This is to certify that the

thesis entitled

CLINICAL JUDGMENTS MADE BY SPEECH PATHOLOGISTS AND STUDENTS UNDER VARYING INFORMATION CONDITIONS

presented by

Michael J. Flahive

has been accepted towards fulfillment
of the requirements for

Ph.D. degree in Audiology
and Speech Sciences

Major professor

Date 9 November 1979

O-7639

CLINICAL JUDGMENTS MADE BY SPEECH PATHOLOGISTS AND STUDENTS UNDER VARYING INFORMATION CONDITIONS

by

Michael J. Flahive

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

Department of Audiology and Speech Sciences

1979

ABSTRACT

CLINICAL JUDGMENTS MADE BY SPEECH PATHOLOGISTS AND STUDENTS UNDER VARYING INFORMATION CONDITIONS

By

Michael J. Flahive

Investigations of examiner bias among graduate and undergraduate students training in speech pathology have been equivocal. The question of potential bias among working speech pathologists has not been examined. Therefore a study was designed to explore the issue of bias across several populations. Seventy-five subjects -- twenty-five undergraduate students, twenty-five graduate students and twenty-five professional speech pathologists -- participated in the present study. Subjects provided scaled ratings of speech samples of eighteen speakers in a repeated measures format with experimental conditions varying as a function of case history information, negative case history information and neutral information. Ratings were made on a seven-point equal-appearing interval scale. Judgments included an initial normal/non-normal determination followed by scaling the degree of severity of the problem if it were determined one existed. An additional judgment relative to the disorder was to determine whether the primary speech production problem was one of articulation or of voice. Speech samples of the same eighteen speakers were then re-ordered and presented a second time following a 10-12 minute distraction

time. The second presentation included fabricated case history statements.

Results indicated consistency in mean group assignment of speech samples to categories across presentations, although a greater number of "problem" samples were identified than actually existed. Measurement of category assignment resulted in high levels of agreement. Accuracy of categorical assignment varied as a function of training and experience with graduate students functioning most accurately and undergraduate students least accurately. Voice problem samples were the more frequent error selection with working professionals demonstrating greatest difficulty in identifying problems of this type on both presentations. Of the proportion of subjects accurately identifying the appropriate category on both presentations, sixty-one percent altered ratings of severity across presentations as a result of case history influence. At the same time it did not appear that the case history type consistently caused judgments to be altered in the suggested direction of the statements. Evaluation of severity rating behavior on the reliability samples of Presentation I indicated poor intra-subject agreement. Ratings of severity for all experimental subjects were consistent, but varied considerably from the values assigned by expert judges.

Generally it appeared that experimental subjects were influenced by case history information and presumably by

the demand characteristics of the experimental task. Results are discussed in light of previous research, and implications are stated for the training of students and working professionals regarding background information and its function in the evaluative process.

TABLE OF CONTENTS

TABLE OF CONTENTS

LIST OF TABLES

LIST OF TABLES

LIST OF FIGURES

LIST OF FIGURES

INTRODUCTION

In 1897, while discussing scientific thinking, T. C. Chamberlin wrote: "If our vision is narrowed by a preconceived theory as to what will happen, we are almost certain to misinterpret the facts and misjudge the issue." During the years following these remarks, the disciplines of psychology and education have been concerned with the issue of a preconceived theory or notion, particularly in the process of evaluation. The term "bias" has been used to describe this predisposition. Plutchik (1974) describes bias as "any fact or factor which contributes to an erroneous conclusion or which makes the conclusion ambiguous."

Experimenters in psychology have conducted a host of studies to determine the effect of various personal attributes upon judgments that are made about individuals. These attributes range from physical appearance to socioeconomic and intellectual status to cultural and ethnic background. An understanding of how this descriptive information influences objectivity in the diagnostic and appraisal process is critical to identifying sources of potential bias. Research efforts in this direction have traditionally used the terms "experimenter bias" or "examiner bias" to describe errors which consistently vary from a true value and which relate

to characteristics of the observer situation (Rosenthal, 1968). Friedman, Kirkland and Rosenthal (1965) differentiate experimenter bias and experimenter effect as follows:

experimenter bias - occurs when the experimenter obtains results from the subject that he expects to obtain.

experimenter effect - occurs when different experimenters obtain different data from the same subject.

As early as 1907, Wells noted the psychological perceptual error known as the "halo effect" which refers to a rating based upon overall impressions of goodness or badness. Social behaviors such as cultural or ethnic stereotyping are examples of the halo effect.

A number of studies have been performed in education to investigate examiner bias. Rosenthal and Jacobson's work in the mid- and late 1960's represents some of the best known and most controversial.

Rosenthal, a social psychologist, reported a series of experiments involving classes designated as "fast," "medium," or "slow" in reading at each grade level from first through sixth in a single elementary school in San Francisco. He administered a test described as a device which would identify "bloomers" among the population after which he told the teachers that a number of children would probably experience an unusual forward spurt in academic and intellectual performance during the school year. The true case, however, was that roughly 20% of the children had been randomly assigned to this condition. Results showed significant differences

favoring children who had been labeled as "bloomers," prompting Rosenthal and Jacobson to conclude "...that teachers' favorable expectations can be responsible for gains in their pupils' I.Q.'s and for the lower grades these can be quite dramatic" (1968, p. 98). The phenomenon of teacher expectancy was labeled by Rosenthal as the "Pygmallion effect," and it received wide attention in academic and social media during the late 1960's.

Barber and Silver (1968) critically analyzed 31 studies which attempted to demonstrate the examiner bias effect. One conclusion they reached was that Rosenthal and other proponents of the term had overstated the issue and that examiner bias was less pervasive and more difficult to demonstrate than had been suggested. They further indicated that subsequent studies of the potential effect should take care to address several methodological issues they found remiss in many papers they reviewed. These include failure to determine the reliability of the criterion instrument, failure to check for the effectiveness of the independent variable manipulation and failure to use control groups. Barber and Silver also raised the issue of means for inducing bias and the need to clarify the role played by these various sources.

In Pygmallion Revisited (1971), Elashoff and Snow summarized nine major attempts to replicate Rosenthal and Jacobson's work (including two in which Rosenthal himself was a co-author) and sixteen related studies. None of these studies was fully able to replicate the original findings. Elashoff and Snow

severely critized Rosenthal and Jacobson's methods, statistical analysis, conclusions and generalizations. Based upon what they believed to be overwhelming evidence, they concluded that

1. teacher expectancy probably does not affect pupil I.Q.
2. teacher expectancy probably affects pupil achievement
3. teacher expectancy probably affects observable teacher and pupil behavior, if the expectancy condition occurs naturally or provides a moderate-to-strong manipulation of inducement (pp. 61-62).

The debate between Rosenthal and those in opposing camps underscores the professional concern regarding objectivity in measurement as well as treatment. An addition confounding element in the study of these questions is what is known as the "Hawthorne effect." This phenomenon suggests that experimental changes can be observed that are not a function of any independent variable but are instead due to the attention received by the subject. It is essentially an observer-subject interaction effect (Roethlishberger and Dickson, 1939). In evaluating reports of any "treatment" effect of difference between methods, it is important to bear this psychological phenomenon in mind.

The evaluation process, whether in psychology, education or speech pathology, involves measurement of some category of an individual's performance and the judgment of that performance along some reference dimension by the examiner.

Johnson, Darley and Sprietsersbach (1963) discuss a philosophy of diagnosis and appraisal in speech pathology wherein they state that the clinician should observe impartially, precisely and reliably.

> He should observe enough, and he should do it by techniques that will permit his information to be compared satisfactorily with that obtained by other observers. It is important that he distinguish between what he observes and what he concludes from his observations. He must distinguish, in other words, between fact and inference. It is not easy to make this kind of distinction and to communicate the results of observation with appropriate objectivity. Without our even recognizing what is happening, our own interests, personal biases, and convictions distort our perceptions, our conclusions and our reports (pp. 3-4).

They further specify that the clinician's constant goal should be to preform as objectively as possible and to acknowledge and reduce the distorting elements in observation and reports.

Few studies have been completed in speech pathology relating to objectivity in the evaluation process or to examiner bias. Beasley and Manning (1973) reported an experiment conducted with graduate students in which several levels of case history information were given. This information was categorized as negative, positive and incomplete, or none. The evaluators then measured language samples on several objective and subjective scale measures. Objective measures included mean length of response, the five longest responses and a type-token ratio. Subjective measures consisted of four seven-point scales of language performance designed after Elliott et al. (1967). The purpose of the experiment was to

see whether elements of self-fulfilling prophecy (a biasing effect) would affect the outcome of the measurement task. Their findings failed to show any biasing effects as a result of the case history information. The authors suggested that this may have been due to the use of group mean scores which would have disguised individual variability. Likewise, their study was conducted with graduate students whom they suggested might be more resistant to induced bias than speech pathologists in other settings. With regard to differences between the objective and subjective measures used by the evaluators, it was noted that while there were not significant differences between groups, the subjective scores were variable. This suggests a greater likelihood of bias occurring when the task is essentially subjective.

Meitus, Ringel, House and Hotchkiss (1973) also explored the potential effects of false case history information on judgments of students regarding severity of the speech disorder and the formulation of a prognostic hypothesis. Several biasing parameters were manipulated in order to derive three categories: positive, negative and no case history. Among the elements altered were factors of intelligence, family status, emotional status, medical and attitudinal history. Students then judged videotaped samples of verbal behavior on a formal phonetic inventory and rated performance on a five-point scale. in addition, students completed a four-point scale regarding prognostic and therapeutic judgments. Results were reported in terms of mean errors reported by each

group, and little variability was evident. Close agreement was also found on the scales relating to prognostic and therapy-related questions. Generally, there was no bias due to case history information. The authors interpreted this finding by noting that students should be influenced by case history information; the tone of their discussion was one of disappointment at not finding some "bias." They indicated that the case history needs to be a more useful tool than a comment on the past. They state that "The information gleaned through history taking must have relemance to the present or else it becomes just another meaningless exercise the student is required to fulfill" (p. 150). They further indicate that the formulation of a clinical impression is one of the cornerstones of clinical practice. In interpreting these comments, it is apparent the authors' use of the term "bias" differs from that of other writers. Clearly the literature in speech pathology, as well as in other fields such as psychology, uses "bias" to refer more closely to a notion expressed by Noll (1970): "In any situation where one is assessing some aspect of behavior, inevitably the particular bias of the evaluator can influence his judgments."

Lass, Browning and Brown (1975) further pursued the question of bias in the clinical judgments of speech pathologists. They sought to explore the effects of experience and educational status of the examiner, as well as the case history information. The population in the Lass et al. study included

three groups of student speech clinicians. One group contained beginning undergraduate students with minimal coursework in speech pathology. The second group was comprised of advanced undergraduate students, each of whom had at least 26 credits in speech pathology and a minimum of 120 clock hours of supervised practicum experience. The third group included advanced graduate students at least half way through their graduate coursework who had a minimum of 100 clock hours of practicum at the graduate level. Their task was to rate 17 speakers whose speech samples were presented on audio tape on two different occasions. Under one condition, the students were given no background whatsoever and were asked to rate the degree of severity on a four-point scale. The second rating session was preceded by the distribution of case history information which related to the speech parameters under study, i.e., case history data were fabricated to suggest the presence of specific types of speech problems. In some cases, the implied or suggested disorder actually did exist, whereas in others it did not. Results indicated that the students having the least amount of experience and coursework rated speech most severely and tended to be most influenced by case history information. Significant differences existed across all 10 parameters of speech investigated and among the 17 speakers, the two sessions and the three groups of student clinicians. In addition to experience, another explanation for the differences might have been the parameters under investigation. Lass et al. suggested that 10 parameters

may have been too many to permit reliable evaluation. Perhaps most important, the type of information given to the clinicians was more directly disorder-oriented as opposed to the kinds of social and educational data given in previous studies in speech pathology. Nevertheless, the authors suggested that predisposing information can bias speech pathologists' judgments and that the topic is worthy of continued research.

