluator to become quite familiar with its evaluating system (through a complicated and somewhat confusing training session), so that he can make reliable judgments regarding voice quality.

Previous studies have investigated perceived harsh voice quality, but have not dealt with categorizing degrees of harsh voice quality. Rees (1958) studied the influence of vowels, selected consonant environments, and vowel initiation on perceived harsh voice quality. She had 32 listeners rate syllables of twelve speakers with clinically diagnosed

---

[1]Fletcher, Harvey, Speech and Hearing In Communication (New York, 1953), p. 62-3.

harsh voices on a seven-point equal-appearing intervals scale. Rees considered the results to be "satisfactorily reliable."

In a similar study, Sherman and Linke (1952) studied the influence of certain vowel types on the degree of harsh voice quality. The particular vowel categories chosen for study were: front, back, high (or short), low (or long), tense, and lax. The following conclusions were drawn:

1. "High vowels are perceived as less harsh than low vowels. Since high vowels are shorter in duration than low vowels, the assumption that short vowels are, in general, perceived as less harsh than long vowels, seems reasonable."
2. "Lax vowels are perceived as less harsh than tense vowels."

Results indicated that controlled categories of vowel factors could be rated as to perceived harshness by a seven-point interval scaling method.

Sherman (1954) evaluated the method of obtaining scale values of severity of harshness and of nasality with recorded speech samples played backwards. The intent was to eliminate irrelevant judgment variables such as articulation and semantic information. A seven-point equal-appearing interval scale for rating voice quality was applied. A Pearson r of .89 between results of forward and backward playing indicated that scale values by the two methods to be about equally reliable.

In the previously mentioned studies, the attempt was to generate specific statements concerning associated components of harsh voice quality, or its general perception.

These studies could represent basic stepping stones toward identifying the "fuzzy", harsh voice quality. Yet, these studies have not been followed with further investigation attempting to pinpoint harsh voice quality.

Investigations provide strong evidence that psychological scaling methodologies have been successfully used to rate articulation (Morrison, 1955; Sherman and Cullinan, 1960; Jordan, 1960; Prather, 1960), language (Shriner and Sherman, 1967; Sherman and Silverman, 1968; Galloway, 1972), stuttering (Sherman and Lewis, 1951; Sherman and Trotter, 1956; Young, 1961) and voice (Sherman and Linke, 1952; Sherman, 1954; Rees, 1958; Spriestersbach, 1955; Spriestersbach and Powers, 1959; Lintz and Sherman, 1961; Dickson, 1962; Dudley, 1970).

Sherman and Moodie (1957) compared equal-appearing intervals, successive intervals, paired comparisons, and constant sums scaling methods to find the most reliable method for scaling defectiveness of articulation. Scale values obtained by the method of paired comparisons were demonstrated to lack internal consistency according to a statistical test used to evaluate the validity of assumptions made regarding the distribution of scale values. Scale values obtained by the method of constant sums were different from the values derived by the other three scaling procedures in that there was a clustering of scale values at the extremes of the scale. On the basis of reliability of scale

values, ease of computation, and close agreement with internally consistent scale values obtained by the method of successive intervals, the equal-appearing intervals appeared to be the most useful for scaling articulation defectiveness. Cullinan, Prather, and Williams (1963) compared the results of severity of stuttering ratings by six variations of the equal-appearing interval method, and found interjudge reliability coefficients ranging from .95 to .97. Sherman and Silverman (1968) compared equal-appearing interval, successive interval, and direct magnitude estimation scaling methods in assessing language development and found that scale values for the methods differed very little. They preferred the equal-appearing intervals scaling method due to its simpler computational procedures. Dudley (1970) concluded that the equal-appearing interval scaling method served as a reliable and practical measurement tool for quantifying voice quality. He reached a reliability level of .99 with a population of 143 untrained observers rating 42 segments of harsh voice quality.

Of the various psychological scaling methods available, the method of equal-appearing intervals appears to be the most widely used method for quantifying listener ratings. The method of equal-appearing intervals was originally described by Thurstone and Chave (1929). In 1954, Guilford presented some advantages for using equal-appearing interval rating methods. They were:

1. EAI requires much less experimental time than either pair comparisons or ranking methods.

2. EAI can be used with naive raters who have had a minimum of training.

3. EAI can be used when presenting a large number of stimuli.

4. EAI has a wider range of application than do ranking or comparing methods.

5. Some experimenters maintain that best judgments are made when stimuli are presented singly, thus assuming that comparative scales destroy the "aesthetic attitude" of the rater.

The assumption of equal-appearing intervals is that "the intervals into which values are rated are equal." (Sherman and Moodie, 1957) Equal-appearing interval scale values represent interval data. (Guilford, 1954) According to Williams, interval data infers "the assignment of numbers for the purpose of identifying ordered relations of some characteristic, the order having arbitrarily assigned and equal intervals but an arbitrary zero point." (1968) However, in a study by Berry and Silverman, it was concluded that it is "not safe to assume that the intervals of equal-appearing interval scales are subjectively equal," relative to the use of severity scales. (Berry and Silverman, 1972) In this study they evaluated the equality of the intervals on the Sherman-Lewis scale of stuttering severity. The interval widths between scale values on that severity scale were not found to be equal. This suggests that the stuttering severity scale represents ordinal data, which is

merely rank ordered. Therefore, it is only safe to assume that the scale values representing harsh voice quality also represent ordinal data, which is characterized by "the assignment of numbers or symbols for the purpose of identifying ordered relations of some characteristic, the order having unspecified intervals." (Williams, 1968)

CHAPTER III

METHODOLOGY

This chapter discusses the subjects, equipment and procedures used for this investigation.

Choice of scaling method.

The psychological scaling method of equal-appearing intervals was chosen for this study on the basis of its positive results in experimentation with speech disorders. (c.f. pp. 9-12)

Preparation of stimuli.

A modified Strandberg procedure (1969) was used to collect samples of harsh voice quality. In the Strandberg study, the voice quality samples were elicited by questions regarding a favorite T.V. program, a most enjoyable summer activity, or most enjoyable part of school. The children were enrolled as first graders in public schools in the East Central Illinois communities of Charleston, Mattoon, and Sullivan. Each child had been diagnosed as having harsh voice quality by one of five East Central Illinois speech pathologists. Harsh voice quality stated by Rees (1958) as defined by Curtis as "an unpleasant, rough, rasping sound, often heard in people for whom voice production

seems to be a considerable effort or strain," was diagnosed. At the respective schools, Strandberg collected tape recorded conversational speech samples of the children who had been referred. Strandberg used an Ampex, Model 602 tape recorder and recorded the speech samples at a tape speed of seven and one-half inches per second. Recordings were made on Scotch Magnetic Tape, silicone lubricated 1.5 mil acetate backing.

