1970

A Comparison of Three Psychological Scaling Methods for Evaluating Voice Quality

George C. Dudley
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.

May 11, 1970
Date

Author

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

_____

_____

_____ _____

Date Author

A Comparison of Three Psychological Scaling Methods for Evaluating Voice Quality

(TITLE)

BY

George C. Dudley

THESIS

SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF

Master of Science

IN THE GRADUATE SCHOOL, EASTERN ILLINOIS UNIVERSITY
CHARLESTON, ILLINOIS

1970
YEAR

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

May 11, '70
DATE

May 11 '70
DATE

TABLE OF CONTENTS

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. Wayne L. Thurman, my advisor for both my undergraduate and graduate study and chairman of my thesis committee, I wish to extend a sincere word of appreciation. I should also wish to extend my gratitude to Dr. Jerry Griffith, committee member, and to Dr. B. F. McClerren, who although a professor in another department, graciously accepted the responsibility of being a member of my thesis committee.

To Mr. Lynn E. Miner, I wish to extend a special "Thank you" for generous time and excellent constructive guidance which allowed me to refine and to handle to statistical procedures involved in this research.

To the university professors who allowed me to use their classes and to the students who sacrificed class time to perform the scaling tasks.

My final expression of appreciation goes to Ellen Scott, my patient typist for her skillful preparation of this manuscript.

Chapter I

STATEMENT OF PROBLEM

A simple communication situation arises during interaction between a speaker and a listener. The message involved during such interaction is a "perceptual event." (Young, 1969) Assuming that the auditory channel of the listener is intact, the conductive medium is free from excessive ambient noise, and the content of the message is within the linguistic concepts of the listener; the amount of interference in the reception of the message is in the listener. Interference to the listener may depend largely upon the speaker's articulation, fluency, language usage, or voice quality. Since interferences are perceptual events, the amount or type of perceived interference may vary from listener to listener. "To depend on observers for measurement is to recognize that classifying speech as defective requires the judgment of an observer." (Young, 1969) Thus a logical research approach to measuring perceived interference in a spoken message would be to quantify judgments of a listener population.

Edwards (1957) has described a general psychological scaling method used by Thurstone which could be applied to measurement of a perceptual event such as speech by a listener

population. Essentially this method uses an observer population to judge a given statement, not in terms of agreement or disagreement, but rather in terms of degrees of favorableness or unfavorableness. The result is a scaling of that statement about a "psychological object" onto a continuum of varying degrees of favorableness or unfavorableness by a judging population. A psychological object is "any phrase, slogan, person, institution, ideal, or idea toward which people can differ with respect to positive or negative affect." (Edwards, 1957) A simple illustration of the Thurstone equal-appearing interval continuum is illustrated in Figure 1. Varying degrees of unfavorableness toward a given statement are represented by letters A, B, C and varying degrees of favorableness toward the statement are expressed by letters E, F, G. Thus one may visualize the formation of a psychological continuum representing a range of degrees of attitudes expressed toward the presented statement. The D point, or the "neutral" (Edwards, 1957, p. 84) interval is essentially a zero point on the continuum.

FIGURE 1. Thurstone equal-appearing interval continuum

The cumulative judgments of a population of observers for each particular statement can be converted to scale values.

These scale values indicate the proportion of judgments made in each category of degrees ranging from least to most favorable.

Application of psychological scaling methods to research in speech pathology is relatively new. The first published study (Lewis and Sherman, 1951) reported use of a nine-point equal-appearing interval scale to measure stuttering severity.

Since that initial study, subsequent studies have used listeners, both trained and untrained, to rate severity of articulation, stuttering, language, and voice quality. Observer methods have differed only in the manner in which judgments and scale values have been obtained. Thus acoustical events can be judged and classified by listener responses that represent a validation for judgment or measure of severity of a given perceptual event.

Although scale values for disordered speech have been obtained from the classical scaling usages, there are important differences. "The stimulus dimensions of disordered speech are nonmetric and multidimensional." (Young, 1969) Speech stimuli may differ from speaker to speaker, from conversational speech to reading, and even from varied speaker stimuli when reading word lists. (Young, 1969) Previous research, (Jordan, 1960), cites that dimensions to be measured such as articulation defectiveness, are affected by other related dimensional paramenters such as frequency or severity of error when rated by an observer population. However an

articulation defective sample can be numerically documented for frequency and type of error by recording from live speech, tape recorded speech, or transcription. (Curry, Kennedy, Wagner, and Wilke, 1943; Henrikson, 1948; and Barker, 1960) Listeners, although receiving a multidimensional interference when rating stuttering severity can document severity by numeric measures such as frequency of repetitions (Lewis and Sherman, 1951; Sherman and Trotter, 1956; and Young, 1961) and speech rate (Bloodstein, 1944 and Johnson, 1961). A listener given the task of rating language development may listen for and document syntactical structure, vocabulary, mean length of response (Johnson, Darley, Spriestersbach, 1952, p. 167), length - complexity (Shriner, 1967), transformations (Menyuk, 1963), and other measures of language development.

Voice quality appears to represent the ultimate in multidimensionality. The listener given the task of judging voice quality faces multiple stimuli interference from articulation, fluency, language, and the message content. Furthermore he is judging a perceptual event and has no transcription record available.

One major task facing the listener lies in the actual perception of the presented voice quality. Each listener may perceive the same speech sample as representative of different voice qualities. In other words, each has listened to the vocal quality but has perceived various characteristics in the same sample. One listener judge may describe the perceived

sample as being representative of "harshness" while another listener might refer to the same sample as "husky." This perceptual problem has resulted in a long list of adjectives describing the same voice sample.

Unlike the situations in articulation, stuttering, and language judgment, no measures of severity have been found that can be applied to judgment of voice quality. Voice quality is a perceptual event. Hence each listener has his own internal reference points as to when voice quality is deviant, as to when it interferes with communication, and as to the nomenclature of what he perceives.

In a scientific reference, experiments are performed to evaluate hypotheses. Thus the primary purpose of this study is to evaluate the following hypothesis. Stated in the null form: There is no significant difference among reliability of measures of data gathered in judgments of voice quality problems by equal-appearing intervals, successive intervals, and direct magnitude estimation.

Secondly, an experiment could indulge the experimenter's curiosity. Questions to be answered in this study are:

1. Can naive or untrained listeners reliably judge the severity of samples of voice quality deviations?

2. If scaling methods can be used to rate severity of voice quality deviations, which method, equal-appearing intervals, successive intervals, or direct magnitude estimation, will be most reliable for evaluative purposes?

Thirdly, an experiment should attempt a new technique or approach, should strive to improve a current or known technique or represent an extension of an old technique into new areas. The equal-appearing intervals scaling technique has been used for rating articulation, stuttering, language, and voice. Chapter II will reveal studies which have compared various scaling techniques for the purpose of searching for improved means for rating articulation, stuttering, and language performance. Voice quality still is rated by the equal-appearing intervals method. No known study has compared scaling methodologies in attempting to seek an improved means for rating voice quality in terms of observer reliability, in experimental practicability, and in manipulating computational data.

An extension of techniques from this study would yield scale values of paramenters of voice representing degrees of perceived voice quality which may be applied to training listeners for judging similar perceptual events.

Chapter II

REVIEW OF LITERATURE

Psychological Scaling Literature.

When an experimenter uses psychological scaling methodologies to evaluate speech production, he is essentially asking listeners to make comparative judgments of the presence or absence of acoustical characteristics which affect communication of the speaker. Young (1969) states that, "observers are frequently used in clinical and experimental settings to evaluate speech disorders on a variety of perceptual dimensions." Review of the literature indicates that psychological scaling methodologies can be applied to research in speech pathology. This is a useful procedure because listener judgments of perceptual events can be quantified to represent a single judgment of severity for a presented speech sample.

The three psychological scaling methods frequently employed in communications research are: (1) equal-appearing intervals, (2) successive intervals, and (3) direct magnitude-estimation.

Equal-appearing intervals.

Sherman and Moodie (1957) describe this method as one in which "the observer is instructed to assign numbers to the stimuli in relation to an equal-appearing scale of severity."

The principle assumption underlying this method is that the observer can reliably equate intervals or distances between responses to stimuli. The equal-appearing intervals (EAI) scaling method was chosen for comparison in this study because of its common use in experimentation with speech disorders as evidenced in Chapter I.

Thurstone and Chave (1929) originally described the method of equal-appearing intervals. They assumed that a judge's attitudes toward the object being scaled would not affect reliability. Edwards (1957) indicated that this method required each observer to make only one comparative judgment for each stimulus presented.

Guilford (1954) presents some advantages for using EAI rating methods: 1. EAI requires much less experiment time than either pair comparisons or ranking methods. 2. EAI can be used with "psychologically 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 much wider range of application than do ranking or comparing methods. 5. EAI is assumed to yield interval data, which is a higher form of data than nominal or ordinal data. 6. Some experimenters maintain that best judgments are made when stimuli are presented singly; comparative scales destroy the "aesthetic attitude" of the rater.