Wilson and Gasek (1975) explored the question of whether pre-information would influence speech clinicians' ratings of a single child's articulation. They also were interested in seeing whether experienced or inexperienced clinicians were more susceptible to bias given the pre-information.

Experienced clinicians were defined as employed speech clinicians with at least one year of paid professional experience, whereas the inexperienced sample was composed of undergraduate students majoring in speech pathology.

Subjects in both groups were assigned to one of two treatment conditions. In one the final sentence of a written case summary contained a statement indicating the child's articulation problem was of a mild-to-moderate type, whereas the other condition specified moderate-to-severe. A video tape of the child responding to an articulation inventory was presented as the stimulus, and subjects were asked to rate severity on a nine-point equal-appearing interval scale.

Results indicated that bias was induced as a result of the different pre-information statements. These differences

in ratings were most noticable in the population of experienced speech clinicians. The authors concluded that

> ...such imprecise written descriptions as 'he has a mild articulation problem' or 'he has a moderate articulation problem' or 'he has a moderate-to-severe stuttering problem' if used with no definite standards for application of the descriptive terminology, may well influence the clinician receiving the information (p. 21).

They recommended that in the clinical exchange of information attention be given to detailing specific behavior and to avoiding the use of subjective descriptions.

The Wilson and Gasek study is the only known published work involving working professional speech pathologists. It is also only the second study of examiner bias in speech pathology in which bias has been demonstrated. For these reasons it bears close scrutiny.

Wilson and Gasek did not address the issue of possible biasing effects which may be introduced through the use of video tapes. Several authors in psychology, among them Auffrey (1975), have shown that a number of qualitative judgments are made based on the appearance of the test subject. In addition to Wilson and Gasek, the Meitus et al. study also failed to control for the influence of physical appearance.

A second point in question are the authors' dexcription of an "experienced" clinician. They either failed to gather or failed to report information concerning the educational level of the participating professionals. Considering the

implications of their findings, this appears to be a critical point.

A third issue is the direction of the biasing statements. Wilson and Gasek employed varying degrees of negatively biasing statements and did not explore the possibility of shifting judgments in a positive direction based on information suggesting the absence of problems. Likewise, they did not employ neutral conditions or other forms of control.

Considering the major design problems in the Wilson and Gasek study, their results must be viewed with some skepticism. At the same time they did employ a strategy similar to that of Lass et al. in using biasing statements directed at speech functioning and did find biasing occurring. They noted, as had Beasley and Manning, that the more subjective the measure, the greater the likelihood of bias occurring. Their work provides an additional stimulus for the design of the present study.

Wilson and Gasek did use a group of working professionals in their study and noted the presence of bias as a function of pre-information. However, several issues concerning the study's design raise questions regarding the validity of their findings. The result is that the professional population Beasley and Manning (1973) and Lass et al. (1975) recommended be examined has not yet been approached using a carefully controlled experimental design.

Recently Naremore and Hipskind (1979) reported results of experimentation on the evaluation of the speech and

language performance of educable mentally retarded children by graduate student testers. The authors' concerns were for internal stereotyping behavior because

> While speech and hearing professionals may have been trained to disregard the results of previous tests in observing behavior, it is likely that they will be unaware of their own stereotypes and thus unable to excape their invluence (p. 28).

The study was designed to examine whether stereotyping did occur and whether this form of bias would affect the evaluations made on normal children and mentally retarded children.

The graduate student subjects rated "expected" speech and language performance on a set of bipolar characteristics based on a short case history-like paragraph. Two such descriptions were given for normal children and two for educable mentally retarded children. One month later the same graduate students listened to four tape recorded speech samples, two of which were normal speakers and two that were mentally retarded. According to the authors, all four had language skills similar to those of other normal and mentally retarded children in their respective age groups. They indicated that none of the children on the tape evidenced articulation or grammatical errors. About the only difference, according to the authors, was that the educable children were less fluent, with one child having a high incidence of repeated words and phrases. The students who acted as judges were only informed that there were both retarded and normal children on the tapes.

Analysis of the data indicated that the speech of the educable mentally retarded children was judged to be less correct, less fluent and less complex than that of their normal counterparts. In short, the judges had stereotypic ideas about the speech and language skills of both groups; and this was evidenced in their judgments of the speech samples. In all instances children identified as retarded were rated lower than normal children.

The authors raised several issues relative to the notion of stereotyping. They inquired about the extent to which predisposing information should alert the clinician to various concerns, much as Meitus et al. had expressed in their study. They did note, too, that labeling may confound the sppropriate balancing of necessary individual information and generally recognizable characteristics of various populations. They underscored the need to be cognizant of the possible existence of this contaminant in evaluation and remediation.

The Naremore and Hipskind study approached the issue of bias from a different perspective. This view is beneficial in that it underscores the need to be aware of several possible sources of influence on clinical judgments. Its perspective was one of a predisposition to a class of subjects or category of behavior, and the graduate student population was found to be influenced by that. It would be interesting to examine responses of other groups varying as a function of experience in diagnosis and remediation, for example, the

working professional. Likewise it would be interesting to see whether the same relative level of stereotyping existed across several clinical populations.

The idea of examining various facets of bias among working professionals appears to be a viable one. In addition to the aforementioned Wilson and Gasek study, the only other investigation of bias among professionals was a paper given at the Michigan Speech and Hearing Association annual meeting by Flahive and Magistro (1974). The professional exercise described had been part of a county speech and hearing association workshop. Participants were public school clinicians with various amounts of work experience. The thirty-three subjects had been randomly assigned to one of three treatment groups (positive, negative and no case history information conditions). Experimental groups met in different rooms and listened to tapes of a youngster responding to an articulation test. Their task was to develop and record diagnostic/prognostic impressions. Prior to the presentation of the tape sample, case history information was distributed. Subject responses were scaled with values from 1-5. Results of this nonrigorous exercise suggested that the groups of speech pathologists were not biased by the predisposing information.

In a study not specifically related to the delivery of speech services, Auffrey (1975) used speech pathologists as one of three groups of professionals who evaluated mentally impaired program candidates. The physical attractiveness of

the mentally impaired individual was the source of potential bias. The evaluation of the retardate consisted of judgments of personal qualities and general diagnostic, prognostic and program placement determinations. Results revealed significant differences in evaluation as a function of the physical attractiveness of the candidate and also as a function of the professional group of the subject. Higher recommendations for program placement and higher scores on a projective diagnostic statement were assigned to the more attractive mentally impaired persons. Auffrey noted that differences in evaluation were a function of training and experience. The speech pathologists in this study performed in a similar fashion as work-study coordinators. Counseling trainees differed in their responses by giving higher score values, a fact which would suggest that, while bias existed as a function of attractiveness, it was greater in the less experienced counseling trainee.

There are several problems with the Auffrey study. His description of speech pathologists indicates a wide variety of years of experience and educational level. Approximately thirty percent of his speech pathologists had Master's degrees' the remainder were bachelor's level subjects. Since critical variables were educational level and years of experience, a more detailed description of responses should have been given. However, his study does suggest a need to consider controlling factors related to physical appearance when exploring clinicians' ratings of performance.

Generalization to speech pathologists as a whole is further confounded by the fact that the task was not typical of professionals in the discipline. While speech pathologists may occasionally function as a member of a habilitation team working with adult retardates, it is not a common setting; and the responsibilities associated with making vocational potential judgments are foreign to them. Given the minimal level of training of the speech pathologist sample and given the fact that the majority of these individuals were public school clinicians, Auffrey's judgments must be viewed with caution.

In summary, the studies of Beasley and Manning (1973) and Meitus et al. (1973) indicated that students could not be biased by case history information of primarily socioeconomic, educational and intellectual types. Lass et al. (1975) induced bias by presenting students with a repeated measures task wherein the second presentation was accompanied by information which "...was fabricated in such a manner as to implicate the presence of a specific type of speech disorder in the speaker" (p. 108). It appears, therefore, that case history information which implies a speech disorder may contribute to biasing the examiner. Although Beasley and Manning and Lass et al. differed in their findings, both studies suggest the need to study the professional speech pathologist. Beasley and Manning indicated that, until further research is carried out, speech pathologists should be cautioned concerning their diagnostic activities, particularly "in settings

where time and/or administrative policy simply do not permit the speech pathologist to administer a battery of objective speech and language measures. An example of such a setting is the public schools" (p. 100). Lass et al. underscored the need to consider the professional based on a difference that educational training settings and professional work environments are inherently different.

The present study was designed to replicate and extend previous work. Several goals influenced the development of the experimental questions:

1. The first goal was to modify the Lass et al. design to include a population of well-defined speech pathologists in addition to the population of graduate and undergraduate students.
2. A second goal was to modify the Lass et al. biasing strategy by addign positive and neutral case history conditions to the existing format which had only negative case history statements. At the same time, a larger number of normal speaking samples were included.
3. A third consideration was the inclusion of equal-appearing interval scales to measure evaluator responses. The approach was employed in three of the research studies cited as investigating speech pathologists and the issue of bias. Specifically, elements from both the Meitus et al. and the Lass

et al. studies served as the basis for the development of equal-appearing interval scales for the present study.

4. A Further extension beyond previously reported investigations was the evaluation of judgment task reliability. Inclusion of this dimension in studies of examiner bias had been recommended by Barber and Silver (1968).

Research Questions

In order to examine critical issues regarding examiner bias in speech pathology, the following research hypotheses were posited:

1. It is hypothesized that experimental subject groups will not differ in the accuracy of identification of speech sample type given a "no case history" condition.
2. It is hypothesized that experimental subject groups will not differ in the accuracy of identification of speech sample type given "case history" conditions.
3. It is hypothesized that case history statements will not have an effect of the ratings of severity of speech problems across presentations.
4. It is hypothesized that subject groups will not differ from expert judges on ratings of speech sample severity.

METHODS

Introduction

Two critical variables in the evaluation of speech performance are educational level and the amount of clinical experience. Previous studies relied on students in training as subjects for experimentation. These individuals had varying amounts of educational background and no paid professional experience. In the single study reporting responses of working speech pathologists, the authors failed to identify the levels of educational training of the participants. The present study utilized students at two specific levels of academic experience and a group of working professionals satisfying several criteria related to level of academic preparation and years of professional experience. The population of professionals was selected from similar settings in the public schools.

Collectively articulation and voice problems represent a percent of typical caseloads for many public school clinicians. Likewise, students in training are often assigned young clients exhibiting problems in either of these categories. As a result the present study utilized elementary school age speakers demonstrating either normal voice and articulation, voice problems or articulation problems.

Several attempts have been made to quantify speech production or attributes of the process. Methods of scaling have been examined and identified as potential psycholphysical methods applicable to this task. The method of equal-appearing intervals is a scaling procedure which has been shown to be effective in making judgments concerning articulation proficiency and is simple and reliable (Morrison, 1955; Sherman and Morrison, 1955; Sherman and Moodie, 1957; Prather, 1960). It has also been recommended as a source of quantifying various attributes of voice production (Wilson, 1979). This strategy was utilized in the present study to allow listeners to attach a numerical value to various speech behaviors.

Previous authors in speech pathology have relied on fabricated case history information in order to potentially bias their experimental subjects. In several cases this information took the form of negative socio-economic statements. In these instances bias was not able to be induced. One additional study used case history statements which related to the presence or absence of speech problems. In this instance bias was generated the present study employed three levels of information relative to the existence or absence of speech production problems.

Experimental Subjects

Three groups of subjects were used: experienced public school speech pathologists, graduate students and undergraduate students training in speech pathology. There were

twenty-five subjects in each of the three samples for a total of seventy-five subjects.

An "experienced speech pathologist" was defined by the following criteria:

(1) present employment in a public school setting with responsibility for speech therapy activities;
(2) a minimum of three consecutive years of experience;
(3) possession of a Master's degree (minimum) in speech pathology.