Voice quality samples for the present study were elicited by asking kindergarten through fourth grade children questions about their families, their favorite T.V. programs, and what they like best about school. The children were enrolled in public schools in the East Central Illinois communities of Newton, Decatur, Neoga, Paris, Effingham, Herrick, and Altamont. The public school speech pathologists in the previously mentioned communities were given the following task:

> "Identify any kindergarten through fourth grade child with a voice quality which is aesthetically unpleasant to listen to, and consequently calls the unfavorable attention of most listeners. It is a voice quality which hygienically exhibits excessive laryngeal tension, evidenced by the speaker's apparent strain and effort in vocalization. The voice may have the accompaning characteristics of breathy quality, glottal fry, and/or low pitch. Overall, it is a hard, flat, inefficient voice."

At the respective schools, in a "quiet" room with only one child and the experimenter present during the recording, conversational speech samples of the children who had been

referred were tape recorded. The recordings were made at seven and one-half inches per second on a Rheem Caliphone, model 70-TC tape recorder. Concert Tape, with a silicone lubricated 1.5 mil acetate backing was utilized.

The tape for the untrained observer ratings was constructed using ten second continuous vocalization segments of each of the previously collected harsh voice quality samples. The first ten seconds of continuous vocalization, free of apparent pauses, of each sample of harsh voice quality, was extracted, then all of the extracted ten second segments were compiled to form the taped composite of the various degrees of harsh voice quality previously collected. Sixty samples of harsh voice quality were collected. This sampling contained only samples of those voice qualities previously identified by public school speech pathologists as presenting some degree of harsh voice quality. The ten second sampling was chosen on the basis of research by Sherman and Moodie (1957), and Sherman and Lewis (1951). In the Sherman and Moodie research, it was concluded that observers, using interval scales, could rate articulation severity of five and ten second segments as reliably as with one minute samples of continuous speech. The Sherman and Lewis study, concluded that in rating stuttering severity, six second samples were too short, fifteen second samples were "unnecessarily prolonged", but ten second samples were of optimum length. The ten second sample also proved successful in the

Dudley (1970) study.

The tape included a five second interstimulus interval to allow for observer judging and recording. Each speech segment was preceded by a two second pure tone, to aid the observer in preparing to listen to the upcoming speech segment. A respective segment number was displayed on a 5" X 8" card, in the front of the testing room.

Construction of Training Tape.

The thirty-three untrained observers' (enrolled in an introductory speech pathology class at Eastern Illinois University) ratings were transferred from the answer sheets to IBM data cards from which statistical computation was made. The mean scale value and semi-interquartile range for each of the sixty stimuli was computed. Four segments that had calculated mean scale values falling nearest each of the one to seven proposed whole integer values were extracted. Then, the speech segments having the smallest SIQ (semi-interquartile range) values, among those samples previously extracted as having mean scale values falling nearest each of the whole integer scale values, were selected to represent each level of harsh voice quality on the master tape of the severity of harsh voice quality. Only one speech segment met the desired mean scale value criterion, for representation of each Level 1 and Level 7, on the master tape of harsh voice quality. Therefore, these segments were accepted to represent Level 1 and Level 7, without

further SIQ consideration.

Ten speech samples were randomly chosen from the remaining samples. They were numbered consecutively one to ten, and placed on individual tape reels. It is realized that the trained observers were asked to classify the speech samples as representing one of the whole integer scale values, when the samples likely represented transitional scale values. This factor was of little consequence in this study, for reliability of rankings was not based on exact scale values, but on the relationship of the rank ordering of the speech segments obtained from the trained observers' ratings. The rank ordering, thus ordinal data was assumed on the basis of the results of the Berry and Silverman study. (1972)

All tapes were made using a Rheem Caliphone, model 70-TC tape recorder, with a tape speed of seven and one-half inches per second. Recordings were made using Concert Tape, silicone lubricated 1.5 mil acetate backing.

Instructions to Judges.

The instructions to the untrained judges were extracted from the Dudley (1970) study. They may be found in Appendix I.

The trained observers were given very little instruction. Each was placed in a room, alone, with the master tape and ten, randomized tape reels, containing the speech samples. The observers were given a list of aspects (found in Appendix III), drawn from a pilot study, characterizing

possible approaches one might follow during the categorization task. Briefly, in the pilot study, two Speech Pathology and Audiology majors at Eastern Illinois University, with at least a bachelor's degree, were asked to categorize five randomly chosen segments of harsh voice quality (extracted from the previously collected samples of harsh voice quality) according to the one to seven whole integer scale values on the master tape of harsh voice quality, using whatever method they preferred. When this task was completed, the experimenter and the pilot judges, after discussion, summarized the procedures used during the categorization task.

The actual trained observers in this study were given no specific procedure or time limit to follow in training themselves to use the master tape to categorize harsh voice quality. The "procedural hints" were given to the judges to help orient them to the task. The ten tape reels, containing the speech samples, were placed in random order on a table. The observers were instructed to evaluate the tapes in the same order as they were placed before them. Each was asked to categorize the individual tape samples as best representing one of the whole integer categories, as presented on the master tape.

The observers' behaviors were video taped during the training and categorizing task, to facilitate in the procedural analysis.

Selection of Judging Panel.

The present study involved thirty-three untrained judges (those having no previous knowledge of voice disorders) and seven trained judges (those having previous knowledge concerning voice disorders).

The untrained judges for this study were the judges who were asked to judge sixty samples of harsh voice quality, in relation to a seven-point scale of "unpleasantness." These untrained judges were undergraduate students enrolled in an introductory speech pathology class at Eastern Illinois University. These judges were naive as far as knowledge of voice disorders.

The "initial" (untrained) judges were untrained for two specific reasons:

1. experienced judges were not available
2. "Ignorance of the areas of speech pathology and language development may constitute an experimental safeguard against particular biases or expectations." (Siegel, 1962)

These reasons were first proposed by Gerald Siegel in reference to the preference of untrained over trained observers of articulation. However, they are equally influential factors to consider when selecting judges to evaluate voice quality.