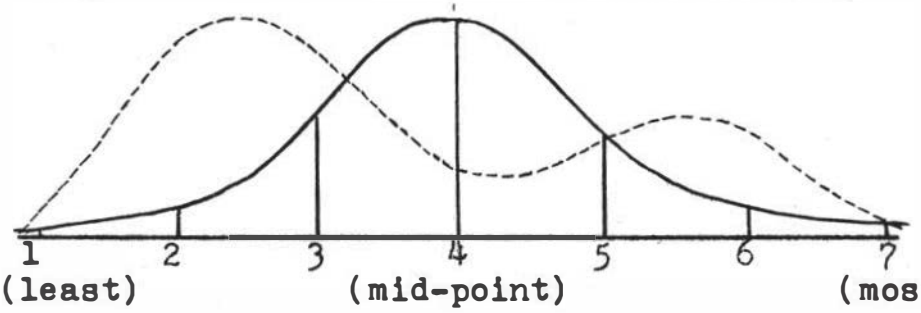
When rating by EAI, observers make judgments about the presented stimuli, usually in reference to their own anchor

points, such as least unfavorable or least severe to most unfavorable or most severe. This particular procedure measures observer's internal standards in relation to their preconceived attitudes of least to most severe. However, good EAI scaling usually ties down end points by initially presenting the entire range of attributes to be scaled. Thus, cumulative observer judgments can be used as a yardstick to measure the given range of presented attributes. The center interval ideally represents the mid-point of the distribution of assigned values along the continuum. Each point is of equal distance from the adjacent point. Thus, if an observer assigned the first stimulus a value of "three", theoretically a stimulus of "six" should be twice as severe as the former stimulus. A stimulus value of "seven" should theoretically be exactly one point more severe than an assigned stimulus value of six. Figure 2 provides a graphic illustration of the assumption of equal-appearing intervals.

This scaling method can, however, have one obvious disadvantage. The resulting stimuli assignments can produce an end-effect, or a piling-up of judgments at one or both ends of the scale. For instance, an observer instructed to rate a series of stimuli on a seven-point scale might hear a stimulus that represents the most severe sample he has heard according to his own concept or anchor point. He would probably assign this particular stimulus a value of seven. However, during the course of the experiment he might hear a stimulus

that appears to be more severe than the stimulus previously heard and rated seven. This situation might occur several times during the experiment and result in the distribution of judgments toward the upper range of the continuum. Thus, the scale values are not of equal distance along the range of judgments. Instead there is an abundance of values at extremes of the scale rather than at the mid-points of the scale.

Fig. 2. Normal curve distribution with assumed equal-appearing intervals. (Guilford, 1952, p. 34.)



Lewis and Sherman(1951) applied a nine-point equal-appearing intervals scale to measurement of severity of stuttering. A graphic illustration of the number of samples in each of the eight severity intervals showed a distribution of ratings far from normal. There was a definite peaking at the least severe end with a marked dip at severity values of three and four. In other words, there was an end-effect. The results of their study are illustrated by the broken lines in Figure 2.

True equal-appearing intervals scaling procedure should require two presentations of the same stimuli. The observer population should merely listen during the initial presentation to perceive the end-points of the continuum. The actual rating

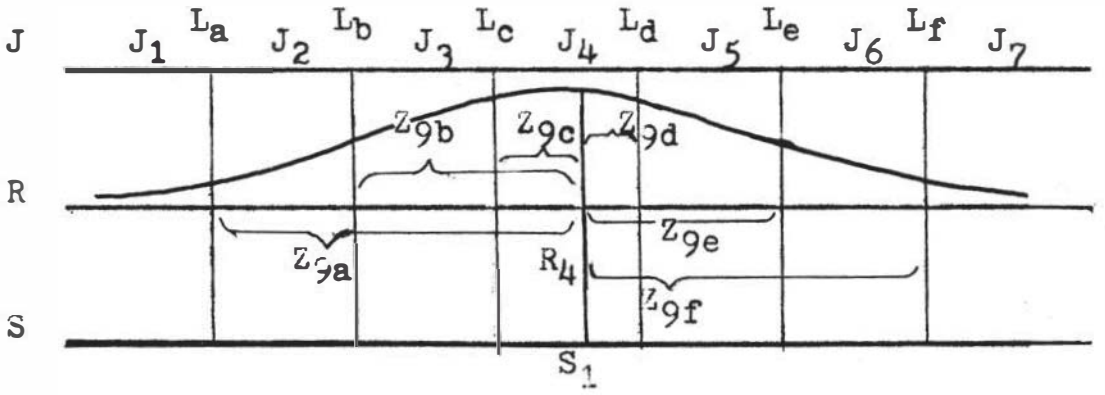
should be performed during the second presentation. Lewis and Sherman may have experienced the end-effect in their study as the result of failing first to present the taped samples prior to the actual rating task.

Despite the mentioned disadvantage, equal-appearing intervals scaling has been used extensively. The method does offer simple computational procedures.

Successive intervals.

Sherman and Moodie (1957) describe successive intervals as being aimed at reducing the end-effect produced by equal-appearing intervals scaling methods. According to Guilford (1954), the experimental operation in successive intervals is essentially "that of judging each of several stimuli as belonging in one of a limited number of categories differing quantitively along a defined continuum." He continues: "No assumption is made concerning the psychological equality of category intervals." The only assumption made is that the "categories are in correct rank order and that their boundary lines are stable except for sampling errors." Figure 3 offers a graphic illustration of the concept of the successive intervals methodology. (Guilford, 1952, p. 34.)

Figure 3. Discriminal dispersion extends over seven successive categories of judgments, J_1 to J_7, with limits between categories, L_a to L_f. The distances from these limits are given by the respective standard measures Z_{9a} to Z_{9f}.



The seven categories are labled J_1 to J_7. Within the seven categories there are six limits, L_a to L_f. Stimulus S_1 is shown to be dispersed through all seven of these categories. The mean of the distribution on R has its "modal discriminal process" (Guilford, 1954), at R_4. If one assumes a normal distribution of the deviations from R_4 by knowing the proportion of judgments in each category limit, one can express that distance of each category limit from R_4 in terms of a z value. After determining the distances of all limits from R_4, the common reference point, one may find by subtraction the distances between limits themselves. By this process one can determine whether widths of categories are equal, and if they are not, can see what the relative widths are. The successive intervals method is essentially interested in the number of judgments that occur within previously assumed equally distant spaced categories.

One apparent advantage of successive intervals is that scale values can be applied to equal-appearing intervals data. Sherman and Moodie (1957) and Silverman and Sherman (1967) made such application of successive intervals to equal-appearing intervals. Guilford (1954) briefly evaluated successive intervals:

> The experimental operations for obtaining judgments in successive categories (successive intervals) are so simple and economical from the standpoint of both investigator and observers that from this point of view the method has everything in its favor.

Silverman and Sherman (1967) somewhat disagree with Guilford's statement about economy of investigator time. They report that the procedure used to derive successive interval scale values is far more complex and time consuming than deriving equal-appearing intervals scale values.

Direct magnitude-estimation.

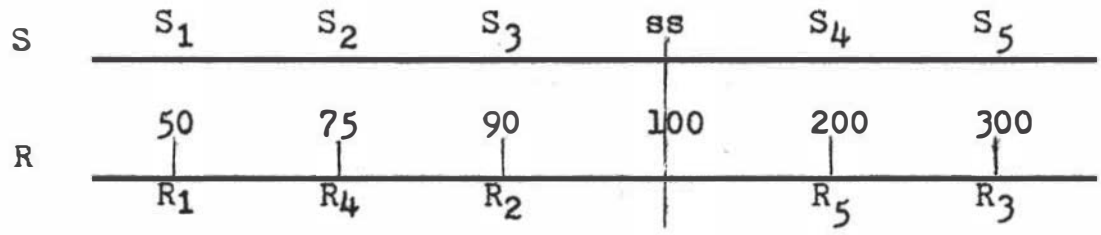
The four levels of measurement listed in an ascending level order from lowest to highest are nominal, ordinal, interval, and ratio. The naming or assigning of frequency values to data such as a two, three, or six in categories represents a nominal level of measurement. Ordinal data represents a rank order value level of measurement. For example, results of a horse race represent ordinal data. Interval level measurement yields a comparative distribution of data, assumed to be in equal intervals, along a continuum in relation to normal. Ratio level measurement uses an absolute zero and value scores are reported in relation to that absolute.