A "graduate student training in speech pathology" was defined by the following criteria:

(1) present enrollment in a speech pathology graduate training program;
(2) successful completion of at least:
    (a) 20 semester or 30 quarter hours of academic coursework;
    (b) 50% of the practicum hours required for completion of the degree program; and
    (c) no more than one year of postgraduate work experience.

An "undergraduate student training in speech pathology" was defined by the following criteria:

(1) completion of basic coursework in phonetics, a survey of speech and language disorders, and basic information on voice and articulation disorders;
(2) at least 15 observation hours but no experience in independent diagnosis of speech disorders.

A general requirement for any subject was that he/she be naive to the purpose of the study. In addition, because of teaching responsibilities of the examiner, an independent, paid test administrator was employed for a portion of the study. This was done to control for possible examiner effects.

All participants were either attending school or employed within the State of Michigan. Subjects were volunteers who were assured of total anonymity throughout the experiment. Student subjects were obtained from the Department of Audiology and Speech Sciences at Michigan State University and the Department of Communication Disorders and Sciences, Wayne State University. Speech pathologists were solicited from throughout the state. Principally they were employees of the Detroit Public Schools and Ingham and Macomb Intermediate School Districts.

Ethical Issues

The question of administering an experimental task to subjects who are naive to the total purpose of a study raises certain ethical questions. The American Psychological Association (APA) Guidelines, "Ethical Principles in the Conduct of Research with Human Participants" (1973), suggest that any subject should be informed of all features of research that

> reasonably might influence a willingness to participate and to explain all other aspects of the research about which the participant inquires (p. 29).

Likewise, the guidelines stress honesty and openness:

> When the methodological requirements of a study necessitate concealment or deception, the investigator is required to ensure the participant's understanding of the reasons for this action and to restore the quality of the relationship with investigator (p. 29).

One guideline recommended for experimenters to consider is to weigh the benefits of a particular project versus the potential risks. As a result of the APA guidelines and University and Department Guidelines, the following procedures were implemented to ensure the rights of all participants were adequately protected:

(1) Potential subjects were informed about the purpose of the study as fully as possible without contamination. Subjects were told that the study was an investigation of determinants of clinical judgments under different conditions of information sufficiency.

(2) All subjects remained totally anonymous. The initial subject response sheets contained general demographic information and an experimenter-assigned number that identified the individual in subsequent responses.

(3) A summary report was made available to all participants.

(4) The present study was in compliance with all requirements of both the University and Department Guidelines on Research with Human Subjects.

Speech Sample Selection

A master stimulus tape was made with samples of thirty-one children who ranged in age from six to thirteen. Seven of these had no previously reported speech production problems, nine were reported to have voice problems and twelve were said to have articulation problems. All youngsters identified as having speech production problems were enrolled in speech therapy.

Eash ov the thirty-one children was tape recorded under quiet conditions. A high quality tape recorder (Sony TC-106-A) and microphone (Shure, Unidyne III) were employed. Intensity was held constant by means of the automatic level control feature of the recorder. High quality recording tape (Scotch, Low Print/Low Noise) was used for all recordings.

The Sounds-in-Sentences sub-test of the Goldman-Fristoe Test of Articulation was administered to all children. Specifically, the Story of Jack and Ricky was used as the stimulus material. Directions were given per the examiner's manual, and each child was required to repeat the story with sequenced picture stimulation.

Stimulus Tape Preparation

The master tape containing the thirty-one samples was played for three speech pathologists who acted as judges. Each of the three had extensive experience in diagnosis and the training of evaluation skills. Specific criteria for

selection of judges included

(1) eitht years (minimum) of experience as a speech pathologist;

(2) three years (minimum) of experience teaching diagnostics and/or supervising speech evaluations.

The three judges who participated in this study were university professors who averaged sixteen years of professional work past their Master's degree; all had Ph.D. degrees, taught diagnostics and supervised clinical evaluations. The average length of experience in teaching diagnosis and evaluation was eleven and a half years.

Each judge listened to the thirty-one samples using the Sony TC-106-A tape recorder with an AVID 8-jack audio distribution unit and standard AVID H/88 headsets. This system was identical to the one the experimental listeners were to use. Following completion of each individual child's passage, the judge stopped the tape and completed a response form. Judges were asked to determine whether or not the speech sample was normal. If not, the task was to rank the primary speech production problem on a seven-point equal-appearing interval scale with one being the least severe and seven the most severe. This ranking procedure was similar to the experimental task. (Detailed information regarding judges' protocols may be found in Appendix A.) The instructions indicated that children with problems were selected in either the voice or articulation categories; however, judges were told to make notations

concerning any voices about which they had questions about. In addition to the ranking task, the judges were asked whether the quality of the samples themselves were adequate to make judgments and whether the speaker's voice was so unusual that they might be recognized in a second presentation thirty minutes afterward.

From the responses of the judges, six children in each of the three categories were selected for inclusion on the stimulus tape. These eighteen samples represented the six highest interjudge agreements in each of the three categories (normal, voice and articulation). While the variable of degree of severity was not controlled in the stimulus sample procurement process, the majority of problem speakers were rated in the "mild" range with values from 2 to 5. Table 1 contains results for the eighteen speakers selected for the stimulus tape. Data included represent values for each of the three judges, for each of the eighteen speakers and the corresponding means and standard deviations. The average rating for speech problems was 3.6, the middle value on the seven-point scale.

The eighteen speech samples were removed from the master tape and extraneous noise and "dead" space was spliced out by hand. This was done to reduce noise and passage length cues which might be in effect during the second presentation. At the same time, it reduced passages to a more uniform length, the range being 43 to 59 seconds with a mean of 50 seconds. Several lists of random numbers from one to eighteen were

Table 1. Scores of the three judges for the eighteen speech samples selected for use on the stimulus tape.

Original Order Number	Assigned Experiment Number	Speech Production Type	Judge's Ratings 1	2	3	Mean	S.D.
2	1	Voice	5	4	5	4.7	.58
3	2	Articulation	1	3	2	2	1
4	3	Normal	N	N	N	-	-
5	4	Normal	N	N	N	-	-
6	5	Voice	3	3	4	3.3	1.8
7	6	Normal	N	N	N	-	-
8	7	Normal	N	N	N	-	-
9	8	Voice	6	5	4	5	1
11	9	Normal	N	N	N	-	-
12	10	Articulation	4	6	4	5	1
13	11	Articulation	2	3	3	2.7	.57
16	12	Voice	4	4	5	4.3	.57
17	13	Articulation	5	7	6	6	1
19	14	Voice	3	3	3	3	0
20	15	Voice	2	3	3	2.7	.57
25	16	Normal	N	N	N	-	-
30	17	Articulation	2	3	3	2.7	.57
31	18	Articulation	2	3	2	2.3	.57
						$\bar{X}$=3.6	S.D.=1.27

generated on the Digital Equipment Corporation PDP 11/40 computer which is housed in the Department of Audiology and Speech Sciences. A program entitled "RANORD" was used for this purpose. There were several reasons for a variety of lists. First, since the design indicated repeated presentations, there was a need for at least two lists. Second, the design also called for a replication of two speakers from each of the three groups (normal, voice and articulation) to be randomly selected and systemmatically introduced in the first presentation as a measure of interjudge reliability. These replications occurred in every fourth position with the restriction that the same sample could not occupy an adjacent position. The addition of the reliability component resulted in a total of twenty-four speakers in the first presentation of the experiment. Table 2 contains the order of Presentations I and II of the experiment. Appendix B contains a detailed description sample randomization and ordering procedures.

Case Histories

The second presentation of speech samples was preceded by the introduction of case history information designed to induce bias. Lass et al. (1975) suggested that if the nature of case history information deals with the specific speech disorder under consideration, the result is a much stronger possibility of bias. Thus, they were able to bias all three of their student subject groups. One shortcoming of their approach was to use only negatively biasing information.

Table 2. Randomized list of speakers for presentations I and II.

PRESENTATION I			PRESENTATION II	
Presentation Order	Experimental Format	Speaker Type	Presentation Order	Speaker Type
1	1	Articulation	25	Normal
2	2	Voice	26	Normal
3	3	Voice	27	Voice
4	Replication	Normal	28	Articulation
5	4	Normal	29	Articulation
6	5	Articulation	30	Voice
7	6	Articulation	31	Normal
8	Replication	Articulation	32	Normal
9	7	Voice	33	Voice
10	8	Normal	34	Articulation
11	9	Voice	35	Articulation
12	Replication	Normal	36	Normal
13	10	Voice	37	Voice
14	11	Normal	38	Voice
15	12	Normal	39	Articulation
16	Replication	Articulation	40	Voice
17	13	Articulation	41	Articulation
18	14	Normal	42	Normal
19	15	Normal		
20	Replication	Voice		
21	16	Voice		
22	17	Articulation		
23	18	Articulation		
24	Replication	Voice		

A similar strategy to Lass et al. was employed in the present study. As indicated, the second presentation was preceded by case history statements. These were specific to the speech production problems listed on the response sheets (articulation and voice). An expansion of the Lass et al. approach included the addition of positive and neutral background information conditions. The purpose was to determine whether a shift in rating of speech samples could be brought about through the use of information of several different types. The following guidelines were used in fabricating the case history statements:

(1) Positive statements would suggest general well-being and the lack of apparent problems;

(2) Negative statements would suggest problems should exist;

(3) Neutral statements would contain ambiguous, irrelevant or incomplete information;

(4) Specific histories would be developed following the random assignment of various samples to specific case history conditions.

The final guideline relates to the many parameters which could have been addressed under the general definition of "bias." The issue of the strength of various statements in any category was determined to be a factor related to specific samples and the degree of change, if any, evidenced during the second presentation.

Experimental Procedures

Subjects participating in the present study were volunteers solicited from populations known to meet the general criteria established for each group. The project was described to prospective participants as one which involved "listening to samples of young children's speech and making judgments about severity of any problems." Subjects were informed that the time involved would be approximately one hour and that anonymity of responses would be maintained throughout.

Subjects participated in groups ranging in size from one to six persons with the average group consisting of four people. All were seated comfortably around a table which contained a tape recorder (Sony TC 106-A) for playback and an audio distribution unit (AVID 8-plug) with sufficient headphones for each participant (AVID H-88) and the examiner. The experimental task was conducted at several locations, all of which contained adequate seating and lighting and were relatively free of background noise.

After being seated, a response packet for Presentation I was distributed to each subject. This packet contained three pages of orientation materials and instructions, twenty-four response sheets and the final sheet which was the consent form. In the orientation protocol found in Appendix D note that demographic information was obtained from the

subjects on the first page of the orientation/instruction section. Following the description of the project, subjects were asked to remove the back sheet (consent form), read and sign it. The signed consent forms were then collected and specific instructions for responding to the samples were presented. The examiner read the instructions to each group. A copy of the presentation protocol may be found in Appendix D.

As part of the instructions, each subject was told that several decisions were to be made in response to each speech sample: were the speech production characteristics normal or not and, if not, how would they rate the degree of severity on a seven-point equal-appearing interval scale? The scale was arranged so that a value of one was to be assigned to the least severe and seven to the most severe. It was further indicated that they were to be concerned with the "primary" problem and therefore to score in only one category. Following judgments of severity, they were to respond to four questions which related to diagnostic/prognostic impressions of the child. These more "subjective" questions also employed a seven point equal-appearing interval scale. A sample response form can be found in Appendix D.

The examinaer indicated that the tape would be stopped following each sample to allow all participants to complete scoring. It was noted that following the first few samples, all groups readily adapted to the response format and almost all scoring was done as the samples were being given.

After the final speech sample of Presentation I, response packets were collected and placed in a large carton in obvious random fashion. A short break was announced in conjunction with changing the tape. The actual purpose was to provide time for a distractor task. At the beginning of the break period, the examiner indicated that a voluntary, anonymous questionnaire was going to be distributed and that their cooperation in responding would be appreciated. The announced purpose was to provide the examiner, a college training program director, with information about perceived professional needs austensibly to assist the examiner in strengthening his training program. Questions varied somewhat between the undergraduate, graduate and professional groups; however, all were directed at the time-consuming, distraction purpose. The task was announced as strictly voluntary; however, seventy-four of the seventy-five participants filled out questionnaires. The one individual who elected not to participate in that task was seen reading a novel for the time between presentations, and this was determined to be sufficiently distracting.