Trained clinicians were chosen to manipulate the master tape in order to categorize samples of harsh voice quality, due to the desired level of sophistication. The five trained clinicians were selected from the Speech Pathology and

Audiology majors at Eastern Illinois University. Each had at least a bachelor's degree in Speech Pathology and Audiology.

The minimum level of reliability for this study was set at .95, due to the desired stable rank ordering and reliability of the data. The number of untrained observers needed to reach the .95 level of reliability was determined by Silverman's principle of sequential sampling. (1968) In this procedure the experimenter establishes a minimum level of reliability desirable for his scale values. Next he has a small number of observers rate the stimuli. He then estimates the reliability of the scale values, which can be derived from the ratings of these observers. If the level of reliability attained is greater than or equal to the desired level, then no observers are added. However, if the level of reliability attained is less than the desired level, the experimenter would then have additional observers selected from the same population of observers rate the stimuli. This same process would be continued until the desired level of reliability is reached. In this study the desired level of reliability was reached with the initial group of thirty-three judges.

Presentation of Stimuli.

The stimuli were presented to the observers by a Rheem Caliphone, model 70-TC tape recorder, at a tape speed of seven and one-half inches per second. The stimuli were

presented at a comfortable listening level. The presentation was made to thirty-three untrained judges, in their respective classroom.

The trained clinicians were given Rheem Caliphone, model 70-TC tape recorders, to use in their training process.

Analysis of Judge's Ratings.

The thirty-three untrained observers' ratings were transferred from the answer sheets to IBM data cards from which statistical computations were made. The mean scale value and semi-interquartile range for each of the sixty stimuli was computed. Four segments that had calculated mean scale values falling nearest each of the one to seven proposed whole integer values were extracted. Then, the speech segments having the smallest SIQ (semi-interquartile range) values, among those samples previously extracted as having mean scale values falling nearest each of the whole integer scale values, were selected to represent each level of harsh voice quality. Only one speech segment met the desired mean scale value criterion, for representation of each Level 1 and Level 7, on the master tape of harsh voice quality. Therefore, these segments were accepted to represent Level 1 and Level 7, without further SIQ consideration.

The trained clinician's ratings were analyzed by the intraclass correlation coefficient for averages and the intraclass correlation coefficient - reliability of indivi-

dual numbers (adjusted for trend) and (unadjusted for trend), by Winer, 1962. Each clinician's training procedure was analyzed by the clinician and the experimenter on the basis of direct observation, video tape review, and self-description of techniques. Then, prominent procedural characteristics among the various training processes were identified.

All statistical analyses were computed on an IBM 360 computer. A mean, median, and semi-interquartile range was computed for each of the sixty stimuli from the untrained observers' ratings, and for the ten stimuli from the trained observers' ratings.

CHAPTER IV

RESULTS

The purpose of this study was to compile a master tape of the severity of harsh voice quality on a seven-point equal appearing interval scale, and to then determine if specifically trained clinicians could reliably use the tape. This chapter reports the statistical computations and interprets the results.

Reliability of Untrained Observers' Ratings.

An intraclass correlation coefficient for averages (Winer, 1962) was computed to evaluate the reliability of the untrained observers' scale value ratings. A reliability level of 0.97 was obtained with a population of 33 observers rating sixty stimuli. This reliability level surpassed the 0.95 reliability level desired at the onset of this study. Approximately 94% of the variance was accounted for. The untrained observers did use reliably a seven-point equal appearing interval scale to rate the severity of harsh voice quality. This was in agreement with the findings of the Dudley (1970) study, in which a reliability level of 0.99 was reached, with a population of 143 observers rating 42 segments of diagnosed harsh voice quality. It was concluded,

as in the Dudley study, that the equal-appearing interval scaling method serves as a reliable and practical measurement tool for quantifying harsh voice quality.

When attempting to develop a scale for evaluating harsh voice quality it is important to have a range of stimuli presented on that scale. Inspection of median scale values indicated that the observers, in this study, perceived a range of voice quality. More specifically, the median scale values varied from a low of 1.42 to a high of 6.55.

Reliability of Trained Observers' Ratings.

The trained clinician's ratings were analyzed by the intraclass correlation coefficient for averages, reliability of individual numbers (adjusted for trend), and reliability of individual numbers (unadjusted for trend), by Winer, 1962. In this study the examiner was interested in knowing to what extent the panel of judges could assign the same rank ordering of scale values to the stimuli, and to what extent each judge could assign the same absolute scale value to each stimulus on a test-retest basis.

Intraclass correlation coefficients are statistical measures of association or of group agreement. They are very stringent and highly controlled measurements. The three types of intraclass correlation coefficients are:

1. intraclass correlation coefficient for averages - It supplies a highly accurate estimate of a Pearson r; that is, it estimates the amount of agreement between groups, if a second set of judges are asked to rate a set of stimuli, and

a correlation between group one and two is computed.

2. intraclass correlation coefficient - reliability of individual numbers (adjusted for trend) - It supplies a correlation of the ratings between the individual judges; in other words, it reveals to what extent the individual judges can rank order the stimuli in the same manner. It is a rank order correlation.

3. intraclass correlation coefficient - reliability of individual numbers (unadjusted for trend) - It reveals to what extent the judges can assign the same scale values to the stimuli. It is an exact number correlation.

The resulting intraclass correlation coefficient (for averages) was 0.94 indicating that clinicians could reliably scale harsh voice quality segments with a minimal amount of variance (12%) unaccounted for, on a test-retest basis. That is, the panel of trained observers, as a group, were consistent in their perceptual reactions to the voice quality stimuli. Apparently, they were all applying the same criterion for making perceptual judgments. This reliability coefficient was obtained with a population of five trained observers. It was determined by Silverman's principle of sequential sampling (1968), that one more trained observer's ratings would have been needed to reach the desired 0.95 reliability level stated at the onset of this study.