The experimenter chose the direct magnitude-estimation psychological scaling method for the purpose of applying a ratio scale to rating voice quality and for the purpose of comparing a ratio scale to interval scales. Ratio data should theoretically yield a higher level of measurement than interval data. Prather (1960) and Sherman and Silverman (1968) report that a ratio scale, compared to interval scales, has the advantage of an absolute zero, a feature which permits use of ratios of scale numbers in all numerical and statistical operations. This feature makes results more meaningful in that judgments are not made on an interval scale but are made in proportion to an absolute zero.

Prather (1960) states that this method involves presenting stimuli one at a time to a group of observers. The experimenter may assign a number to the first stimulus which is to be used as the standard. For succeeding samples observers assign numbers for respective stimuli in proportion to the standard along the continuum of measurement. For example, the experimenter may first present a stimulus which he has assigned a standard of 100. He will continue to present each stimulus to the observer one at a time and have that observer assign whatever numbers represent the relative position of each stimulus on the continuum in proportion to the standard stimulus of 100. If the observer perceives the first stimulus to be twice as severe as the standard, he would then assign a value of 200 to that stimulus. If the second presented stimulus

appeared to be only half as severe as the standard stimulus, the observer would assign a stimulus value of 50. There is no limit placed upon observer assignment of scale values. Stevens (1956) stresses that when using direct magnitude-estimation scaling the observer should be "completely free to decide what number he will assign to the variable." Figure 4 illustrates direct magnitude-estimation.

Figure 4. One observer's ratings of five stimuli by DME. Let S represent the stimuli presented and S_1 to S_5 represent each stimulus. Line R represents the observer response with R_1 to R_5 indicating the severity of S in proportion to the Standard Stimulus (SS = 100)



Prather (1960) essentially found no diffenence between judgments made when the standard was presented to observers at the beginning of the experiment and when the standard was presented after every fifth sample.

Speech pathology literature.

Previous investigations provide strong evidence that psychological scaling methodologies have been successfully used to rate articulation, language, stuttering, and voice. Furthermore, of the various methods available, the method of equal-appearing intervals appears to be the most widely used method for quantifying listener ratings.

Articulation severity has been scaled by the equal-appearing intervals and direct magnitude-estimation scaling methods.

Morrison (1955) concluded that equal-appearing intervals scale values could be used to reliably judge articulation severity from both five- and ten- second speech samples. Sherman and Morrison (1955) did a follow-up study to determine whether they could obtain reliable intervals scale values of articulation defectiveness from ratings of one-minute speech samples by trained individual observers. Judges, trained by the two tape recorded severity scales from Morrison's (1955) study, rated one-minute speech samples. The investigators concluded that trained observers, using equal-appearing intervals scales, could rate articulation of five- and ten-second segments as reliably as with one-minute samples of continuous speech. That is, observers tended to rank order the stimuli in the same manner for three different intervals of presentation.

Sherman and Cullinan (1960) had 14 graduate students majoring in speech pathology to rate severity of articulation defectiveness for 50 one-minute tape-recorded samples of children's speech. The observers used a nine-point equal-appearing intervals scale to rate consecutive 10-second segments from each one minute sample; mean scale values were computed for each observer. The same 50 one-minute speech samples were scaled on a nine-point equal-appearing intervals

scale by 11 additional judges who rated each sample as a whole. Pearson rs were used for comparison of (a) judging segments at consecutive intervals, (b) judging one-minute samples as a whole, and (c) judging randomized segments. The latter were mean scale values obtained by Sherman and Morrison's (1955) study. The Pearson r for estimating the relationship between the 50 mean scale values derived from judgments made at consecutive intervals and the 50 mean scale values values derived from judgments of samples as a whole was .99. The Pearson r for estimating the relationship between the 50 mean scale values derived from judgments of randomized segments, consecutive intervals, and judgments of samples as a whole was, in each case, .98. The high correlation (.98) indicated a strong relationship between any two sets of measures obtained by judging at consecutive intervals, whole samples, or randomized segments. Consequently, they concluded that any one of the above stated methods can be used to rate reliably severity of articulation defectiveness.

Jordan (1960) studied the relationship between articulation test measures and listener ratings of articulation defectiveness. By means of multiple regression analysis, he evaluated relationships between 22 measures obtained by phonetic analysis of 150 children's articulation test responses and measures of defectiveness of articulation obtained by observer ratings of their connected speech. One hundred fifty tape recorded 30-second speech samples were rated on a nine-point

equal-appearing intervals scale by 36 observers. Results essentially indicated that observer's reaction to articulation defectiveness are primarily dependent upon frequency (.90) and severity (.70) of articulatory deviations.

Prather (1960) evaluated the usefulness of the method of direct magnitude-estimation (DME) for scaling defectiveness of articulation. Twenty seven five-second continuous samples of children's speech were rated by 200 students enrolled in an elementary psychology course. The total observers, subdivided into five groups, participated in six different experimental conditions: Condition I, standard of medium severity, assigned a value of 100, presented only at the beginning of the experiment; Condition II, standard of medium severity, assigned as 10, presented only at the beginning of the experiment; Condition III, standard designated as 100, presented before every sixth stimulus; Condition IV, same standard stimulus as Condition I, II, III, no specific point assignment; Condition V, same as Condition I, with same observers who had participated in Condition IV exactly one week later; Condition VI, standard of mild severity assigned as 10, presented only at the beginning of the experiment. Under each condition observers rated samples four times to compare effects of several sequences and to evaluate effects of practice. The high correlation range (.94 to .98) evidenced that neither sequence of presentation of practice effects had any important effects on obtained scale values. Scale values did not depend upon

the assignment of specific standard stimulus values or whether the observer made his own point assignments. However when the assigned stimulus was 10 points, the scale was relatively extended at the upper end as compared to the assigned stimulus of 100 points. Finally, there was no apparent advantage in frequent presentation of the standard stimulus over a single presentation at the beginning of the experiment.

The following summary statements may be made regarding the application of psychological scaling methodologies to rating articulation severity. Both equal-appearing intervals and direct magnitude-estimation methods have been successfully used to rate articulation severity. The nine-point equal-appearing intervals scale appears to be the most commonly used scale for rating articulation severity.

The equal-appearing intervals psychological scaling method also has been applied to observer rating of language development. The Shriner and Sherman (1967) study shows the relevance of psychological scaling to language development. Three hundred language samples consisting of 50 responses to picture stimuli or to examiner questions were used in this study. The following measures were obtained for each of the 50 responses: mean length of response, mean length of the five longest responses, number of one word responses, standard deviation of response length by number of words, number of different words, and structural complexity score. Stimuli were presented to 104 judges, who were students in Speech Pathology and Audiology

and who had previously been enrolled in a course in language development. The stimuli were presented in typed, mimeographed form. Samples were rated on a seven-point equal-appearing intervals scale with one representing the least development of language and seven representing the most development of language. A multiple R of .85 was obtained when a multiple-regression analysis in which all six predictor variables were used. This was interpreted to mean that the above predictors of language development cannot be used reliably to assess language development. Mean length of response had a higher correlation (.80) with obtained scale values than did any other predictor variable. Thus it would appear that mean length of response, if used as a single measure for assessment of language development, would be most useful among those studied.

Sherman and Silverman (1968) compared equal-appearing intervals, successive intervals, and direct magnitude-estimation scaling methodologies for usefulness in measuring language development in samples of children's speech with reference to 'intricacy of language usage.' Their stated operational definition was "the intricacy of the arrangement of words for the purpose of conveying information." Fifty language samples, typed mimeographed form, were presented to 62 university students who rated the 50 language samples on a seven-point equal-appearing intervals scale. None of these observers had had extensive course work in language development of children. Successive intervals computational procedures were applied

to equal-appearing intervals data. For the method of direct magnitude-estimation, the same 50 language samples, arranged in a different random order, were rated by an additional 42 naive observers. The standard sample was assigned a stimulus value of 100. A comparison between equal-appearing intervals and successive intervals methodologies (r = 0.995) revealed that both sets of scale values rank ordered the 50 samples in "almost identically the same manner." Comparison between equal-appearing intervals and direct magnitude-estimation yielded a correlation of 0.92. Sherman and Silverman concluded that scale values obtained by the three methods did not appear to differ in their usefulness for the kind of stimuli presented. However because of simpler computational procedures, equal-appearing intervals is preferred for obtaining scale values for rating intricacy of language.

There is particular significance in the relevance of psychological scaling methods to rating stuttering severity. The first application of psychological scaling to speech pathology was in rating stuttering severity. Lewis and Sherman (1951) applied a nine-point equal-appearing intervals scale to measures of stuttering severity. Thirty elementary psychology students, employing the equal-appearing intervals scale, rated 240 samples of stuttered speech. Ninety six of the original 240 samples were then presented to 106 elementary psychology students to rate in order to check internal consistency; that is, whether the scaling method yielded the same

results on successive application. The obtained Pearson $\underline{r}^s$ of .98 and .97 "strongly indicated that the scale of severity obtained in the study was a rather precise one."