The distractor task lasted between ten and twelve minutes. Subjects were then asked to replace the headphones and listen to a second group of children. Response packets for Presentation II were distributed. Each contained eighteen response sheets. Once again the subjects were told that they were public school speech pathologists and the groups of children they were about to hear were transfers into their

responsibility area for the upcoming Fall. They were told that on the top of each response wheet they would find a summary statement concerning the child which had been "lifted" from his/her accompanying school records. They were directed again to consider whether the production was normal or not and to rate those they felt were abnormal on the seven point equal-appearing interval scale. The tape recorded samples were played in similar fashion to Presentation I. The examiner stopped the tape after each sample to allow scoring to be conpleted. Following the last sample, the response packets were collected and the subjects were thanked for their cooperation. In addition they were asked not to discuss the procedures with their colleagues who had yet to participate. Data were collected over a three month period of time

RESULTS

Introduction

The present study was designed to examine the effect of potentially biasing information on the responses of undergraduate students, graduate students and working professionals on an experimental task involving the rating of children's speech samples. The task varied as a function of speech sample type and background information. Subjects were asked to make several judgments for each speech sample given. The first of these was whether they believed the speech production to be normal or non-normal. If the subjects determined a problem existed, they rated the sample on a seven-point equal-appearing interval scale with the value one representing the least severe and seven the most severe. The instructions further specified that they were to rate the primary speech production problem only and that the problem would either be one of articulation or voice. Eighteen speech samples were selected from a larger pool of thirty-one speakers. The final stimulus tape was composed of six normal speaking children, six children with articulation problems and six children with voice problems. Judgments of deviant speech for the stimulus tape were made by a panel of speech

pathologists having competencies in both diagnosis and appraisal of speech disorders and in the training of students in these skills. The entire evaluation protocol for the judges may be found in Appendix A.

Data regarding the judges' evaluations for the eighteen stimulus speech samples are also included in Appendix A. Samples selected for use were assigned two random orders, one for each presentation of the experimental task. A high quality stimulus tape was prepared and twenty-five subjects from each of the three experimental groups (undergraduate students, graduate students and working professionals) participated in the evaluation task.

In examining the research hypotheses several experimental questions were asked:

1. Are the experimental groups consistent in judgments asross presentations?
2. What level of interjudge reliability exists for categorical judgments for each subject by experimental group?
3. Do subject groups differ in their ability to identify accurately the speech sample type on Presentation I (no case history condition)?
4. Do subject groups differ in their ability to identify accurately the speech sample types on Presentation II (case history condition)?
5. Does the accuracy of categorical judgments for Presentation II vary as a function of case history type?
6. What are the average severity ratings by subject group for each Presentation?
7. Do severity ratings vary between Presentation I and II because of the introduction of case history statements?
8. Does the type of case history information affect change in the direction of that case history type (positive, negative, neutral)?

9. What level of agreement exists for ratings of severity between experimental subject groups and the panel of expert judges?

Experimental Subject Groups

Undergraduate subject volunteers were obtained from the training programs at Michigan State University and Wayne State University. All twenty-five satisfied the academic coursework and clinical practicum requirements outlined in the instructions. All were naive to the purpose of the study, and none reported known hearing loss.

Graduate student subjects were all enrolled in the Michigan State University speech pathology program, and at the time of the experiment were in the last academic term of their program. These participants had attended a number of undergraduate training programs and twenty-three held Bachelor of Arts degrees, whereas the remaining two held Bachelor of Science degrees. All satisfied academic and clinical criteria specified in the instructions. None of the graduate students had any previous professional work experience, and one of the group indicated a known hearing loss. This individual indicated that the audio presentation system provided sufficient intensity and clarity for her to make adequate judgments.

The working professionals were volunteers from a number of school systems throughout the State of Michigan. All had at least a Master's degree in speech pathology. Sixteen indicated their degree to be a Master's of Arts, five a

Master's of Science, three a Master's of Education and one participant had a Ph.D. in audiology with a Master's degree in speech pathology. This individual indicated his work history included several years as an itinerant speech professional and present responsibility as a supervisor of an intermediate school district's speech pathologists. Data regarding professional experience were gathered by assignment to categories: three to five years experience, six to eight years experience and more than nine years of experience. Responses indicated that of the twenty-five, four persons had three to five years experience, nine had from six to eight years of experience and the remaining twelve had nine or more years of work experience. All were employed in the public schools at the time of the experiment. Seventeen of the respondents indicated they held Certificates of Clinical Competence from the American Speech-Language and Hearing Association. Three individuals indicated known hearing loss. One person described the loss as "very mild," and all three reported the audio presentation system to be adequate in both intensity and clarity to allow good judgments.

Data Reduction/Statistical Analysis

Data for the first experimental question which is concerned with consistency of group judgments, correct or incorrect, across presentations are reported as percent of error judgments. The reliability of categorical judgments by subjects in each group is reported as correlation coefficients.

This analysis involved results on the replication of samples of each type during Presentation I. Categorical judgments under the "no case history" condition (Presentation I) are reported as percent correct judgments by subject groups. A one-way between subjects analysis of variance was performed for each sample category to determine whether group accuracy differed as a function of training and experience.

Experimental questions four and five are concerned with group accuracy on Presentation II which involved case history information. Again, data presented according to group accuracy for each sample, and these samples were also categorized by case history type. Two-factor mixed design with repeated measures on one factor analyses of variance were performed to determine whether differences existed between groups for various sample and case history conditions.

The severity ratings for each sample are presented as means for each experimental group with ranges included. The table containing these data also reflects the average judgments of the group of expert judges whose determinations formed the basis of speech sample selection. A comparison of the ratings of these judges and the experimental subject groups is referred to in question nine.

The question of whether ratings of severity change as a result of the introduction of case history information is addressed in two ways. First, data are provided regarding average group severity ratings. These have been computed for all subjects making correct categorical judgments on

both presentations and include voice and articulation sample types. T-tests for correlated means were performed for each subject group on each sample pair between presentations. Results are reported by group and sample type. Secondly, results of a sign test are reported. This statistical measure reflects directional changes (more severe rating, less severe rating) which may have occurred between presentations as a function of the type of case history information provided. Tables and figures reflecting all analyses are contained in the body of the text. Supplemental figures and raw data are contained in Appendix E.

Subject Consistency

In the response format the initial question asked of each subject for each sample was this: "are the speech production characteristics normal or not?" The seventy-five judges responded to samples of twenty-four voices on the first presentation and eighteen on the second for a total of 3150 judgments. Six samples were introduced during Presentation I for purposes of reliability measurement. Responses to these replications are discussed in a subsequent section of this chapter.

Raw data on responses of all seventy-five subjects to the original eighteen speech samples for Presentations I and II are reported in Appendix E, Tables 1 and 2. These tables indicate the number of judgments of non-normal from the total number presented in the binary choice paradigm. In the first

presentation 984 samples or 72.9% were judged as non-normal. In the second presentation 1012 or 74.9% of the samples evaluated were judged as problematic. In both of these there were 1350 trials, two thirds (67%) of which were from children previously identified as having speech production problems. The totals for each group are given in the extreme right hand column of these tables. Undergraduate students consistently identified the largest number of problems and the working professionals the least number, although all experimental groups identified greater totals than were actually problem samples.

In Table 3 of Appendix E the data for each presentation are given according to the matched pairs between Presentations I and II. Judgments of normal/non-normal were consistent for the three groups between presentations. Correlation coefficients for subject groups were undergraduate students .96, graduate students .97 and working professionals .98. These values reflect consistency of judgments relative to the existence of a problem and do not involve issues of accuracy or sensitivity.

Figures 1 and 2 of Appendix E provide graphic representation of this information as a function of speech sample type. In all graphic displays the letters UN represent the undergraduate students, GR the graduate students and WP the working professionals. Note that the undergraduate students identified twice as many non-normal speech samples on the actual normal speakers as did either of the other two groups

of subjects for both presentations.

Reliability Measurements

In the present study reliability involved the question of whether subjects from each experimental group identified the same speech sample in the same category given several opportunities. In employing a repeated measures design, it is imperative that internal consistency be evaluated so that differences in responses between the two presentations may be inferred to be a function of other variables such as case history and not the result of confusion about normal speech production, articulation disorders or voice problems. In order to measure reliability, two speech samples from each sample category were randomly selected and introduced into Presentation I of the experimental task. Replication samples were interjected in every fourth position of the presentation with the primary restriction that the same sample could not occupy either the preceding or succeeding adjacent position. The specific protocol for generation of random numbers and development of the stimulus tape are found in Appendix B.

Data for these replications were collected and analyzed by combining accuracy judgments for each sample type across subject groups. Undergraduate student agreement was .70 on normal speech samples, .72 for voice problem replications and .70 on articulation samples. Graduate students' results were .80, .90 and .78 for normal, voice and articulation types respectively. Working Professionals agreement values were .62 for normal samples, .80 on voice problem types and .82

on articulation problem replications. These values are displayed in Table 4 of Appendix E.

Subject Accuracy

In the present study accuracy of judgments relates to the determination of speech sample categories by the experimental subjects. Categorical judgment refers to the selection of the speech production category which is consistent with that made by the panel of expert speech pathologists who originally rated all samples. Results are reported both according to responses of subject groups and by speech sample type. Figure 1 provides an overview of error response patterns for each of the three subject groups for Presentation I. Note that undergraduate students made categorical errors involving normals speech samples in over one-third of the cases. Likewise, the working professionals incorrectly categorized voice problem cases approximately twenty-five percent of the time.

Figure 2 summarizes the percent of categorical judgments for Presentation II. Again the undergraduate students have the largest error rate on normal speech samples and working professionals are highest on voice samples. The overall rate of errors for all subjects increased in Presentation II, while a decrease in the category of voice problems was evidenced.

Table 5 of Appendix E contains raw data on categorical errors for each of the six samples of each speech production type by subject group for Presentation I. The table also

includes the average number of errors across categories by subject type. Table 6 in Appendix E reports this average number of categorical errors for Presentation I as a percent of incorrect judgments. This is similar to the values found in Figure 2. Tables 7 and 8 of Appendix E provide a similar report for speech sample types and subject groups for Presentation II.

A one-way between-subjects analysis of variance was performed on the mean percent of agreement on normal speech samples and the results of that analysis are found in Table 5 of the text. As indicated, the F-ratio of 5.95 is significant at the .05 level, indicating significant differences exist between groups on responses to normal speech samples. The ω^2 of 0.116 is indicative of a moderate strength of association in that approximately twelve percent of the variance can be accounted for in the present experimental design.