While the intraclass correlation coefficient for averages suggested that the group ratings were internally consistent, the development of a severity rating scale requires a more stringent statistical analysis. For purposes of this experiment it was felt necessary to look not at group be-

havior but rather at individual clinician behavior. The question of interest was: To what extent does each individual clinician rank order the voice quality stimuli in the same manner? Emphasis was placed on individual clinician behavior rather than on group behavior because clinical judgments are made by individuals and not by groups. Therefore, the correlation value of greatest interest in this study was the intraclass correlation coefficient - adjusted for trend. The resulting value of 0.76 indicated that there was considerable response variability among the individual observer's rank ordering the stimuli. A correlation value of this magnitude indicates 58% common variance among the individual observers, with 48% of the variance unaccounted for, which in part may be due to the multidimensional nature of the stimuli. The specific way in which the other voice quality variables interact with perceived harshness remains unanswered. Further psychological scaling studies with this master training tape might well suggest some solutions. Specifically, follow up studies should include having these voice quality stimuli rated for such attributes as breathiness, tension, appropriateness of pitch level, and glottal fry. Correlations between scale values derived for these attributes and the scale values resulting from this study should suggest the influence of other voice quality variables upon perceived harshness.

The intraclass correlation coefficient (unadjusted for trend) was 0.69, revealing that the observers were unable to

reliably assign the same scale values to the stimuli. Fifty-two percent of the variance was unaccounted for. This means that the individual judges did not consistently apply the same absolute scale values to each voice quality stimulus.

In scaling the ten segments of harsh voice quality, the confusion seemed to be greater at the midpoints along the scale, and lesser at the ends of the range. This response variability, at the midpoints, might present a clinical barrier, especially if one is attempting to evaluate progress made during voice therapy. Apparently, the individual observers had less difficulty differentiating between stimuli which fell at the ends of the scale as compared to the middle of the range. This variability may be due to less distinct features characterizing the midpoints along the scale (Levels 2-6), however, further investigation is warranted to justify the increased variability at the midpoints.

Training Procedures.

At the onset of this study, the following question was posed: What training procedures does a clinician use to train herself to use the master tape of harsh voice quality? Through clinician-experimenter discussion, three prominent training procedures were identified. They were:

1. Procedure 1 - The clinician listened to the master tape at least two times, making internal and/or written notations regarding features (such as glottal fry, breathy quality, and tension) of each level. With this body of knowledge, the clinician listened to and rated (assigned a master tape severity scale value) each of the ten segments of harsh voice quality,

without further reference to the master tape. She listened to each sample segment approximately twenty seconds.

2. Procedure 2 - The clinician listened to the master tape at least two times, making internal and/or written notations regarding features of each level. The clinician then proceeded the same as defined in Procedure 1, except that when unsure of the appropriate scale value to assign, she returned to the master tape to make acoustic comparisons between the sample segment and specific levels (chosen by the clinician) on the master tape. When returning to the master tape, the clinician listened to a specific level only one time.

3. Procedure 3 - The clinician listened to the master tape at least two times, making internal and/or written notations regarding features of each level. She then, made a gross comparison, for example, Level 3 or Level 4, and then narrowed the comparison to a single level, for example, Level 3. Before assigning a scale value, the clinician listened to specific master tape levels and the sample segment several (four to six) times.

Following is a composite list of features identified by the trained observers (clinicians) as characterizing each level of harsh voice quality on the master tape.

Level 1
1. no truly distinguishable characteristics
2. difficult to distinguish from "normal" voice quality

Level 2
1. emergence of breathy quality - slight
2. tension toward the end of sentences
3. lack of breath support - too much speech per breath stream

Level 3
1. more apparent breathy quality
2. inconsistent glottal fry
3. increased tension

Level 4
1. constant state of tension and strain

2. evident glottal fry
3. continued breathy quality

Level 5
1. constant state of tension and strain
2. evident glottal fry
3. continued breathy quality
4. deeper pitch level
5. inconsistent pitch breaks
6. inconsistent loss of voicing, due to excessive breathiness

Level 6
1. excessive tension
2. excessive pitch breaks
3. evident glottal fry
4. excessive breathy quality
5. more intense vocalization
6. moderate intelligibility

Level 7
1. frequent pitch breaks
2. extreme glottal fry
3. extreme breathy quality
4. whisper quality, due to voicing difficulties
5. frequent aphonia
6. minimal intelligibility

Inspection of the features for each level of harsh voice quality, suggests that breathiness is a feature characterizing six of the seven levels. Observer comments indicated that it was particularly apparent in the lower levels (Level 2 and Level 3) and somewhat less obvious in the higher levels (Level 6 and Level 7). Nevertheless, it was consistently a characterizing feature. This seems to suggest that harshness may not be a discrete voice quality, but overlaps with breathiness. Further studies on the reliability of clinicians to categorize types of voice quality disorders seem indicated.

Conclusions.

The research hypothesis posed at the onset of this investigation was: A master tape can be compiled to represent reliably the severity of harsh voice quality on a seven-point equal-appearing interval scale, and specifically trained clinicians can use the tape to rank order reliably degrees of harsh voice quality. The research hypothesis was rejected; that is a master tape was compiled to represent reliably the severity of harsh voice quality on a seven-point equal-appearing interval scale, however, specifically trained clinicians were unable to use the tape to reliably rank order degrees of harsh voice quality. Even though clinicians were unable to reliably rank order degrees of harsh voice quality, strides were made in identifying harsh voice quality, by assuming an analytic rather than a general view of the voice quality.

CHAPTER V

SUMMARY AND CONCLUSIONS

In recent years, great stress has been placed upon objective verification of the results of speech therapy. Various charting and tallying methods have been applied for evaluation of articulation and language; however, little effort has been directed toward obtaining an objective evaluation of voice quality. Clinicians develop their own personal systems for evaluating voice quality, which seem "somewhat" meaningful to themselves, yet communicate little with others. Thurman's sonographic analysis technique (1953) and the Jewish Hospital Voice Profile (Wilson, 1972) both provide a rather complex and somewhat confusing system for evaluating voice quality. The present lack of an adequate tool for evaluating voice quality and the intensive focus on objective verification of therapy results, creates a need for a brief, sensitive, yet adaptable tool to facilitate inter- and intra-clinician communication regarding voice quality.

The primary purpose of this investigation was to compile a master tape of the severity of harsh voice quality on a seven-point equal-appearing interval scale, and to then determine if specifically trained clinicians could reliably use the tape. The general procedure consisted of: 1) col-

lecting a sample of the voice quality which public school clinicians label as "harsh", (2) and constructing a master tape, composed of a range of degrees of the "harsh" voice quality previously identified, (3) then, utilizing this defined range of harsh voice quality in voice analysis. Specifically, the following questions were posed at the onset of this study:

1. Can untrained observers reliably use a seven-point equal-appearing interval scale to rate the severity of harsh voice quality?