Sherman and Trotter (1956) used a nine-point equal-appearing intervals scale to compare listener judgment of stuttering severity and frequency. They found a close correlation (.81) between the two factors. In other words, scale values tended to increase as judgments of severity and frequency of stuttering increased. This obtained correlation however did not indicate a one-to-one relationship between the measures.

Young (1961) presented 50 tape recorded samples of speech, 200 words in length, to 48 listeners. The listeners were divided into three categories: Group I (stutterers), Group II (clinicians), and Group III (laymen). Scale ratings were compared to predicted measurements of disfluency and rate of utterance. Listener agreement was measured by means of intraclass correlations. The coefficient for Group I was .79, Group II was .83, Group III was .87, and the combined reliability measure was .83. The types of disfluencies that appeared to be associated with judgmental ratings were syllable or word repetition, sound prolongations, broken words, and words involving apparent or unusual stress or tension.

The following summary might be stated regarding application of psychological scaling methods to rating stuttering severity. Scaling methods have been successfully employed to rate stuttering severity. The obtained scale values from

psychological scaling have in stuttering studies, as in articulation and language studies, provided a validation of other predictor measures for severity.

Finally, investigators have used psychological scaling methods, particularly equal-appearing intervals, to rate severity of perceived voice qualities. The following studies are offered as evidence to application of scaling methods to voice quality.

Sherman and Linke (1952) first applied equal-appearing intervals scale values to determine whether variations of vowel content in controlled speech samples had any effect upon perceived harshness. 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. This method was used to eliminate irrelevant judgment variables such as articulation. She used a seven-point equal-appearing intervals scale for rating both harshness and nasality. 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. Sherman concluded that although some irrelevant judgment variables had been eliminated by backward playing of speech samples, no advantage was gained in judgment reliability.

Rees (1958) had 32 listeners rate syllables of twelve speakers with clinically diagnosed harsh voices on a seven-point equal-appearing intervals scale. The mean Q-value for the 1080 scale values was .79 which Rees considered to be "satisfactorily reliable." She concluded that the method of scaling could be used to study the influence of vowels, selected consonant environments, and vowel initiation on perceived harsh voice quality.

Spriestersbach (1955) used a seven-point equal-appearing scale to investigate the influence of articulatory defects upon judgments of nasality. Thirty-second speech samples of 50 cleft palate children with cleft palate speech were obtained. Judgments of severity of nasality were made when the samples were presented forward and when presented backwards. Judgments of defectiveness of articulation and effectiveness of pitch variation were made when the samples were played forward. Results indicated that trained observers were able to make "stable" judgments of severity of nasality when the samples were presented backwards (.90) but articulation defectiveness appeared to affect severity of nasality when samples were played forward (.69).

Spriestersbach and Powers (1959) evaluated the relationship between connected speech and isolated vowels on perceived nasality. Recordings were made of seven vowels and of connected speech (played backwards) produced by 50 children with cleft palates. These recordings were scaled for severity on a

seven-point equal-appearing scale by 30 judges who were advanced students in speech pathology. The correlation coefficients for the severity judgments ranged from .47 to .60. The investigators concluded that severity of nasality in connected speech is related to severity of nasality for each isolated vowel studied.

Lintz and Sherman (1961) studied the influence of vowel quality and consonant environments upon nasality. Twenty adult male subjects recorded vowels and consonants in isolation and in CVC syllables. Judges, 35 advanced students with training in voice quality deviation diagnosis, rated perceived nasality on a seven-point equal-appearing intervals scale. A correlation of .89 was obtained as an estimate between the two sets of scale values for the first 100 samples. The investigators concluded from the scaling method that "degree of perceived nasality varies with fundamental frequency, duration, and intensity of vowels."

Dickson (1962) made an acoustic study of nasality. The The vowels /i/ and /u/ in the words "beet" and "boot" were recorded for each of 60 subjects. Each stimulus was rated by five experienced phoneticians using a seven-point equal-appearing intervals scale of nasality. Each judge rated each word twice, thus providing a means of estimating the reliability of the participating judges. Rank order correlations between the two ratings ranged from .63 to .81. The sound spectrograph was then utilized to analyze the stimuli for acoustical

determinants of nasality. Sound spectrograms which appeared to represent acoustical determinants of nasality were correlated with judgmental ratings. The acoustical determinants of nasality were correlated with judgmental ratings. The acoustical-perceptual correlations were .79 for the /i/ and .64 for the /u/. In other words, there appeared to be "little relationship between the initial classification of subjects as normal or functionally nasal and the degree of judged nasality on the two vowels studied."

In summary, the equal-appearing intervals scaling method has been the only known method applied to judgmental rating of perceived voice quality. The seven-point scale has been used exclusively in previous voice studies. Irrelevant judgmental variables such as articulation still are believed to effect judgments by the listening population assigned the task of rating severity of voice quality. (Sherman, 1954) Other than Shermsn's attempt to eliminate irrelevant judgmental variables by backward playing of the stimuli, no studies have been applied to the need for more reliable means for rating voice quality severity. Furthermore, no known study has used, or compared the use, of different scaling methodologies. There is no logical basis to assume EAI is preferred method to scale voice quality.

To this point, the review of previous research has cited evidence that psychological scaling methodologies have been applied to various parameters of speech pathology. Investigators

have not only used equal-appearing intervals, successive intervals, and direct magnitude-estimation methodologies in rating speech but have compared scaling methodologies in articulation, language, and stuttering.

Several studies have attempted to compare psychological scaling methods for purposes of quantifying attributes of disordered communication. Comparison among methods for rating severity of articulation first will be reviewed.

Sherman and Moodie (1957) compared equal-appearing intervals, successive intervals, pair comparisons, and constant sums scaling 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, they concluded, was most useful for scaling articulation defectiveness.

The following study compared scaling methods in attempting to find the most reliable means for assessing attributes of language development. Sherman and Silverman (1968) compared

equal-appearing intervals, successive intervals, and direct magnitude-estimation. Observers rated typed samples of speech, one sample for each of 50 children. The two sets of scale values derived from the same data by equal-appearing intervals and by successive intervals ranked the 50 samples almost identically. The correlation between the two sets of values was .995. This correlation was of the same magnitude as was reported between equal-appearing intervals and successive intervals scale values for other stimuli as reported by Silverman and Sherman (1967). They found a correlation of .92 between direct magnitude-estimation values and the mean scale values of equal-appearing intervals. Sherman and Silverman concluded that "scale values obtained by the three methods appear to differ very little in their usefulness, at least for the kind of stimuli used in this study. They stated that because of simpler computational procedures, equal-appearing intervals scaling techniques are often preferred.

The following study compared scaling methods to determine the best technique for assessing severity of stuttering. (Cullinan, Prather, and Williams, 1963) They compared the results of severity of stuttering ratings by six variations of equal-appearing intervals and by those from direct magnitude-estimation. Stimulus material, consisting of 27 20-second tape recordings representing the continuum of severity of stuttering from very mild to very severe, were rated by 128 undergraduate students enrolled in a communication skills class.

Samples were rated for: severity of stuttering on a five-point scale (I); seven-point scale (II); nine-point scale, little definition of points (III); seven-point scale, points defined at length (IV); "likeness to normal speech" (V); "easiness to listen to" on a seven-point scale (VI); severity by direct magnitude-estimation (VII). A different group of judges was used for each of the seven rating conditions. Interjudge reliability coefficients for the equal-appearing intervals rating ranged from .95 to .97 but the interjudge reliability coefficient for the method of direct magnitude-estimation was lower (.90).

Research comparing the usefulness of rating articulation severity found equal-appearing intervals generally to be the most practical, with successive intervals, and direct magnitude-estimation also yielding reliable judgments. The study, (Sherman and Silverman, 1968), that compared equal-appearing intervals, successive intervals, and direct magnitude-estimation found all three yielding reliable judgmental ratings for evaluating language development. However, Sherman and Silverman preferred using equal-appearing intervals because of simpler computational procedures. Comparison of equal-appearing intervals and direct magnitude-estimation in rating severity of stuttering found equal-appearing intervals gave higher judgment reliability.

Although investigators have compared, and attempted to determine the best, and most reliable scaling method for the

above speech disorders, none have compared scaling methods to find the best method for rating severity of voice quality.

A summary of the review of previous studies concerning perceptual rating of voice quality leads to the finding that the majority of scaling studies of voice quality disorders has been done by the method of equal-appearing intervals. Cullinan, Prather, and Williams (1963) compared five-, seven-, and nine-point equal-appearing intervals scaling methods to rating stuttering severity. These investigators concluded that there were essentially no differences among interjudge reliability ratings obtained from either three of these psychological scaling methods. Apparently, stimuli rank order themselves in the same manner regardless of the EAI scale length. On the basis of the above mentioned studies, this investigator decided to use a seven-point equal-appearing intervals scale to rate degree of "unpleasantness" of voice quality samples.