Table 6 (p. 49) summarizes results of a one-way analysis of variance for the mean percent of agreement on articulation problem samples for Presentation I. The F-ratio of .487 was not statistically significant (α=.05). This indicates there were no significant differences between subject groups' mean percent of agreement on articulation speech samples for the first presentation. Table 7 (p. 50) presents results on an ANOVA for the mean percent of agreement on voice problem samples from Presentation I. The F-ratio was significant (α=.05) 6.07, indicating differences did exist between groups on percent of agreement for judgments on voice samples. The ω^2

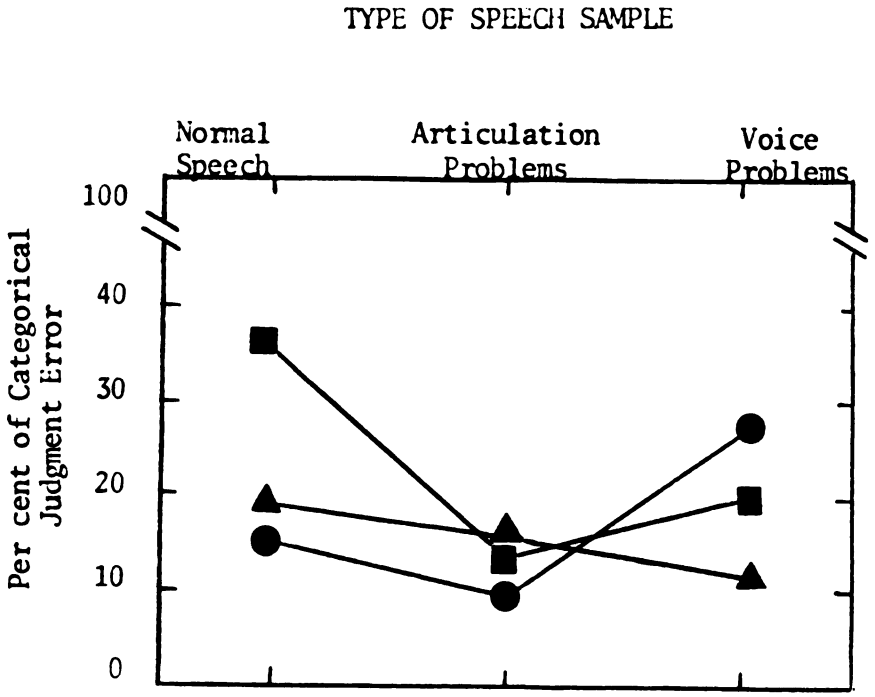

Figure 1. Mean percent error scores for each experimental subject group under each speech sample condition for Presentation II.

Key:

■ Undergraduates

▲ Graduates

● Professionals

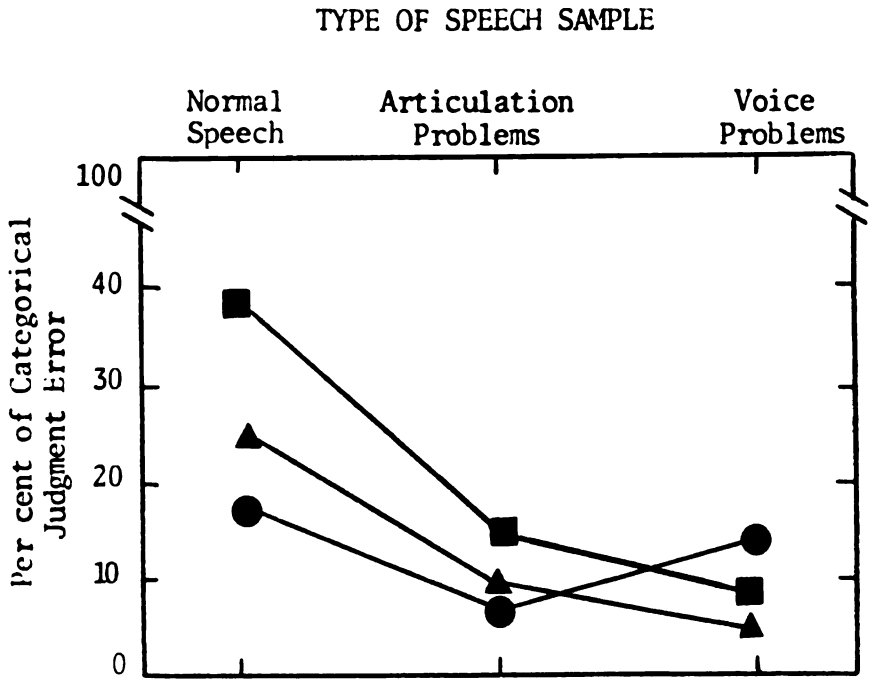

Figure 2. Mean percent error scores for each experimental subject group under each speech sample condition for Presentation I.

strength of association value of 0.118 is moderate with approximately twelve percent of the variability being accounted for in the present design.

In Presentation II the issue of average correct categorical judgments by subject group was also examined and statistically analyzed. A two-factor mixed design with repeated measures on one factor analysis of variance was performed for each of the speech sample type conditions. In these analyses the conditions of positive, negative and neutral case history were compared across subject groups. On Presentation II subjects had two trials under each case history condition with possible accuracy outcomes of 0%, 50% and 100%. For the statistical treatment of comparing mean percent correct, certain subjects from each group were excluded from the computation as the ANOVA format does not accomodate zero values. Table 8 (p. 52) indicates the number of subjects in each analysis that were included in the statistical computation from the original sample of twenty-five subjects in each cell. When these raw data are converted to percent of subjects in each cell the range extends from 68% to 100% participation in the analysis with a mean of 90.2%. The smallest cell (lowest number of correct judgments) was 17 by undergraduate students on normal samples. With the exception of the relatively low accuracy (68%) the remaining groups and judgments were all above 84%.

Table 9 (p. 53) contains results of the two-factor mixed design, repeated measures on one factor analysis of variance

Table 5. Summary table for a one-way between-subjects ANOVA for mean percent of agreement on normal speech samples in Treatment I.

SOURCE	SS	df	MS	F	ω^2
Treatment	5209.7	2	2604.8	5.95*	0.116
Error	31530.6	72	437.9		
Total	36740.3	74			

*significant at .05 level

Table 6. Summary table for a one-way between-subjects ANOVA for mean percent of agreement on articulation problem speech samples in Treatment I.

SOURCE	SS	df	MS	F
Treatment	230.4	2	115.18	.487
Error	17041.8	72	236.69	
Total	17272.2	74		

Table 7. Summary table for a one-way between-subjects ANOVA for mean percent of agreement on voice problem speech samples in Treatment I.

SOURCE	SS	df	MS	F	ω^2
Treatment	2965	2	1482.5	6.01*	0.118
Error	17756.9	72	246.6		
Total	20721.9	74			

*significant at .05 level

for the six normal speech samples of Presentation II. Results indicate that significant differences existed between subject groups and across case history types. At the same time, the results of the interaction condition were not significant.

Table 10 contains results of a similar ANOVA treatment of the mean percent correct values for the articulation problem samples of Presentation II. Again there were a total of six samples, two each of the positive, negative and neutral case history types. Results of this analysis did not reach a level of significance (α=.05) for the between subjects condition. However, there were significant performance differences found as a function of case history type (trials). The interaction condition of case history and subject groups yielded a value that did not achieve significance (α=.05).

Results of the third analysis are found in Table 11. This computation was performed on the mean percent correct values for each subject group on the six voice problem samples of Presentation II. Results of the between-subject analysis were significant (α=.05) indicating differences as a function of training and experience. Likewise, significant differences existed on the within subject condition which represented the varying case history types. As in the other two analyses, the trials by conditions interaction did not reach a level of significance.

In summary, results of the two-way ANOVA's on the mean percent correct judgments of category for Presentation II

Table 8. Number of subjects in each group involved in the computation of ANOVA results for case history conditions and speech sample types on Presentation II.

SUBJECT GROUP	SPEECH SAMPLE CONDITIONS		
	Normal	Articulation	Voice
UN	17 (68%)	24 (96%)	23 (92%)
GR	22 (88%)	24 (96%)	24 (96%)
WP	23 (92%)	21 (84%)	25 (100%)

Table 9. Two-factor mixed design: repeated measures on one factor analysis of variance results for normal speech samples on Presentation II.

SOURCE	SS	df	MS	f	P
TOTAL	110873.7	185			
Between Subjects	50873.7	61			
Levels of Experience	5490.8	2	2745.4	3.57*	0.0334
$Error_b$	45382.9	59			
Within Subjects	60000.0	124			
Case History Condition	10672.09	2	5336.0	13.1*	0.0001
Levels X Case History	1186.65	4	296.7	0.73	0.5757
Error	48141.26	118	408.0		

*significant at α=.05

Table 10. A two-factor mixed design: repeated measures on one factor analysis of variance on results for articulation problem samples on Presentation II.

SOURCE	SS	df	MS	f	P
TOTAL	86099.5	215			
Between Subjects	32766.167	71			
Levels of Experience	1006.7	2	503.35	1.0936	0.3427
$Error_b$	31759.467	69			
Within Subjects	53333.33	144			
Case History Condition	2245.3	2	1122.65	3.06*	0.0486
Levels X Case History	484.65	4	121.16	0.33	0.858
Error	50603.38	138	366.69		

*significant at α=.05

Table 11. A two-factor mixed design: repeated measures on one factor analysis of variance results for voice samples on Presentation II.

SOURCE	SS	df	MS	f	P
TOTAL	72705.3	206			
Between Subjects	27705.3	68			
Levels of Experience	5140.8	2	2570.4	7.52*	0.0015
Error	22564.5	66			
Within Subjects	45000.0	138			
Case History Condition	4806.8	2	2403.4	8.38*	0.0006
Level X Case History	2326.0	4	581.6	2.03	0.0928
Error	37867.0	132	286.9		

*significant at α=.05

indicated significant differences between subject groups for the normal and voice problem samples but not the articulation samples. In addition, significant differences existed under each speech sample category as a result of case history type. Finally, in all three analyses the interaction of trials and conditions failed to reach significant levels.

Another component of the accuracy issue concerns each individual speech sample and its relative degree of difficulty. While the expert judges varied little in categorical identification and ratings of severity, responses of experimental groups were not as consistent. Figures 3 to 8 in Appendix E present the categorical errors in percent by speech sample type. These histograms reflect the general error patterns of all seventy-five subjects taken collectively. Note that the articulation samples for both Presentation I and II and the voice samples for Presentation II have few errors. In contrast, the number of errors on normal samples was high for both presentations.

Figures 9 to 14 in Appendix E present data on the percent correct for each speech sample type by subject group. In contrast to the error analysis depicted in Figures 3 to 8, these histograms reflect the specific accuracy of judgments for each sample. Note the low accuracy on normal speech samples by the undergraduate student group and the relative difficulty working professionals had with the second voice sample in Presentation I. The histograms are arranged by case history type for Presentation II. Generally all three subject

groups appear to have had greater difficulty identifying normal speakers in the presence of case history statements of any kind. Accuracy for categorical judgments of articulation or voice problems appears to increase given case history information. The specific type of background statement does appear to be a critical variable influencing accuracy.

An additional component in the analysis of group accuracy concerns error selections. Figures 3, 4, and 5 (pp. 59-61) summarize the raw data for Presentations I and II. A total of 150 judgments for each subject group are present when collapsing across all samples in a category, twenty-five for each of six samples. In Figure 3 judgments for the normal speech samples are presented. Note that in both Presentations I and II the designation of "voice problems" was the most frequent error selection by all three subject groups. Totally, this designation accounted for seventy-five percent of the errors on Presentation I and nearly eighty percent on Presentation II. During both presentations undergraduate students evidenced greatest difficulty in correctly identifying normal speakers. The total error rate for all three subject groups did not vary significantly between presentations, a finding suggesting that errors of category judgment were a function of factors other than case history.

There are several possible explanations for the high (38%) overall error rate on normal speech samples by the undergraduate students. The first would be their relative

lack of clinical experience and inability to recognize the broad range of normal. It is also likely that several research biasing factors including the Hawthorne effect and demand characteristics influenced the performance of all three subject groups to a certain degree and that the undergraduates were most effected by these biasing influences. These experimental phenomena are addressed in detail in the discussion chapter.

In Figure 4 results of judgments are given for the articulation speech samples. Again, the total number of error selections vary little between Presentation I (54 errors) and Presentation II (53 errors). The distribution by group is also similar in that the undergraduate students have the highest rate of error and the inappropriate identification of "voice problems" were the most frequent error type. At the same time, it should be noted that undergraduate students and working professionals performed more accurately in making judgments on articulation samples than with the normal or voice problem types. Graduate students performed equally well on voice and articulation problem judgments.

Figure 5 contains raw data for responses to voice problem samples. Graduate students performed most effectively on both Presentations I and II. Undergraduate students and working professionals had similar numbers of judgment errors on Presentation I. However, given case history information on the second presentation, the professionals had the greatest number of errors with judgments on voice problems.