2. Can specifically trained clinicians reliably use the tape to rank order the severity of harsh voice quality?

3. Specifically what does a clinician do to become trained? In other words, what training procedures does a clinician use to train herself to use the master tape of harsh voice quality?

Voice quality samples for the present study were elicited by asking kindergarten through fourth grade children questions about their families, their favorite T.V. programs, and what they like best about school. The children were enrolled in the East Central Illinois communities of Newton, Decatur, Neoga, Paris, Effingham, Herrick, and Altamont. The public school speech pathologists in the previously mentioned communities were asked to:

> "identify any kindergarten through fourth grade child with a voice quality which is aesthetically unpleasant to listen to, and consequently calls the unfavorable attention of most listeners. It is a voice quality which hygienically exhibits excessive laryngeal tension, evidenced by the speaker's apparent strain and effort in vocalization. The voice may have the accompaning char-

> acteristics of breathy quality, glottal fry, and/or low pitch. Overall, it is a hard, flat, inefficient voice.

At the respective schools, in a "quiet" room with only one child and the experimenter present during the recording, conversational speech samples of the children who had been referred were tape recorded. The recordings were made at seven and one-half inches per second on a Rheem Caliphone, model 70-TC tape recorder. Concert Tape, with a silicone lubricated 1.5 mil acetate backing was utilized. From these samples, a stimulus tape and a master tape were prepared.

The EAI stimulus tape was constructed by extracting the first ten seconds of continuous vocalization from each of the sixty, previously collected, harsh voice quality samples.

The tape included a five second interstimulus interval to allow for observer judging and recording. Each speech segment was preceded by a two second pure tone, to aid the observer in preparing to listen to the upcoming speech segment. A respective segment number was displayed on a 5" X 8" card, in the front of the testing room.

Thirty-three untrained judges were asked to rate the sixty samples of harsh voice quality according to a seven-point EAI scale. These untrained judges were undergraduate students enrolled in an introductory speech pathology class at Eastern Illinois University. They were naive as far as knowledge of voice disorders.

The EAI stimulus tape was presented in the student's

respective classroom. Each observer heard the stimulus tape twice. The first presentation proposed to allow each observer to listen only and to formulate his own concept as to the least and most severe voice quality perceived on the tape. The actual scaling task was performed during the second stimuli presentation.

An intraclass correlation coefficient for averages (Winer, 1962) was computed to evaluate the reliability of the untrained observers' scale value ratings. A reliability level of 0.97 was obtained with a population of 33 observers rating sixty stimuli. This reliability level surpassed the 0.95 reliability level desired at the onset of this study. Approximately 94% of the variance was accounted for. The untrained observers did use reliably a seven-point equal appearing interval scale to rate the severity of harsh voice quality. This was in agreement with the findings of the Dudley (1970) study, in which a reliability level of 0.99 was reached, with a population of 143 observers rating 42 segments of diagnosed harsh voice quality.

The master tape of harsh voice quality was constructed by selecting the speech segments having the least amount of variance (SIQ) and mean scale values falling nearest each of the whole integer scale values (as determined from the untrained observers' ratings) to represent each level of harsh voice quality.

Five clinicians selected from the Speech Pathology and Audiology majors, with at least a bachelor's degree, at

Eastern Illinois University were presented ten randomly chosen samples of diagnosed harsh voice quality, and asked to rate each sample as best representing one of the whole integer categories, as presented on the master tape.

The trained clinician's ratings were analyzed by the intraclass correlation coefficient for averages, reliability of individual numbers (adjusted for trend), and reliability of individual numbers (unadjusted for trend), by Winer, 1962. The resulting intraclass correlation coefficient for averages was 0.94 indicating that clinicians could reliably scale harsh voice quality segments with a minimal amount of variance (12%) unaccounted for, on a test-retest basis. That is, the group ratings were internally consistent.

For purposes of this investigation, emphasis was placed on individual clinician behavior rather than on group behavior because clinical judgments are made by individuals and not by groups. Therefore, the correlation value of greatest interest in this study was the intraclass correlation coefficient - adjusted for trend. The resulting value of 0.76 indicated that there was considerable response variability among the individual observers' rank ordering the stimuli. Also, the intraclass correlation coefficient - unadjusted for trend (0.69), revealed that the individual judges did not consistently apply the same absolute scale values to each voice quality stimulus.

Each clinician's training procedure was analyzed by the clinician and the experimenter on the basis of direct obser-

vation, video tape review, and self-description of techniques. Three promoinent training procedures were identified and a composite list of features identified by the clinicians as characterizing each level of harsh voice quality on the master tape, was generated.

The research hypothesis posed at the onset of this study was rejected; that is, a master tape was compiled to represent reliably the severity of harsh voice quality on a seven-point equal-appearing interval scale, however, specifically trained clinicians were unable to use the tape to reliably rank order degrees of harsh voice quality. Even though clinicians were unable to reliably rank order degrees of harsh voice quality, strides were made in identifying harsh voice quality, by assuming an analytic rather than a general view of the voice quality.

Implications for Future Research.

Inspection of the results of this study suggest several features and applications of the master tape of harsh voice quality which warrant further research. Follow up studies might include:

1. An investigation into the specific way in which other voice quality variables (breathiness, tension, appropriateness of pitch level, and glottal fry) interact with perceived harshness.

2. An investigation to justify the increased variability in differentiating stimuli representing midpoints of the range of harsh voice quality.

3. An investigation to evaluate the reliability of clinicians to categorize types of voice quality disorders.

APPENDIX I

INSTRUCTIONS TO UNTRAINED OBSERVERS

You are asked to judge a series of children's voices which are presented to you in a tape recorded form. You are asked to judge each voice sample in relation to a seven-point scale of "unpleasantness." Unpleasantness, for purposes of this experiment, is interpreted to mean that the quality is bad enough to call the unfavorable attention of most listeners to the child's voice.

Quite obviously, not all children's voices sound alike. Some voices are more pleasant than others. The voices you will hear were previously judged by speech pathologists to represent varying degrees of unpleasantness. Your task is simply to rate the degree of unpleasantness each voice represents.