This investigator reviewed the literature to determine whether trained or untrained observers should be used to rate the voice quality samples to be presented in this study. Some investigators compared the reliability of observations of untrained listeners versus the reliability of observations made by trained listeners. Perrin (1952) investigated the question whether untrained observers could use the method of paired comparisons to rate functional articulation defects. Untrained observers were enrolled in a basic psychology course. The trained observers were enrolled in a course in

clinic methods in speech correction. Perrin found that the observers did not differ significantly (.82) in their evaluation of severity of articulation defects.

Morrison (1955) had both trained and untrained observers rate samples of severity of articulation defectiveness. Each group used a nine-point scale to rate both five- and ten-second speech samples. The differences between the two groups of observers were small and nonsignificant (0.11).

Young (1961) essentially used trained and untrained observers when he had clinicians, stutterers, and laymen rate severity of stuttering samples. The reliability for the combined three groups was 0.83. This indicated that both trained and untrained observers tended to agree when rating stuttering severity.

Siegel (1962)compared "experienced" and "inexperienced" articulation examiners. Two experienced (graduate students in speech pathology) and two inexperienced (women who had been classroom teachers) observers made judgments of correct, incorrect, or unscorable on responses to a modification of the Templin-Darley articulation test. The experienced observers received no training. The inexperienced observers received training after the first listening session. The inexperienced observers correlated (r = 0.92) before training. Correlations among scores of two experienced and two inexperienced articulation examiners on three occasions were .97, .99, and .96 respectively.

No previous research found comparisons between trained and untrained observers when rating voice quality samples. Research by Perrin (1952), Morrison (1955), Young (1961), and Siegel (1962) indicated little or no significant differences between judgments by trained or untrained observers in rating severity of articulation or stuttering. As a result of the findings by the above cited investigations, untrained observers were used in this study upon the assumption that there would be little significant difference between trained and untrained observers in rating voice quality samples.

A review of Chapter II indicates that psychological scaling methodologies can be applied to speech pathology. The need for this study is again emphasized by the following concluding statement. Although there have been comparative studies made in attempt to determine the best, or most reliable technique to rate perceptual judgments of articulation, language development, and stuttering, no study has attempted to determine the most reliable methodology for rating voice quality.

CHAPTER III

PROCEDURES

Preparation of stimuli.

The voice quality samples were elicited from 42 children, 27 boys and 15 girls, enrolled as first graders in public schools. These children had been selected from a population of first grade children from the East Central Illinois communities of Charleston, Mattoon, and Sullivan. Each of the 42 subjects had been diagnosed as having harsh voice quality by one of five speech clinicians serving those respective communities. All subjects used in this study had been identified in a previous study. (Strandberg, 1969). None of the children had yet been enrolled in voice therapy. The public school clinicians had identified these children by evoking a minimum of 15 seconds of spontaneous speech from each child. Clinicians had used the Curtis definition of harsh voice quality as stated by Rees (1958): 'Harsh voice quality has an unpleasant, rough, rasping sound. It is often heard in people for whom voice production seems to be a considerable effort or strain.'

Four of the five clinicians who had assisted Strandberg in the original identification had attained the M.S. in Speech Pathology and Audiology and had at least one year of professional

practice. The fifth clinician held the B.S. in Education with a major in Speech Pathology and Audiology, 18 semester hours graduate work toward the M.S. in Speech Pathology, and had three years of professional experience in public schools.

Strandberg (1969) recorded the original speech samples which were used as stimulus material in this study. She recorded a minimum of one-minute speech samples of each identified first grader. Her collection of continuous speech samples was similar to the technique used by Morrison (1955). Each child spoke about his favorite T.V. program, an activity during the summer which he thought was most fun, and what he liked most about school. Each subject's verbal output was recorded in the speech therapy room of his respective school.

Samples were recorded on an Ampex, Model 602 tape recorder at a tape speed of seven and one-half inches per second. To obtain optimum fidelity, she used Scotch Magnetic Tape, silicon lubricated 1.5 mil acetate backing. The child was seated so that the distance from his mouth to the microphone could be controlled at six inches.

Strandberg had collected the speech samples as soon as possible after identification by clinicians "to eliminate possible intrusion of extraneous factors which might have influenced and changed the voice quality heard by the public school speech correctionist. . ."

Since retrieval of stimuli from original recordings should be done with consistent methodology, the experimenter chose the

first ten-seconds of verbalization of each subject from the original tapes prepared by Strandberg. In some instances continuous ten-second responses could be recorded. However when a subject responded only in one- or two-word utterances, these responses were recorded until ten-seconds of stimuli had been obtained. Lewis and Sherman (1951) had presented varying lengths of samples of stuttering for judging. They had essentially concluded that six-second samples were too short, 15-second samples were "unnecessarily prolonged," but ten-second samples were of optimum length. The Morrison (1955) study compared length of stimuli for rating articulation defectiveness. This study reported that both five- and ten-second speech segments sould be used as reliably as one-minute speech samples. On On the basis of reliability and experiment time, ten-second length speech samples were chosen for the observer rating procedures.

Forty two stimulus segments were selected from the original 44 samples. Two samples were excluded from this study because they had been judged to be normal by at least 80% of a panel of trained speech pathologists in the Strandberg (1969) study.

Preparation of EAI stimulus tape.

The original tapes were played on an Ampex, Model 602 tape recorder and the first ten-second segments were internally dubbed onto silicon lubricated 1.5 mil Scotch Magnetic Tape through a Revox, Model 36-G tape recorder. When recording at seven and one-half inches per second, the Revox displays a

frequency range of 40-18,000 Hz., intensity variation of +2/-3 dB, and tape speed deviation of no more than .03 from seven and one-half inches per second. The experimental segments were dubbed through Channel I and were monitored acoustically by the experimenter using Telex MR-6 earphones. The input to Channel I was monitored visually by the experimenter using the Channel I V.U. meter and attenuator.

The experimenter announced and recorded each respective stimulus number through a Shure microphone into Channel II of the Revox recorder. These stimulus numbers were recorded as closely as possible to the input level as Channel I. Channel II input was likewise monitored acoustically and visually by the respective V.U. meter and attenuator. These assigned stimulus numbers served not only to assist the observer to follow respective items on the response sheet, but also to increase observer attention in preparing to listen for the upcoming stimulus. A five-second inter-stimulus interval was used to allow time for observer judging and recording.

A twenty-second pure tone of 1000 Hz., recorded at the same average input level as Channels I and II, was inserted at the beginning of the completed tape. This tone was introduced by holding an earphone of a Beltone Audiometer 10-C, with the attenuator set at 75 dB, to the microphone connected to the Channel I input. The purpose of this test tone was to enable the experimenter to control the intensity of the stimuli output in various experimental environments. The intensity

range of the experimental tape with the output control held at a constant #3 setting on the Revox was 65 to 85 dB with the average intensity being 75 dB. A Sound Level Meter, General Radio Type 1551-C was used to determine the above output levels.

Preparation of DME stimulus tape.

The second tape, prepared for judgment by direct magnitude-estimation, was constructed in the following manner. One segment from the EAI tape, (Tape I), was extracted to become the standard stimulus for the DME tape, (Tape II). The criteria for selecting the standard stimulus for Tape II was that this stimulus previously must have been judged to represent a mid-point of all samples, and that the segment must be of acceptable acoustical quality and length to be judged.

Four trained speech pathologists rated all 42 stimuli on Tape I. Two of the observers held the Ph.D. and had an average of 15 years clinical experience, another held the M.S. in Speech Pathology with eight years of clinical experience, and the latter held the B.S. in Education with a major in Speech Pathology and three years of clinical experience. The four observers rated Tape I by the method of equal-appearing intervals. The stimuli was presented through the Revox recorder, free-field in a sound treated room. Since the test tone represented the average intensity range for the entire tape, the Sound Level Meter 1551-C was employed to set the test tone level at 65 dB. This setting allowed the tape to be presented at the intensity range of 55 to 65 dB.

Seventy-five percent of the judges agreed that segment #30 of Tape I represented the fourth category, or mid-point of the range of voice qualities presented. Judgmental reliability for this judging group was .84 as computed by the intra-class correlation coefficient formula. (Winer, 1962, p. 198) The experimenter and another member of the judging group agreed that segment #30 met the previously described criteria of acceptance.

Tape II was then dubbed from Tape I using the same technical procedures as were used to prepare Tape I. The standard stimulus was dubbed into the beginning of the tape and after every subsequent fifth speech segment. Because #30 was omitted from Tape II, new stimulus numbers were assigned to segments following number 29. The words "standard stimulus", which were inserted preceeding each standard segment, and the revised segment numbers were inserted through Channel I of the Revox recorder. The completed tape to be used for judgment by direct magnitude-estimation contained 41 segments and nine presentations of the standard stimulus.

Selection of scaling methods.