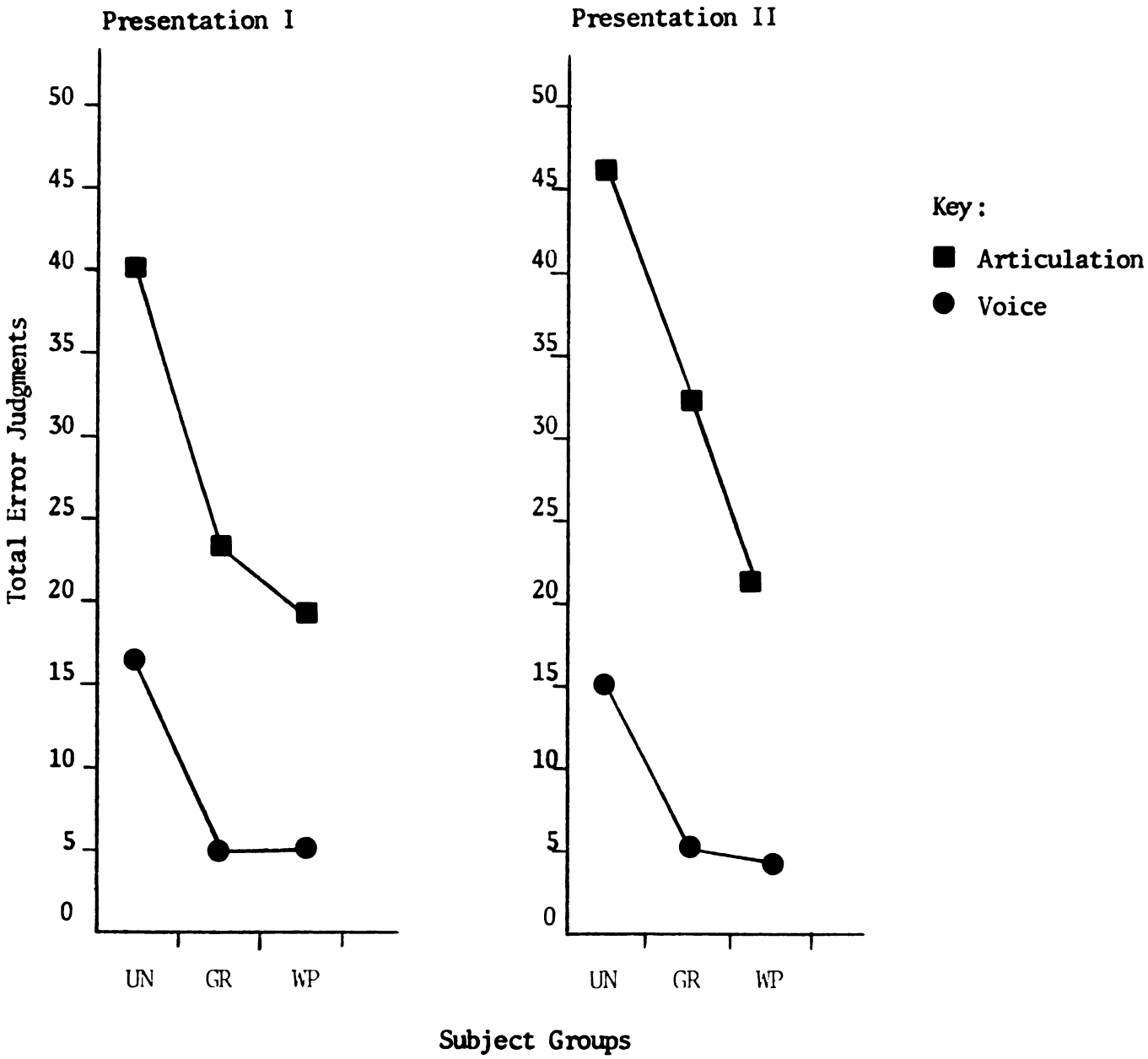

Figure 3. Combined error judgments by subject group for the six normal speech samples on Presentations I and II.

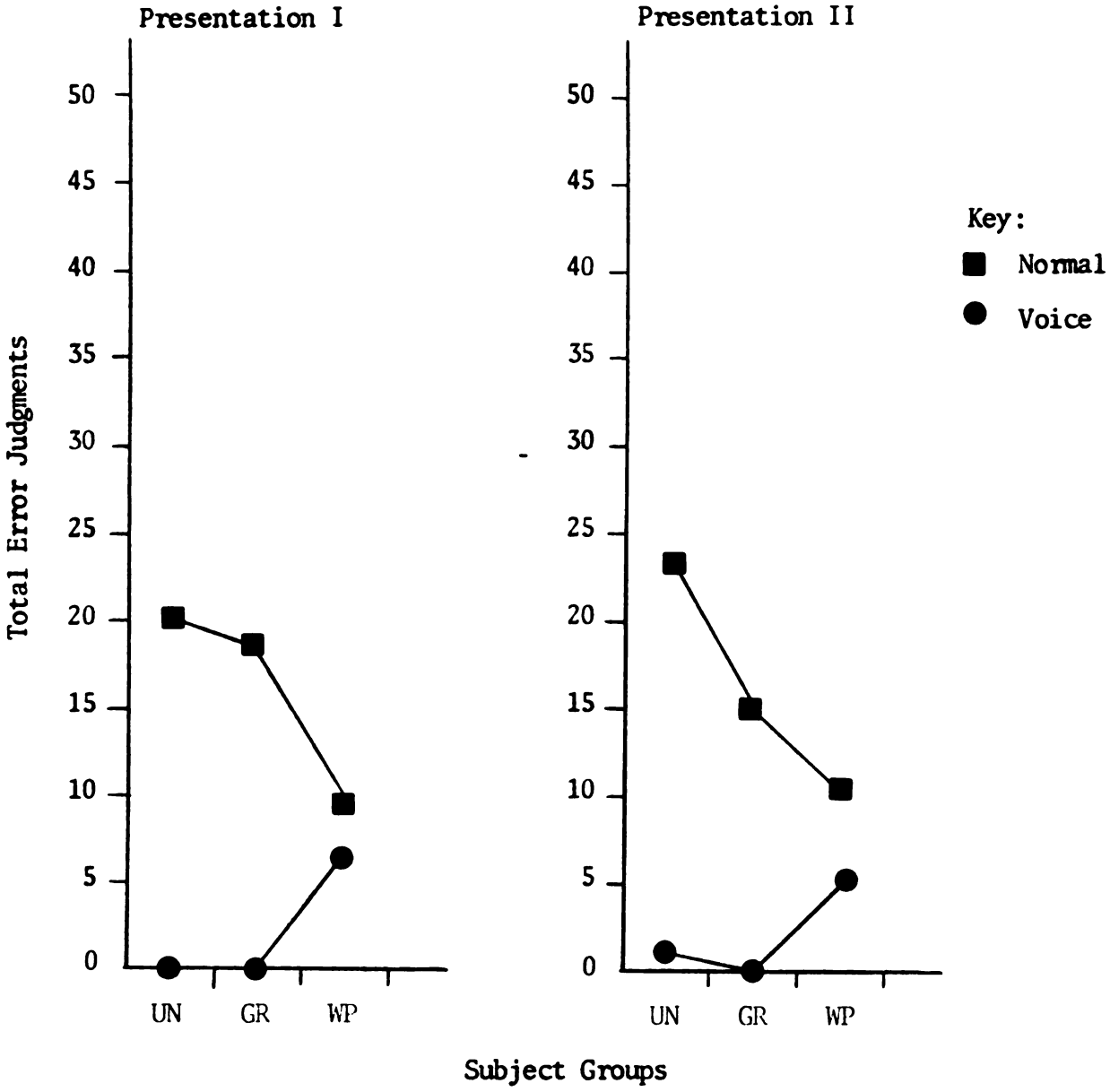

Figure 4. Combined error judgments by subject group for the six articulation speech samples on Presentations I and II.

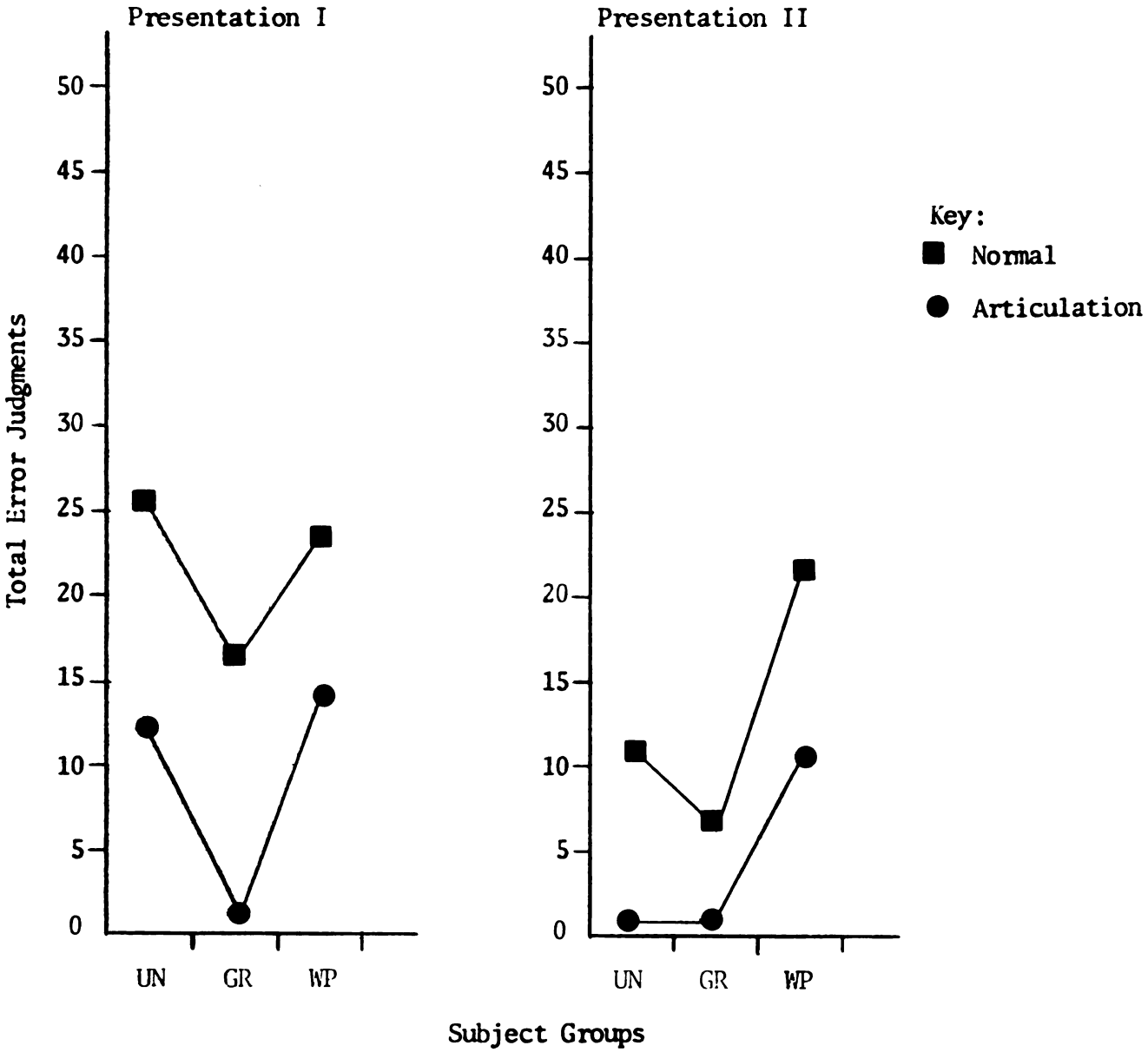

Figure 5. Combined error judgments by subject group for the six voice speech samples on Presentations I and II.