Make your judgment on the basis of each individual voice quality. Avoid being influenced by mispronunciations of words, poor grammar, or usage of vocabulary, but listen to how each child sounds in terms of his voice quality; that is, how unpleasant does each child's voice sound to you.

The rating scale is one of equal intervals, from 1 to 7, with 1 representing the least unpleasant quality you hear and 7 representing the most unpleasantness you hear on the tape; 4 represents the midpoint between 1 and 7 with respect to

unpleasantness. The other numbers fall at equal distances along the scale. Do not attempt to place samples between any two of the seven points, but only at these points. Remember the range is from 1 to 7 with 7 representing the most unpleasant voice you hear on this tape.

Each unpleasant voice quality is preceded by a "signal tone." Your task will be to record your judgment to the right of the identifying number on the rating sheet. The numbers on the rating sheet run from the top to the bottom of the page.

Following there will be sixty voices to be rated on the 7-point scale. These voice samples were obtained by asking kindergarten through fourth grade children questions about their families, their favorite T.V. programs, and what they like best about school. All responses are to the same set of questions.

Before you record any judgments, you will listen to the voices previously judged to represent different degrees of unpleasantness, in order to acquaint yourself with the experimental task and to the range of voices which you are asked to judge with respect to the degree of unpleasantness. Just listen, form a concept of the least and most unpleasant voices on the tape. As you listen, pay close attention to how each child's voice sounds. Occasionally you will hear some background noise on the tape. Totally disregard this and form your impressions solely on the basis of each child's

voice. Do not record any judgments now. Just listen.

This time I will play the tape and you will judge each child's voice on the rating sheet. Remember, 1 represents the least unpleasant and 7 represents the most unpleasant voice quality you hear on this tape.

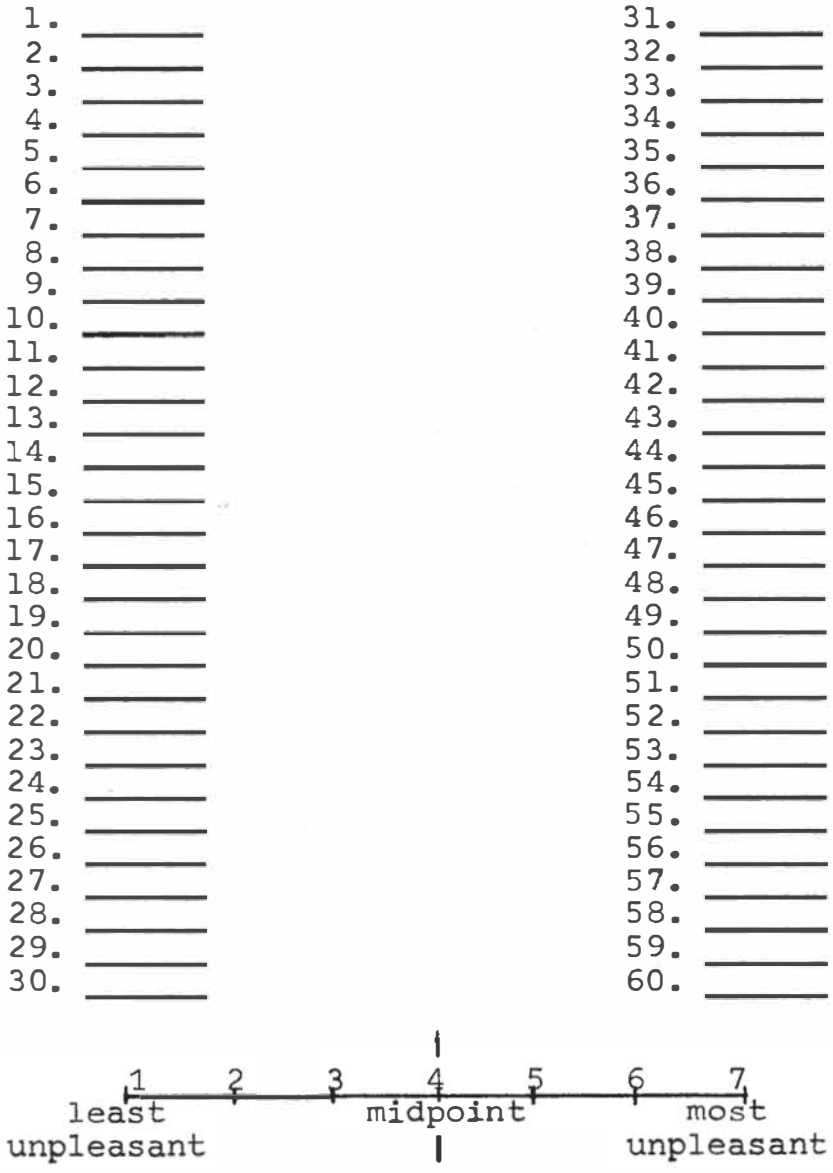
Make a judgment on every sample. If you are somewhat doubtful, make a guess as to the most suitable scale position.

Are there any questions?

APPENDIX II

UNTRAINED OBSERVER RATING SHEET

RATING SHEET



APPENDIX III

"PROCEDURAL HINTS"

You are asked to judge a series of children's voices which are presented to you in a tape recorded form. You are asked to judge each voice sample in relation to a seven-point scale of "harshness", also presented in tape recorded form. Make your judgments on the basis of each individual voice quality. Avoid being influenced by mispronunciations of words, poor grammar, or usage of vocabulary, but listen to how each child sounds in terms of his voice quality.

To help orient you to the task, the following list of hints has been compiled.

1. The acoustic characteristics of breathy quality and glottal fry, and the subjective characteristics of speech intelligibility and effort or strain (subjective in the present situation) in vocalization, have been noted as becoming gradually more severe from Level 1 to Level 7, with these characteristics being least noticable at Level 1 and most interferring with communication at Level 7. The degree of each of these characteristics at each level, requiring individual subjective judgments to be made by each clinician.
2. It may be helpful to listen to the master tape segments several times, until one becomes familiar with each level of harsh voice quality.
3. In categorizing a speech sample, it may be helpful to choose a gross comparison, for example, Level 3 or Level 4, and then narrow the comparison to a single level, for example, Level 3.