The previous Chapter has offered theoretical considerations for selection of the three scaling methodologies. The following summary statements are made about each methodology.

Successive intervals scaling assumes that judges are not able to divide a continuum into equal-size segments. Scale values are ordinal and do not assume to satisfy the criteria

for interval-level measurement. The successive intervals method may be applied to equal-appearing intervals data.

Equal-appearing intervals scaling was selected because of its common use in speech pathology. This scaling method assumes that judges are able to divide a continuum into equal-sized segments. If judges perform the given task as instructed, their judgments should result in scale values which satisfy the criteria for interval level measurement. (Sherman and Silverman, 1968)

Theoretically, direct magnitude-estimation should result in scale values which satisfy the criteria for ratio-level measurement. (Sherman and Silverman, 1968). Scale values should be located in reference to a true zero and thus could be used meaningfully in all arithmetical operations.

Selection of judges.

The experimenter chose to use untrained listeners for this study. Trained listeners form only a small sample from a total population of listeners. Judgments of defective speech primarily come from cultural standards of a society of untrained listeners. Since voice quality is a perceptual event, judgments as to voice quality are subjective and the speech pathologist must rely upon an untrained listener population to quantify judgments as to the severity of voice quality. Siegel (1962) has listed two reasons why it is desirable to use relatively inexperienced persons as articulation examiners. These reasons also appear to be applicable to investigations of voice quality. The

first reason is practicability. An experimenter may not always have experienced examiners available. The second reason is that "Ignorance of the areas of speech pathology and language development may constitute an experimental safeguard against particular biases or expectations." Siegel's second reason should also apply to voice studies from the standpoint of reducing some of the extraneous variables of articulation and language which plague the trained voice judge.

The untrained listener population for this study was selected from speech, psychology, and health education classes at Eastern Illinois University. All of the classes were Freshman level courses except for one psychology class which was at the Sophomore level.

Students selected as judges for this study were checked for hearing acuity. This process was accomplished by checking each judge's Speech and Hearing Screening record at the Department of Speech Pathology and Audiology. One subject was eliminated from this investigation because he had not passed the hearing screening.

The traditional approach for selecting the number of judges for a study arbitrarily predetermines the number of judges to be used. Investigators then compute the reliability of obtained scale values, plot scatter-grams of each method against the other, then finally determine the correlation between scale values. Interpretation of results of this method are unclear. One cannot know whether obtained differences lie

in judgmental reliability or to differences in the scaling methodologies. That is, one cannot conclude from the traditional method whether obtained reliability differences may have resulted from internalized observer reactions to given stimuli or to the scaling method function as a yardstick to measure the range of attitudes along the perceptual continuum.

This investigation employed the principle of sequential sampling as described by Silverman (1968). In this procedure the experimenter sets a minimum level of reliability desired for scale values. He would have a small number of observers rate the stimuli. Next he would estimate the reliability of scale values which could be derived from ratings of these observers. If the level of reliability attained was greater than or equal to the desired level, no observers would be added. However, if the attained level of reliability was less than the desired level, the experimenter would then have additional observers, selected from the same population of observers, rate the stimuli. This described process is replicated until the desired reliability level is attained. With this procedure, obtained differences may be explained as due to methodological variations. The minimum level of reliability for observers scaling by EAI and DME in this study was set at .95. The reliability level was set at .95 for the following reasons. (1) Previous voice quality scaling studies (Sherman and Linke, 1952; Sherman, 1954; Spriestersbach, 1955; Rees, 1958; Spriestersbach and Powers, 1959; Lintz and Sherman, 1961; and

Dickson, 1962) were unsuccessful obtaining judgment reliability over .90 using EAI and the traditional research approach of selecting the number of observers to perform the given scaling task. (2) The second reason was to investigate whether Silverman's (1968) principle of sequential sampling could be applied successfully to reach a high reliability with voice quality scaling methodologies. (3) An alpha level of .05 would indicate that the chances of obtaining similar high judgmental reliability in replicating this study would be .95.

Presentation of stimuli.

The stimuli for rating by equal-appearing intervals and direct magnitude-estimation were presented in the student's respective classroom. Each class contained a maximum of 30 students. The small class grouping allowed the investigator to supervise the experimental session closely. The stimuli were presented on the same experimental schedule for both the EAI and DME judging groups as follows. (1) Before the experimental session the investigator set up the equipment so that the sound source was in front and center of the classroom. (2) The Sound Level Meter, Type 1551-C was used to check the test tone of 75 dB measured from the front row of the classroom. (3) The instruction booklet and response sheet were distributed when class members had been seated. (4) The instructions were read aloud by the experimenter. (5) The tape was presented for judges to listen. The first playing was intended to give them an idea of the task and to give them

the opportunity to perceive the range of stimuli so that they could form their own end-points of the continuum. (6) The remainder of the instructions were read and any questions regarding judging procedure were explained. (7) The tape was played the second time for purposes of marking judgments to stimuli. (8) Judges were asked to give name, class standing, and age on the front of the response booklet. (9) Response booklets were collected. (10) Question and answer session. The entire session averaged 32 minutes.

A copy of the directions and response booklet for both equal-appearing intervals and direct magnitude-estimation may be found in Appendix A and B respectively.

Chapter IV

RESULTS AND DISCUSSION

Scale Values.

The reliability of the scale values obtained by direct magnitude-estimation were assessed by the intraclass correlation coefficient for averages (Winer, 1962, p. 128). The resulting r was 0.93, based upon the judgments of 80 observers. This correlation was interpreted to mean that the stimuli tended to rank order themselves in a similar manner. The scale values, which represent a mean of observer responses for each presented stimulus, range from 82.19 to 149.01 with a mean of 118.11 and a standard deviation of 21.55. The sequential sampling procedure (Silverman, 1968), which determines the number of additional observers from the same population needed to reach the desired reliability was used. However, the obtained reliability level fell slightly short of the pre-determined level of 0.95. It seems reasonable to assume that there would be little difference, if any, in the rank ordering of the stimuli between the obtained reliability level of 0.93 and the desired level of 0.95. A shortage of available observer population hindered addition of observers to attempt to reach this desired level. Although this observer population

was considered to be alike in that they were university students naive or untrained in rating voice quality, some discussion must be given to possible differences within this population.

This total observer population was divided into sub-populations by class, academic course, age, and sex. Table I illustrates inter-group reliability levels for scale values obtained from the 80 observers who rated the voice quality stimuli by direct magnitude-estimation.

Table I. Intraclass correlations obtained for sub-populations by class, academic course, age, and sex rating by DME.

Academic class	N	r	Age	N	r
Freshman	49	.90	Age 17	20	.80
Sophomore	20	.78	Age 18	26	.75
Junior	10	.44	Age 19	16	.59
			Age 20	9	.73
Course			Sex		
Psychology	28	.78	Male	39	.90
Speech	22	.84	Female	44	.85
Health Education	30	.85			

Differences in the magnitude of correlations between groups (Blommers and Linquist, 1960, p. 465), were computed within the academic class and sex sub-populations. There were no significant differences between any of the obtained correlations for these sub-populations comprising the total observer population rating by direct magnitude-estimation. Other comparisons within the age and academic class categories were not made because of the differing sub-sample population size. Since differences between correlations are a function of sample size, and the N

within these categories varied considerably, any significant results, or lack of them, would be impossible to interpret. Since the principle of sequential sampling (Silverman, 1968) is based on the assumption that additional observers are drawn from the same population, it can be concluded that for DME scaling of voice quality academic class and sex are not relevant variables in the selection of additional observers.

The reliability of the equal-appearing intervals scale values for the 42 stimuli was computed by the intraclass coefficient for averages (Winer, 1962, p. 128). A reliability level of 0.99 was obtained with a population of 143 observers. The 42 EAI scale values range from 2.19 to 6.55 with a mean of 5.07 and a standard deviation of 1.27.

The sequential sampling procedure (Silverman, 1968) was again applied successfully to reach the pre-established reliability level of 0.95. Since increased reliability is a function of increased numbers of observers from the same population, fewer observers could have been used for rating the voice quality stimuli by EAI. Hand computation errors in sequential sampling account for over-estimation of additional observers. Because the EAI scaling task was performed prior to DME, this over-estimation contributed to the shortage of available population needed to establish desired reliability for DME.

Sub-populations divided by academic class, course, age, and sex composed the total untrained observer population. Table II illustrates the obtained reliability levels for each

sub-population rating by EAI.

Table II. Intraclass correlations obtained for sub-populations by academic class, course, age and sex rating by EAI.