Sensitivity Ratings

Sensitivity is another critical issue analyzed in the present study. It was defined as the rating of severity in those samples judged to have speech production problems. Since judgments of normal could not, by definition, be assigned a weighted value, the data and discussions are restricted to the samples having voice or articulation problems. In the experiemntal procedure once the initial judgment concerning normal/non-normal speech production was made, subjects were asked to rate the degree of severity. The method of equal-appearing intervals was used with a seven-point scale on which a rating of "1" represented the least severe and "7" the most severe.

The results and discussion of sensitivity involve several components. Initially data are presented concerning the question of whether subject group ratings varied at all from one presentation to the other. This analysis examines the question in a general sense, that is, change regardless of the direction (more severe, less severe) or problem type (articulation or voice).

Following that, specific information is provided for each subject group concerning the percentage of group members who changed ratings between presentations. Results of statistical analyses are presented comparing those values. They also address the question of whether subject groups differ significantly in the proportion of members changing rating values as a function of speech sample type.

One of the major experimental questions concerns the notion of case history type and hypothesized changes in ratings as a function of either positive, negative or neutral predisposing information. The third component of the sensitivity discussion involves analyses of group responses as a function of case history type.

Table 12 contains data on the number of subjects in each experimental group who changed scaled sensitivity values for speech sample pairs between Presentations I and II. The N values represent the number of subjects from the total of twenty-five in each group who correctly identified the appropriate category on both presentations. The percent value reflects the proportion of that N who changed rated values between presentations. Each of the three subject groups had approximately the same percent of change across the twelve pairs: undergraduate students averaged 65.1%, graduate students 58.3% and the working professionals 59.6%.

Table 13 contains results of a one-way analysis of variance analyzing the mean values for change between subject groups. The F-ratio of .55 indicates that there were no significant differences (α=.05) between the percentage of subjects changing ratings across the three groups. Tables 14 and 15 contain ANOVA results of percent of change values by experimental group by speech sample type. These analyses were performed to determine whether the "no differences" conclusion reported in Table 13 was equivocal by speech sample type.

Table 12. Raw score and resulting percentage of subjects per experimental group who changed scaled sensitivity values for articulation and voice speech sample pairs across Presentations I and II.*

ARTICULATION SAMPLE PAIRS

	#1/#39		#6/#28		#7/#34		#17/#29		#22/#41		#23/#35	
	N	%	N	%	N	%	N	%	N	%	N	%
UN	25	60%	21	76%	14	36%	17	76%	24	67%	19	52%
GR	24	71%	24	58%	13	15%	16	69%	23	65%	24	42%
WP	23	70%	24	71%	19	32%	17	59%	24	38%	23	35%

VOICE SAMPLE PAIRS

	#2/#30		#3/#27		#9/#33		#11/#37		#13/#38		#21/#40	
	N	%	N	%	N	%	N	%	N	%	N	%
UN	14	86%	16	88%	19	79%	21	57%	25	48%	23	57%
GR	21	62%	11	64%	21	57%	25	56%	24	58%	23	83%
WP	14	43%	5	80%	20	60%	23	74%	24	75%	15	80%

*N represents the number of subjects from the sample of twenty-five in each group who successfully identified the appropriate speech category on both presentations.

Table 13. Summary table for a one-way between-subjects ANOVA comparing the percent of subjects in each group who changed ratings of severity between Presentations I and II.

SOURCE	SS	df	MS	F
Groups	320.39	2	160.2	.55
Error	9587.17	33	290.52	
Total	9907.56	35		

Table 14. Summary table for a one-way between-subjects ANOVA comparing the percent of subjects changing sensitivity ratings in each group on the articulation samples for Presentations I and II.

SOURCE	SS	df	MS	F
Groups	361.0	2	180.5	.53
Error	5099.5	15	5460.5	
Total	5460.5	17		

Table 15. Summary table for a one-way between-subjects ANOVA comparing the percent of subjects changing sensitivity ratings in each group on the voice samples for Presentations I and II.

SOURCE	SS	df	MS	F
Groups	125.4	2	62.7	.309
Error	3049.5	15	203.3	
Total	3174.9	17		

Resulting F-ratios of .53 and .31 for articulation and voice sample types respectively lead to the conclusion that there were no differences due to sample type in the percentage of each experimental group that changed rating values.

The sign test was employed in order to examine the relationship between the direction of rating change (more severe, less severe) and case history type (positive, negative, neutral). The hypothesis was that there would be no differences in sensitivity ratings between presentations in spite of the introduction of predisposing information prior to the second presentation of the speech sample. The sign test was based on the differences in either positive (less severe judgment) or negative (more severe judgment) ratings on the interval scale. Table 16 summarizes sign test results across subject groups by case history type. The letter "r" denotes the number of times the less frequent sign occurs. For this analysis subjects whose values were similar for both presentations, that is, differences were zero, were excluded from the computation. Included in the analyses, however, are results involving judgments of "normal" on one of the pair of presentations. The shift in either the positive or negative direction was presumed to be the result of the case history information. There are two cells where significant results are found (α=.05). One of these is with undergraduate students and neutral case history information. The direction of this shift is positive. This suggests that in the presence of non-speech related information the undergraduate

Table 16. Sign test results for responses to Presentation II by case history type.

		CASE HISTORY TYPE		
		+	-	~
UN	+ changes	27	28	50
	- changes	32	32	51
	r=	27	28	21*
GR	+ changes	13	20	29
	- changes	39	35	37
	r=	13*	20	29
WP	+ changes	18	28	37
	- changes	24	29	26
	r=	18	28	26

*significant at α=.05

Key: + = positive case history type
- = negative case history type
~ = neutral case history type

students judged the speech samples to be better than their original ratings. The other cell containing significant results is that of the graduate students and positive case history information. In this instance a significant proportion of graduate students rated samples with positive background information as performing poorer in Presentation II. In addition, it appears that graduate students tended to react negatively to any kind of predisposing information as results for the negative case history type nearly reached significance as well. Note that unlike undergraduates and working professionals, the graduate students had greater negative direction changes in all case history categories. In looking at the direction of change by case history type, it is interesting to note that both positive and negative case history resulted in larger numbers of negative changes. Only under the neutral case history condition were there greater positive direction shifts.

Tables 9 and 10 of Appendix E contain average ratings for each experimental group by first and second presentation pairs. To summarize these data, for the six pairs of articulation problem samples undergraduate's ratings averages 4.5 on the seven-point scale for Presentation I. Given case history information this overall average changed slightly in a positive direction to 4.2. Both graduate student's average ratings (4.28 for Presentation I and 4.31 for Presentation II) and working professional's ratings (4.19 for Presentation I

and 4.25 for Presentation II) indicated differences in the opposite direction. However, the differences for all three groups were so small as to be essentially negligible.

The data regarding voice problem samples are somewhat different. Although all three groups are in relative agreement concerning the degree of severity on Presentation I, the graduate students in Presentation II rate voice samples as considerably poorer (Presentation I, $\bar{X}$=4.9, Presentation II, $\bar{X}$=5.73). The change of this magnitude accounts for the significant difference among the groups reported earlier. At the same time undergraduate students and working professionals varied little (undergraduates Presentation I, $\bar{X}$=4.6, Presentation II, $\bar{X}$=4.74, working professionals Presentation I, $\bar{X}$=4.64, Presentation II, $\bar{X}$=4.55) between presentations. These results are likewise recognized when simply comparing total averages between presentations without regard for speech sample type.

Table 17. Average combined sensitivity values using the seven-point rating scale groups for Presentations I and II.

	Presentation I	Presentation II
UN	4.55	4.46
GR	4.59	5.02
WP	4.41	4.40

Table 18 contains results of t-tests for correlated means which were performed with values collapsed across sample pairs

Table 18. Results of T-tests for correlated means of rated severity of speech samples for Presentations I and II by speech sample type and subject group.

	CASE HISTORY TYPE					
	POSITIVE		NEGATIVE		NEUTRAL	
	Articulation Samples	Voice Samples	Articulation Samples	Voice Samples	Articulation Samples	Voice Samples
UN	t=.88	t=-.70	t=.52	t=-1.30	t=3.32*	t=1.90
GR	t=1.33	t=-1.90	t=3.0*	t=-2.14*	t=.78	t=2.48*
WP	t=1.19	t=1.43	t=.57	t=.68	t=2.1*	t=.66

*significant at α=.05

and reported according to case history type. As indicated, there are several cells in which the magnitude of change between presentations was significant. In examining these results by case history type, it is noted that the neutral information condition had the greatest number of significant changes; and there were no circumstances where significance was reached under positive case history conditions.

The final issue regarding sensitivity is concerned with the level of judged severity of each experimental group and the expert judges who had originally rated each sample. As specified in Appendix A, the protocol for the expert judges was essentially the same as that of Presentation I of the experimental procedure. In the methods chapter is was reported that the average overall sensitivity rating by the expert judges was 3.6, approximately mid-way on the seven-point scale. By having the average value in this position, respondents theoretically had approximately equal space in either positive or negative directions for response judgments to shift given case history information. Table 19 contains summary information on ratings from the panel of judges and the experimental subjects for the twelve speech problem samples rated during Presentation I. Note that with one exception all of the ratings of the experimental subjects are considerably higher (more severe) than those of the panel of experts. Ratings among experimental subjects, as previously indicated, were consistent between groups. The consistently more severe ratings of the experimental subjects may be a function of

Table 19. Average sensitivity range values for each experimental group and the expert judges taken from Presentation I.

Subject Group	Sample Type	Sample Number 1 (A)	2 (V)	3 (V)	6 (A)	7 (A)	9 (V)	11 (V)	13 (V)	17 (A)	21 (V)	22 (A)	23 (A)
	UN	3.2	3.4	4.7	4.0	6.6	5.1	5.5	4.9	3.2	3.9	3.8	6.3
	GR	3.1	3.9	4.9	3.5	6.8	5.8	5.5	5.8	2.8	3.5	3.3	6.3
	WP	3.1	3.9	4.6	3.5	6.5	5.7	4.9	5.2	2.8	3.5	3.3	6.0
	EJ	2.7	2.7	3.3	2.3	6.0	4.7	4.3	5.0	2.0	3.0	2.7	5.0

demand characteristics, a notion to be discussed in detail in the next chapter.

Summary

The present study resulted in a greater number of judgments of non-normal speech than actually existed. This was the case for all three subject groups. Groups were, however, consistent in their assignment of samples to categories across presentations. A replication of thirty-three percent of samples from each category during Presentation I resulted in high levels of interjudge reliability.

Significant differences existed in the accuracy of groups assigning speech samples to appropriate categories on Presentation I. Undergraduate students frequently identified normal speakers as having speech problems. The most common selection of disorder type were "voice problems." Recognition of articulation problems was at a high level of accuracy for all three groups in Presentation I. At the same time the working professionals had the most difficulty identifying voice problems accurately.

In evaluating responses to Presentation II it was noted the experimental groups followed a pattern similar to that for Presentation I. All three groups had a relatively high rate of errors on normal samples with the undergraduate students again being the least accurate. The groups performed accurately on articulation samples, and in Presentation II improved their accuracy in identifying voice problems.

In evaluating the error selections made overall, it was noted that "voice problems" were most frequently identified as the alternative to either normal or articulation problem samples for both presentations. Two-way analyses of variance indicated that groups varied significantly as a function of experience for normal and voice problem samples, a pattern seen in Presentation I. In all three analyses the case history condition was also significant. Interpretation of results at this level would suggest that groups do differ in their ability to differentiate various speech problem samples from normal and that this is a function of training and experience. Graduate students appear to be most accurate, whereas undergraduate students performed poorest. Given the significant differences for the three ANOVA's performed across Presentations I and II for the case history conditions, it is inferred that bias exists within all three groups based on case history statements.

The question of bias was examined further through evaluation of the ratings of severity of speech problems. Table 12 indicates the number of subjects from each group who made correct categorical judgments on both Presentation I and II and the percentage of this number who altered their sensitivity ratings from one presentation to the next. Values were consistent between groups with 65% of undergraduates changing, 58% of graduate students, and 60% of working professionals. Overall, 61% of the subjects who were successful in

correctly identifying the appropriate speech problem category on both Presentations altered their scaled severity rating as a result of case history information. These data strongly suggest that all three subject groups were biased by case history information.