APPENDIX IV

MEAN SCALE AND VARIANCE VALUES
(as rated by untrained observers)

speech segment	mean scale	(SIQ) variance
1	3.88	0.24
2	2.21	0.73
3	2.48	0.82
4	4.48	0.86
5	2.79	0.89
6	5.42	0.71
7	3.82	0.97
8	2.55	0.90
9	2.03	0.76
10	4.88	0.75
11	1.76	0.30
12	4.27	0.74
13	1.42	0.13
14	3.64	0.91
15	6.12	0.38
16	5.12	0.96
17	6.55	0.18
18	4.03	0.88
19	4.12	0.72
20	3.48	0.94
21	3.67	1.00
22	4.12	1.44
23	3.27	0.89
24	4.03	1.03
25	1.97	0.31
26	2.85	0.74
27	5.88	0.32
28	6.15	0.23
29	3.00	0.48
30	5.09	0.72

speech segment	mean scale	(SIQ) variance
31	3.94	0.41
32	4.82	0.85
33	5.42	0.66
34	2.52	0.98
35	2.94	0.37
36	3.09	0.74
37	3.30	0.98
38	3.88	0.90
39	3.61	0.85
40	4.61	0.84
41	4.88	0.80
42	2.61	0.73
43	2.64	0.69
44	4.73	0.32
45	2.48	0.76
46	6.12	0.68
47	4.06	0.35
48	5.39	0.87
49	5.55	0.94
50	3.30	0.77
51	4.42	0.77
52	2.76	0.99
53	3.70	0.96
54	2.76	0.74
55	4.27	0.81
56	3.88	0.26
57	3.85	0.28
58	3.79	1.04
59	5.18	0.74
60	6.27	0.67

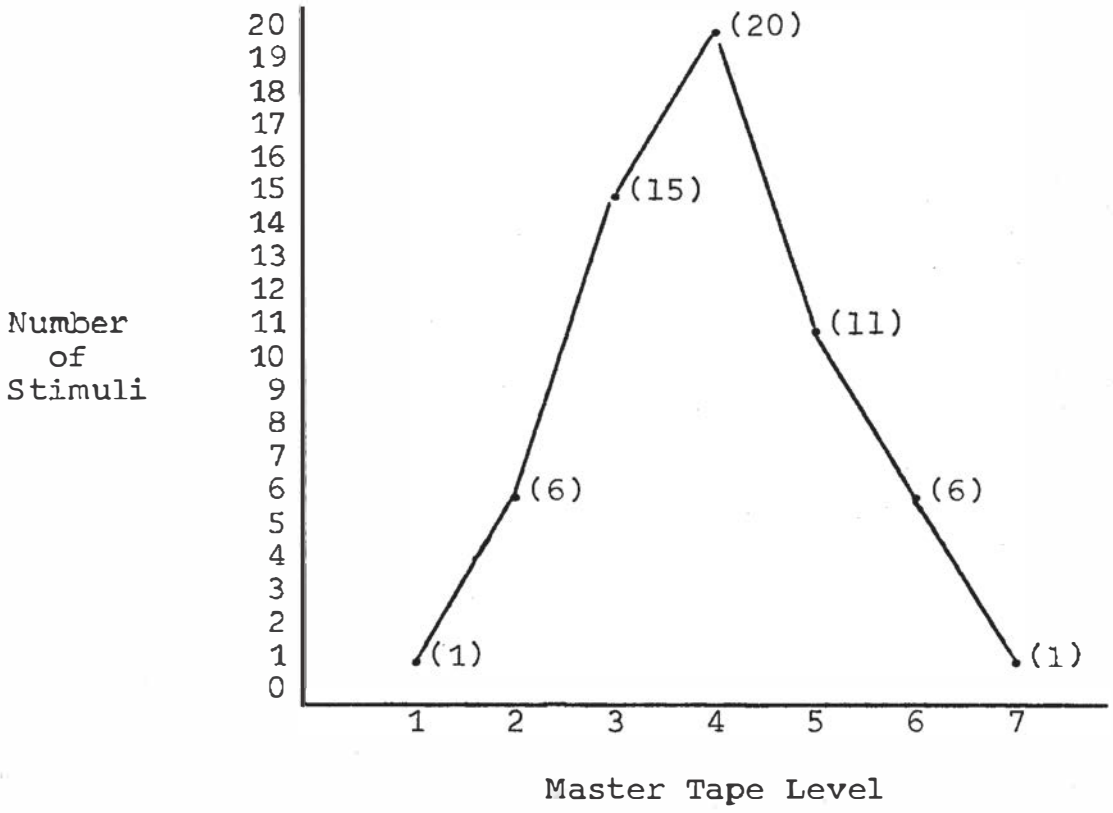
APPENDIX V

MASTER TAPE

LEVEL	speech segment	mean scale	(SIQ) variance
1	13	1.42	0.13
2	9	2.03	0.76
3	29	3.00	0.48
4	47	4.06	0.35
5	30	5.09	0.72
6	15	6.12	0.38
7	17	6.55	0.18

APPENDIX VI

DISTRIBUTION OF STIMULI
(determined by untrained observers' ratings)



APPENDIX VII

AGE, RACE AND SEX DISTRIBUTION OF SUBJECTS

Distribution of the 60 Subjects

GRADE	SEX	
	male	female
kindergarten	2	0
first	11	7
second	15	5
third	7	2
fourth	10	1

		SEX	
		male	female
RACE	Caucasian	41	10
	Negro	4	5

Distribution of the Master Tape Subjects

LEVEL	SEX	RACE	GRADE
Level 1	male	Caucasian	2nd
Level 2	male	Caucasian	2nd
Level 3	female	Caucasian	1st
Level 4	male	Caucasian	2nd
Level 5	male	Caucasian	1st
Level 6	female	Negro	1st
Level 7	female	Negro	1st

APPENDIX VIII

TRAINED OBSERVER RATING SHEET

RATING SHEET

speech segment 1. Level _____

speech segment 2. Level _____

speech segment 3. Level _____

speech segment 4. Level _____

speech segment 5. Level _____

speech segment 6. Level _____

speech segment 7. Level _____

speech segment 8. Level _____

speech segment 9. Level _____

speech segment 10. Level _____

APPENDIX IX

RELIABILITY OF TRAINED OBSERVERS' RATINGS

speech segment	Clin. 1 scale values	Clin. 2 scale values	Clin. 3 scale values	Clin. 4 scale values	Clin. 5 scale values	mean scale values	Vari-ance (SIQ)
1	3	2	3	1	4	2.60	0.92
2	5	2	3	3	4	3.40	0.92
3	7	4	5	6	6	5.60	0.92
4	6	4	4	3	5	4.40	0.92
5	1	2	2	2	1	1.60	0.50
6	6	3	4	3	6	4.40	1.50
7	4	1	2	1	2	2.00	0.83
8	7	6	7	7	7	6.80	0.10
9	2	3	3	4	3	3.00	0.25
10	2	2	3	2	4	2.60	0.75