Academic class	N	r	Age	N	r
Freshman	68	.97	Age 17	12	.91
Sophomore	36	.94	Age 18	37	.95
Junior	19	.88	Age 19	31	.95
Senior	20	.89	Age 20	19	.90
			Age 21	18	.88
			Age 22	8	.77
Course	**N**	**r**	**Sex**	**N**	**r**
Psychology	78	.98	Male	68	.97
Speech	65	.97	Female	75	.97

Differences on magnitude of correlations between sex and academic class (Blommers and Linquist, 1960, p. 465) indicated no statistically significant differences between obtained correlations for these sub-populations. This was not an unexpected finding in view of the extremely high overall reliability level.

Since successive intervals scale values are computed from scale values derived by equal-appearing intervals methodology, it seems reasonable to assume that the reliability of these scale values is of a comparable magnitude as the EAI scaling procedure. The precedent for this assumption is found in the Sherman-Silverman 1968 study. The range of scale values, computed from a table of cumulative proportions based on responses obtained by EAI, was from 0.8 to 3.3, with a mean of 1.61 and a standard deviation of 0.64.

For internal consistency evaluation, cumulative theoretical

proportions of judgments in the 7 intervals for each of the 42 voice quality stimuli were computed and compared with the corresponding observed cumulative proportions. The agreement between the observed and theoretical proportions is close. Only 65 of the 252 theoretical proportions deviate from the observed proportions by more than 0.05. Although the mean deviation is small, 0.26, it is slightly larger than the typical average error reported by others (Edwards, 1957, p. 138). However, a deviation of this magnitude still is a reliable consistency within scale values for successive intervals.

Comparison of scaling methods.

The two sets of scale values derived from the same data by EAI and SI methodologies indicate that both methodologies rank order the 42 voice quality samples in an identical manner. The correlation between the two sets of scale values was 0.99. This correlation is the same magnitude as has been reported for correlations between equal-appearing intervals median scale values and successive interval scale values for other types of stimuli (Silverman and Sherman, 1967 and Sherman and Silverman, 1968). Essentially there is no difference between obtained scale values for the two methods. Because of simpler computational procedures and less computational time, EAI is the preferred scaling method of choice.

The correlation of 0.93 between direct magnitude-estimation mean scale values and equal-appearing intervals scale values is high. In fact, this correlation should be considered especially

high since the two sets of scale values are derived from two different groups of observers rating by different methodologies.

The null hypothesis posed for this investigation was: There are no significant differences among reliability of measures of data gathered in judgments of voice quality problems by equal-appearing intervals, successive intervals, and direct magnitude-estimation. The null hypothesis was confirmed; that is, high and comparable reliability levels were obtained by each of the scaling methods. Moreover, each scaling method yielded a similar rank ordering of the stimuli.

A second question raised at the outset of this investigation was: Can naive, or untrained listeners reliably judge the severity of samples of voice quality deviations? Previous research in articulation (Perrin, 1952; and Morrison, 1955) and in stuttering (Young, 1961) report little or no significant reliability differences between trained and untrained observers. The high correlation (0.93) between EAI and DME suggests that naive, or untrained observers also can be used to reliably rate severity of voice quality stimuli.

Scale values obtained by the three methodologies for the kind of stimuli used in this study appear to differ very little in their usefulness. All three scaling methods, EAI, SI, and DME, tend to rank order the stimuli in a comparable manner. The results of this investigation are compatible with other published research in the speech pathology literature. Since EAI is a practical and reliable measurement procedure and is

the easiest of the three methods to compute, it remains the preferred scaling method.

Implications for future research.

The first step in quantifying the perceptual impact of voice quality deviations upon observers is to select a reliable and practical measurement tool. The results of this investigation suggest that the psychological scaling method of equal-appearing intervals satisfies these criteria. A logical extension of the present research would be to construct a master tape for the purpose of training speech pathologists in making voice quality judgments. Since reliable scale values were obtained, those stimuli having approximately integer values and small Qs could be employed to prepare a severity training tape of voice quality comparable to the Lewis and Sherman scale of stuttering severity. Such a tape would aid the speech pathologist in quantifying voice qualities.

Chapter V

SUMMARY

The primary purpose of this investigation was to evaluate the null hypothesis that there were no significant differences among the reliability of measures of data gathered in judgments of voice quality problems by equal-appearing intervals, successive intervals, and direct magnitude-estimation. Two other questions posed in this study were: (1) Can naive, or untrained listeners reliably judge the severity of samples of voice quality deviations? and (2) If scaling methods can be used to rate severity of voice quality deviations, which method, EAI, SI, or DME, will be most reliable and practical for evaluative purposes? When attempting to quantify the perceptual impact of voice quality upon listeners, the methodological question arises, which scaling method should be employed? This procedural problem must be resolved before one could train observers or construct a master training tape of voice quality deviations.

Equal-appearing intervals has been described by Sherman and Moodie (1957) as a scaling methodology in which "the observer is instructed to assign numbers to the stimuli in relation to an equal-appearing scale of severity." The principle assumption underlying EAI is that the observer can successfully

equate intervals or distances between responses to stimuli. EAI yields interval level of measurement data.

Successive intervals scaling essentially places each of several stimuli into a limited number of categories differing quantatively along a given continuum. No assumption is made that scale values are equi-distant. However, it does assume that "categories are in correct rank order and that their boundary lines are stable except for sampling errors." (Guilford, 1952, p. 34). Successive intervals scaling yields ordinal level of measurement data.

In DME scaling, observers assign scale values in relation to a standard stimulus sample, of a pre-assigned value. Scale values are representative proportions of judgments made in reference to an absolute zero. Derived scale values represent ratio level of measurement.

The stimuli employed in this study were obtained from the Strandberg study (1969). Strandberg had collected the original voice quality samples by recording one-minute speech samples elicited in response to questions regarding a favorite T.V. program, a most enjoyable summer activity, or most enjoyable part of school. These samples were recorded by an Ampex, Model 602 tape recorder at a tape speed of seven and one-half inches per second. From these samples, two stimulus tapes were prepared.

The original tapes were played on an Ampex, Model 602 recorder and the first ten-seconds were internally dubbed onto the EAI tape through a Revox, Model 36-G tape recorder.

Stimulus numbers were recorded on the tape preceeding each respective stimulus. A five-second inter-stimulus interval was used to allow time for observer judging and recording. Forty-two stimuli comprised the EAI judging tape. The second tape for scaling by DME was prepared in like manner to the EAI tape except for the inclusion of a standard stimulus.

The untrained observer population for this study was selected from speech, psychology, and health education classes at Eastern Illinois University. All classes were freshman level courses except for one sophomore level course. All observers passed a sweep check hearing screening test at the university's Speech and Hearing Clinic.

Both EAI and DME stimulus tapes were presented in the student's respective classroom. Each observer heard his stimulus tape twice. The first presentation proposed to allow each observer to listen only and to formulate his own anchor points as to the least and most severe voice quality perceived on that tape. The actual task was performed during the second stimuli presentation.

The reliability of the scale values obtained by DME, assessed by the intraclass correlation coefficient for averages, yielded an r of 0.93 for 80 observers. Although the obtained reliability level fell slightly short of the pre-determined level of 0.95, it seems reasonable to assume that there would be little or no difference in the rank ordering of stimuli. The total observer population was divided by academic class, course,

age, and sex. There were no significant differences between any of the obtained correlations for these sub-populations.

The reliability of EAI scale values, also computed by the intraclass coefficient for averages yielded a correlation of 0.99 based upon 143 observers. Differences on correlations between sub-populations, also divided by academic class, course, age, and sex, indicated no statistically significant differences between obtained correlations for the sub-populations.

Successive intervals were computed from scale values derived from EAI methodology. A check for internal consistency found the mean deviation of 0.26 to be slightly larger than the typical average error reported by previous investigators. However, this slight deviation still indicates a reliable internal consistency within scale values for SI.

The null hypothesis posed for this investigation was confirmed. That is, high and comparable reliability levels were obtained by each of the three scaling methods. The high correlations between EAI and DME suggest that naive, or untrained observers can reliably rate severity of voice quality stimuli. All three scaling methods tend to rank order the stimuli in a comparable manner. Since EAI is a practical and reliable measurement procedure, it remains the preferred scaling method for rating voice quality severity.

Appendix A

INSTRUCTIONS TO JUDGES
FOR EAI SCALING

You are asked to judge a series of children's voices which are presented to you in 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 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; likewise, some voices are more unpleasant 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 only 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 unpleasant 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 1 representing the least unpleasant and 7 the most unpleasant voice you will hear on this tape. I shall play the samples first; do not record the samples--merely listen.

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

Following there will be 42 voices to be rated on the 7-point scale. These voice samples were obtained by asking first grade children questions about their favorite T.V. program, activities during the summer that they thought were most fun, and what they liked best about school. All responses are to the same set of questions.

Before you record any judgments, you will listen to the 42 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 degree of unpleasantness. Just listen, form a concept of the least and most unpleasant voices on 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 answer sheet. Remember, 1 represents least unpleasant and 7 represents 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?

Observer No._____

ANSWER SHEET

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._____
33._____
34._____
35._____
36._____
37._____
38._____
39._____
40._____
41._____
42._____

Appendix B

INSTRUCTIONS TO JUDGES FOR DME SCALING

You are asked to judge a series of children's voices which are presented to you in tape recorded form. You are asked to judge each voice sample in relation to a standard sample of "unpleasantness." Unpleasantness, for purposes of this experiment, is interpreted to mean that the quality is bad enough to call 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; likewise, some voices are more unpleasant 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 only to how each sounds in terms of his voice quality; that is, how unpleasant does each child's voice sound to you?

You are asked to estimate the relative degree of "unpleasantness" of each voice quality segment in relation to a standard

segment which will be played for you soon. You will do this task by assigning the number of points you believe represents the relative degree of unpleasantness for each segment in relation to the standard segment. Now you shall hear what we call the standard segment. (Play it once) You will assign 100 points to this segment. The point assignments you will be asked to make on the succeeding segments should represent the relative degree of unpleasantness of each child's voice quality exhibited in each segment. For example, if you believe that the unpleasantness of the second segment exhibits twice the degree of unpleasantness as the voice quality in the standard segment, you will assign 200 points to the second segment. If you believe that the degree of unpleasantness exhibited in the segment is half that exhibited in the standard segment, you would assign 50 points. Of course, you may use any point assignment you choose to represent the degree of unpleasantness; you need not limit yourself to even fractions and multiples of the 100 points assigned to the standard. You might use the quantity of 85 or 65 or 20 or even 112, or 120 or 215 or any number you choose so long as it represents the degree of unpleasantness exhibited in relation to that exhibited in the standard segment.

Now you will hear the standard segment followed by those segments which you will soon be judging. Do not record judgments--merely listen. You might think about the point assignments you would make if you were recording judgments. Occasionally you will hear some background noise on the tape. Totally disregard

this and form your impressions soley on the basis of each child's voice. (Play tape--just listen)

You are now ready to judge the experimental segments. The first segment is your standard segment. When it is played, listen very carefully and note the 100 assigned on your answer sheet. With the remainder of the segments, you must record the number which represents the degree of unpleasantness exhibited in the segment in relation to the 100 points assigned to the standard segment. The standard segment of 100 will be played after every five judgments that you make. If you are somewhat doubtful about what number to assign, make a guess. You will record your number to the right of the segment number on your answer sheet. (Each segment will be announced by its respective number.) After listening to each segment, you will record the number of points which you think the segment would have in relation to the standard segment of 100 points.

Are there any questions?

Observer No. _____

ANSWER SHEET

(Standard segment 100)

1. _____
2. _____
3. _____
4. _____
5. _____

(Standard segment 100)

6. _____
7. _____
8. _____
9. _____
10. _____

(Standard segment 100)

11. _____
12. _____
13. _____
14. _____
15. _____

(Standard segment 100)

16. _____
17. _____
18. _____
19. _____
20. _____

(Standard segment 100)

21. _____
22. _____
23. _____
24. _____
25. _____

(Standard segment 100)

26. _____
27. _____
28. _____
29. _____
30. _____

(Standard segment 100)

31. _____
32. _____
33. _____
34. _____
35. _____

(Standard segment 100)

36. _____
37. _____
38. _____
39. _____
40. _____

(Standard segment 100

41. _____

BIBLIOGRAPHY

Articles

Barker, Janet O'Neill, "A Numerical Measure of Articulation," Journal of Speech and Hearing Disorders. 25, 1960, 79-88.

Bloodstein, O. N., "The Relationship Between Oral Reading Rate and Severity of Stuttering," Journal of Speech and Hearing Disorders. 9, 1944, 161-173.

Cullinan, W. L., Prather, Elizabeth M., and Williams, Dean. "Comparison of Procedures for Scaling Severity of Stuttering," Journal of Speech and Hearing Research. 6, 1963, 187-194.

Curry, R., et. al., "A Phonographic Scale for Measurement of Defective Articulation," Journal of Speech and Hearing Disorders. 8, 1943, 123-126.

Dickson, David, "An Acoustic Study of Nasality," Journal of Speech and Hearing Research. 5, 1962, 103-111.

Henrikson, E. H., "An Analysis of Wood's Articulation Index," Journal of Speech and Hearing Disorders. 13, 1948, 233-235.

Johnson, Wendell, "Measurement of Oral Reading and Speaking Rate and Disfluency of Adult Male and Female Stutterers and Nonstutterers," Journal of Speech and Hearing Disorders. 1961 Monograph Supplement 7, 1-20.

Jordan Evan P., "Articulation Test Measures and Listener Ratings of Articulation Defectiveness," Journal of Speech and Hearing Research. 3, 1960, 303-319.

Lewis, D. and Dorothy Sherman, "Measuring the Severity of Stuttering," Journal of Speech and Hearing Disorders. 16, 1951, 320-326.

Lintz, Lois B. and Dorothy Sherman, "Phonetic Elements and Perception of Nasality," Journal of Speech and Hearing Research. 4, 1961, 381-396.

Menyuk, Payla, "Syntactic Structures in the Language of Children," Child Development. 34, 1963, 407-422.

Morrison, Sheila, "Measuring Severity of Articulation Defectiveness," Journal of Speech and Hearing Disorders. 20, 1955, 347-351.

Perrin, Elinor H., "The Rating of Defective Speech by Trained and Untrained Observers," Journal of Speech and Hearing Disorders. 19, 1954, 48-51.

Prather, Elizabeth M., "Scaling Defectiveness of Articulation by Direct Magnitude-Estimation," Journal of Speech and Hearing Research. 3, 1960, 380-392.

Rees, Maryjane, "Some Variables Affecting Perceived Harshness," Journal of Speech and Hearing Research. 1, 1958, 155-168.

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, 312-321.

_____ and Eugene Linke, "The Influence of Certain Vowel Types of Degree of Harsh Voice Quality," Journal of Speech and Hearing Disorders. 17, 1952, 401-408.

_____ and Sheila Morrison, "Reliability of Individual Ratings of Severity of Defective Articulation," Journal of Speech and Hearing Disorders. 20, 1955, 352-356.

_____ and William D. Trotter, "Correlation Between Two Measures of the Severity of Stuttering," Journal of Speech and Hearing Disorders. 21, 1956, 426-429.

_____ and Catherine E. Moodie, "Four Psychological Scaling Methods Applied to Articulation Defectiveness," Journal of Speech and Hearing Disorders. 22, 1957, 698-706.

_____ and W. L. Cullinan, "Several Procedures for Scaling Articulation," Journal of Speech and Hearing Research. 3, 1960, 191-198.

Sherman Dorothy and Franklin H. Silverman, "Three Psychological Scaling Methods Applied to Language Development," Journal of Speech and Hearing Research. 11, 1968, 837-841.

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 and Motor Skills. 27, 1968, 1333-1334.

_____ and Dorothy Sherman, "Equal-Appearing Interval Scale Values and Successive Interval Scale Values Derived from the Same Set of Ratings," Perceptual and Motor Skills. 25, 1967, 226-228.

Shriner, Thomas, "A Comparison of Selected Measures With Psychological Scale Values of L nguage Development," Journal of Speech and Hearing Research. 10, 1967, 828-835.

_____ and Dorothy Sherman, "An Equation for Assessing Language Development," Journal of Speech and Hearing Research. 10, 1967, 41-48.

Spriestersbach, D. C., "Assessing Nasal Quality in Cleft Palate Speech of Children," Journal of Speech and Hearing Disorders. 20, 1955, 266-270.

_____ and G. R. Powers, "Nasality in Isolated Vowels and Connected Speech of Cleft Palate Speakers," Journal of Speech and Hearing Research. 2, 1959, 40-45.

Stevens, S. S., "The Direct Estimation of Sensory Magnitude-Loudness," American Journal of Psychology. 69, 1956, 1-25.

Strandberg, Twila E., "A Preliminary Evaluation of a Screening Method to Identify Harsh Voice Qualities in First Grade Public School Children," Unpublished graduate study. 1969.

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

Young, Martin A., "Observer Agreement: Cumulative Effects of Rating Many Samples," Journal of Speech and Hearing Research. 12, 1969, 135-143.

_____. "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, 144-155.

Books

Blommers, Paul and E. F. Lindquist. Elementary Statistical Methods in Psychology and Education. Boston: Houghton-Miffin Company, 1960.

Edwards, Allen L. Techniques of Attitude Scale Construction. New York: Appleton-Century-Crofts, Inc., 1957.

Guilford, J. P. Psychometric Methods (2nd Ed.). New York: McGraw-Hill, 1954.

Johnson, Wendell; Darley, Frederic L., and Spriestersbach, D. C. Diagnostic Methods in Speech Pathology. New York: Harper and Row Publishers, 1952.

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

Winer, B. Statistical Principles in Experimental Design. New York: McGraw-Hill, 1962.