INTRACLASS (AVERAGE) CORRELATION = 0.9422
INTRACLASS (ADJUSTED TREND) CORRELATION = 0.7653
INTRACLASS (UNADJUSTED TREND) CORRELATION = 0.6985

BIBLIOGRAPHY

Articles

Berry, Richard C. and Silverman, Franklin H. "Equality of Intervals on the Lewis-Sherman Scale of Stuttering Severity." Journal of Speech and Hearing Research, 15 (March, 1972), 185-90.

Cullinan, W. L.; Prather, Elizabeth M.; and Williams, Dean E. "Comparison of Procedures for Scaling Severity of Stuttering." Journal of Speech and Hearing Research, 6 (1963), 187-94.

Dickson, David. "An Acoustic Study of Nasality." Journal of Speech and Hearing Research, 5 (1962), 103-11.

Dudley, George C. "A Comparison of Three Psychological Scaling Methods for Evaluating Voice Quality." Unpublished graduate study, Eastern Illinois University, 1970.

Galloway, Sheri. "Cross Validation of the Length-Complexity-Index Screening Form." Unpublished graduate study, Eastern Illinois University, 1972.

Jordan, Evan P. "Articulation Test Measures and Listener Ratings of Articulation Defectiveness." Journal of Speech and Hearing Research, 3 (1960), 303-19.

Lafon and Guichard. "Procedés pour measure les résultats de la therapeutique de la voix de la parole et du language." Folia Phoniatrica, 23 (1971), 22-3.

Lintz, Lois B. and Sherman, Dorothy. "Phonetic Elements and Perception of Nasality." Journal of Speech and Hearing Research, 4 (1961), 381-96.

Morrison, Shelia. "Measuring Severity of Articulation Defectiveness." Journal of Speech and Hearing Disorders, 20 (1955), 347-51.

Prather, Elizabeth M. "Scaling Defectiveness of Articulation by Direct Magnitute-Estimation." Journal of Speech and Hearing Research, 3 (1960), 380-92.

Rees, M. "Some Variables Affecting Perceived Harshness." Journal of Speech and Hearing Research, 1 (1958), 155-68.

Sherman, Dorothy. "The Merits of Backward Playing of Connected Speech in the Scaling of Voice Quality Disorders." Journal of Speech and Hearing Disorders, 19 (1954), 312-21.

_____ and Cullinan, W. L. "Several Procedures for Scaling Articulation." Journal of Speech and Hearing Research, 3 (1960), 191-8.

_____ and Lewis. "Measuring the Severity of Stuttering." Journal of Speech and Hearing Disorders, 16 (1951), 320-26.

_____ and Linke. "The Influence of Certain Vowel Types on Degree of Harsh Voice Quality." Journal of Speech and Hearing Disorders, 17 (1952), 401-8.

_____ and Moodie. "Four Psychological Scaling Methods Applied to Articulation Defectiveness." Journal of Speech and Hearing Disorders, 22 (1957), 696-706.

_____ and Silverman. "Three Psychological Scaling Methods Applied to Language Development." Journal of Speech and Hearing Research, 11 (1968), 837-41.

_____ and Trotter, William D. "Correlation Between Two Measures of the Severity of Stuttering." Journal of Speech and Hearing Disorders, 21 (1956), 426-9.

Shriner, Thomas and Sherman, Dorothy. "An Equation for Assessing Language Development." Journal of Speech and Hearing Research, 10 (1967), 41-8.

Siegel, Gerald M. "Experienced and Inexperienced Articulation Examiners." Journal of Speech and Hearing Disorders, 27 (1962), 28-35.

Silverman, Franklin H. "An Approach to Determining the Number of Judges Needed for Scaling Experiments." Perceptual Motor Skills, 27 (1968), 1333-4.

Spriesterbach, D. C. "Assessing Nasal Quality in Cleft Palate Speech of Children." Journal of Speech and Hearing Disorders, 20 (1955), 266-70.

_____ and Powers, G. R. "Nasality in Isolated Vowels and Connected Speech of Cleft Palate Speakers." Journal of Speech and Hearing Research, 2 (1959), 40-5.

Strandberg, Twila D. "A Preliminary Evaluation of a Screening Method to Identify Harsh Voice Qualities in First Grade Children." Unpublished graduate study, Eastern Illinois University, 1969.

Thurman, Wayne L. "The Construction and Acoustic Analyses of Recorded Scales of Severity for Six Voice Quality Disorders." Unpublished Ph. D. dissertation, Purdue University, 1953.

Young, Martin A. "Predicting Ratings of Severity of Stuttering." Journal of Speech and Hearing Disorders. Monograph Supplement, 7 (1961), 31-54.

_____. "Observer Agreement: Cumulative Effects of Rating Many Samples." Journal of Speech and Hearing Research, 12 (1969), 135-43.

_____. "Observer Agreement: Cumulative Effects of Repeated Ratings of the Same Samples and of Knowledge of Group Results." Journal of Speech and Hearing Research, 12 (1969), 144-55.

Books

Fisher, Hilda B. Improving Voice and Articulation. Boston: Houghton Mifflin Company, 1966.

Guilford, J. P. Psychometric Methods. 2nd ed. New York: Mc Graw Hill, 1954.

Mager, Robert F. Goal Analysis. Belmont, California: Fearon Publishers, 1972.

Thurstone, L. L. and Chave, E. J. Measurement of Attitude. Chicago: University of Chicago Press, 1929.

Williams, Fredrick. Reasoning With Statistics. New York: Holt, Rinehart and Winston, Inc., 1953.

Wilson, Frank B. The Sound Of Disordered Voice. (tape recorded series) St. Louis: Dynamic Productions, 1972.

Winer, B. Statistical Principles in Experimental Design. New York: Mc Graw Hill, 1962.