T-tests for correlated means were used to explore the magnitude of changes between presentations. Results of these computations indicated significant differences between Presentations I and II in five of eighteen cells. These identified results and those reported as the percentage of change in Table 12 underscore the notion posited by Beasley and Manning (1973) that computations based on group means, such as the t-tests for correlated means reported in Table 18 may not be sensitive to individual differences and thus bias, because they are groups-based analyses. The computations may, therefore, obscure the identificiation of bias. Reexamination of Table 12 indicates that for articulation speech samples, twelve of the eighteen cells resulted in over 50% of the subjects changing their ratings in either direction (more severe, less severe) as a result of case history statements. For voice problems this increased to sixteen of the eighteen cells having a 50% or greater change rate. Summarizing, differences existed across Presentations I and II for all subject groups suggesting that all were subject to influence by case history information.

The issue of directionality of case history information (positive, negative, neutral) and the resulting changes of severity of ratings was explored. Undergraduate students reacted to a significant degree in a positive direction given neutral case history information. Graduate students reacted to a significant degree in a negative direction given positive case history information. Changes in sensitivity ratings for other conditions are presumably due to random factors.

Comparisons were made between the average severity ratings of the three experimental groups and the expert judges. Experimental subjects rated eleven of the twelve problem sample types more severely than the expert judges. Scale values for all three experimental groups were in close agreement.

DISCUSSION AND CONCLUSIONS

The review and discussion of findings in the present study focus on several areas. The use of equal-appearing interval scales as the dependent variable is discussed. Results of the present study are compared with previous findings. Related issues are identified and implications for training and professional practice and posited. Limitations of the present study are raised and suggestions for further study are made.

Dependent Variable

The dependent variable employed in the present study was a seven-point equal-appearing interval scale. There are two issues which warrant resolution regarding this strategy. The first involves the use of the particular psychophysical method of equal-appearing intervals as opposed to other approaches. The second issue deals with the application of equal-appearing intervals specifically to articulation and voice behaviors.

The method of equal-appearing intervals was first described by Thurstone and Chave (1929). Since that time several investigators in the field of speech pathology have examined the usefulness of this psychophysical method for

judging various parameters of speech behavior. Morrison (1955), Sherman and Morrison (1955), and Sherman and Moodie (1957) each explored the use of this method and found that this scaling procedure could be applied to articulation skills with good reliability. Prather (1960) compared equal-appearing intervals with direct-magnitude estimation. The latter is presumably a more powerful form of scaling as it results in ratio-type data. Prather concluded that for articulation measurements scale values obtained by direct-magnitude estimation were in very close agreement with those obtained using equal-appearing intervals. She concluded that, because of the closeness and linearity of the relationships between the two methods, the limitations of the method of equal-appearing intervals may not be important.

Young and Downs (1968) reiterate the popularity of the method of equal-appearing intervals and reason that this is due to the ease of administration and reliability of scale values and that there are minimal underlying assumptions concerning observers' abilities.

The conclusion is that the method of equal-appearing intervals has frequently been employed in making qualitative judgments of speech performance. The studies on which this investigation is based, namely, Beasley and Manning (1973), Meitus et al. (1973), Lass et al., and Wilson and Gasek (1975) all employ scaling to one degree or another.

As previously indicated, a number of early investigations of the use of equal-appearing interval scales dealt with

ratings of articulation proficiency. Morrison (1955), Sherman and Morrison (1955), Sherman and Moodie (1957) and Prather (1968) all concluded that this method was applicable.

Wilson (1979) recommends the use of equal-appearing intervals for voice evaluations, although he does point out that reliability has been a problem in studies of voice disorders. He has suggested that speech pathologists develop their skill in scaling by rating voices and comparing them with other speech pathologists, i.e., a method for developing an internal referent system. Bradford, Brooks and Shelton (1964) found reliability poor with both experienced and inexperienced speech pathologists who were not specifically trained for the task of rating hypernasality. On the other hand, Schulz, Heller, Gens, and Lewin (1973) found interjudge reliability to be 0.94 when employing a seven-point scale for judging nasal resonance. Lass et al. (1975) employed a repeated measure format to study examiner bias and included the rating of parameters of articulation and voice on a four-point scale. Their task involved rating voice characteristics of hypernasality, hyponasality, husky-hoarse, breathy, weak, pitch and volume. Their findings indicated differences in ratings from one presentation to the next, however, they speculated that too many parameters were being assessed at one time.

To summarize, the method of equal-appearing intervals has been employed in studies of articulation problems and

voice problems. It has been compared with other psychophysical scaling procedures and determined to be of essentially similar accuracy and considerably less complexity in computation. Seven-point scales are most prevalent in the literature. The present study, therefore, utilized a seven-point, equal-appearing interval scaling format for the dependent variable.

Examiner Bias

The present study was designed to explore the extent to which bias may influence the performance of clinical behavior. Inherent in experimentation are the possibilities of bias in the conduct of the task itself. At the outset it is important that various forms of bias be defined in order to determine which may have influenced the experimental results and which were actually under investigation in the experimental questions.

Several characteristics in behavioral research regarding interpresonal interaction have been identified by Gephart and Antonoplos (1969) as potential sources of bias. These are experimenter bias, demand characteristics, the Hawthorne effect, placebo effect and the halo effect. The authors stated that each of these "...acts in a role that possibly confounds the results of research through influencing the data generated and the conclusions reached" (p. 580). They further indicated that these five concepts can be differentiated in terms of the locus of their effect and the nature

of the error contribution. The locus of effect refers to the apparent place the biasing factor is found in the research process. For example, the Hawthorne effect is a frequently cited psychological phenomenon which is associated with unanticipated, disproportionate outcomes in experimentation. Cook (1967) defines the Hawthorne effect as:

> ...a phenomenon characterized by a cognitive awareness on the part of the subjects of special treatment created by artificial experimental conditions. It becomes confounded with the independent variable under study, with the subsequent result of either facilitating or inhibiting the dependent variables under study and leading to spurious conclusions (Gephart and Antonoplos, p. 581).

The locus of the novelty of the artificial experimental environment would typically occur during initial interaction between the subject and procedures. At the same time the awareness of experimental procedures would continue throughout the research process. Gephart and Antonoplos suggested that in these contexts the nature of the error with the Hawthorne effect would be to alter the treatment and provide a potential threat to the internal validity of the test of the hypothesis.

In the present study it does not appear that there was bias as a result of the Hawthorne effect. The overall performance of the subject groups did not appear "striking" nor did they "defy explanation in line with the procedures used and preexisting information" (Gephart and Antonoplos, p. 581). The nature of the experimental task, listening to speech samples and making clinical judgments, is not novel in the

training of speech pathologists, therefore, the effects of the artificial experimental environment were minimized.

The experimenter bias effect deals with the expectations held by the researcher regarding the results and other factors outlined by Gephart and Antonoplos:

> It involves the transmission of that expectancy to the subjects in a way that alters the normal functioning of the subject on the dependent variable central to the research being conducted. It should be added that the discussion here focuses on influence that is subconscious (p. 580).

In the present study several controls were exercised to minimize any effects of this sort. The experimental task was regidly described and implemented. The stimulus items were taped and presented according to the same format for all subjects. Instructions were read and questions and supplementary information which might have functioned as cues were minimized. Because of teaching responsibilities and possible influences on the graduate student population, a paid tester was hired to administer the experimental task to the graduate subjects.

A third form of potential bias outlined by Gephart and Antonoplos is that of demand characteristics. They indicated that according to Orne (1962), an experimental subject interprets the nature of the experimental procedures and then consciously and unconsciously contrives role demands. He specifies demand characteristics to be

> ...the totality of cues which convey an experimental hypothesis to the subject and become significant determinates of subject's

> behavior. We have labeled the sum total of such cues as the "demand characteristics of the experimental situation." These cues include the rumors or campus scuttlebutt about the research, the information conveyed during the original solicitation, the person of the experimenter, and the setting of the laboratory, as well as all explicit and implicit communications during the experiment proper. A frequently overlooked but nonetheless very significant source of cues for the subject lies in the experimental procedure itself, viewed in the light of the subject's previous knowledge and experience.

Given Orne's definition, the present study is actually an examination of the influence of demand characteristics under rigidly controlled experimental conditions. This study sought to explore the perception/performance characteristics of individuals at various levels of training and experience given specific cues. The task was limited to several levels of more routine clinical behavior with the stimulus comprised of "typical" cases for a common work environment. Clinical judgments were evaluated on two levels: the acceptability/non-acceptability of speech and the reaction/over-reaction to cues relating to speech behavior. The rigid control of cueing presumably reduced extraneous influences other than, perhaps, the effect of the actual experimental situation and the expectations of finding problems on the part of experimental subjects.

The result is that the locus of these demand characteristics is continuous. The effect of these characteristics are found at various levels of cueing and varying levels of training and experience. The results of experimental

questions were, in essence, reflective of demand characteristics as applied to the role of the speech pathologist. The result was that discussion of differences between experimental groups will necessarily involve discussion of differences which may exist in the roles of individuals at various levels of training.

Experimental Questions/Accuracy

Responses to the first experimental question dealt with the consistency of subject behavior between presentations. Data were analyzed according to group values and reported primarily as group means. Results for all three subject groups indicated consistent group performance in the number of normal/non-normal judgments between presentations. Correlation coefficients of 0.96 for undergraduate students, 0.97 for graduate students and 0.98 for working professionals all suggest consistent group behavior. It should be underscored, however, that these values do not reflect accuracy or sensitivity of judgments.

The second experimental question was of considerable importance. Reliability of judgments was a fundamental assumption in the present experimental design. In order to test for reliability, two samples of each type were reintroduced into Presentation I. These samples were randomly assigned, one to every fourth position in the order with the restriction that the same sample could not occupy an adjacent position. Given the number of samples rated on Presentation

I (24) and the pre-pubescent status of all speakers, it was assumed that each sample would be rated independently. In addition, the expert judges were asked as part of their rating to indicate whether they thought the individual sample had unusual enough characteristics that it would be identified based on those cues. The final stimulus samples did not have any judgments of this sort. Further cue reduction included consistent sample length and reduction of intersample noise or silence cues through splicing.

Subject's group agreement for normal samples ranged from 0.62 to 0.80, with the working professionals having the greatest difficulty with normal speakers. Presumably the professionals expected that listening to samples under experimental conditions would result in more problematic samples than were actually given. Another possible explanation is that over time these individuals have become more dependent on sources of judgment other than simply listening.

Group agreement for the voice and articulation samples was high. These values, ranging from 0.70 to 0.90, are in general agreement with Morrison (1955) who found values of 0.98 in rating articulation behavior and concluded that:

> Reliable mean scale values of the severity of defective articulation can be obtained for one-minute speech samples from the responses of a trained individual observer (p. 385).

These values are also in agreement with those of Schulz, Heller, Gens, and Lewin (1973) who had obtained 0.94 inter-judge reliability employing a seven-point scaling task with

voice cases. The implication of these moderate and high levels of agreement is that differences in performances between Presentations I and II may be inferred to be the result of manipulated variables such as case history and not due to internal judgment problems.

Accuracy of categorical judgments is addressed in experimental questions three, four, and five. It was found that undergraduate students had significantly more errors on the normal speech samples. This presumably was due to their clinical inexperience and may also have been influenced by the biasing element of demand characteristics. The likelihood is that undergraduate students came to the experimental situation prepared to listen for problems and when given the alternatives to normal production, these students selected "voice problems" as the alternative. This would seem to indicate either uncertainty over voice problems and/or confidence at this level of training in identifying articulation problems or random error. Given the relatively low selection of "articulation problems," it would appear that uncertainty of voice pathology and perhaps experimental bias are reasons for the number and type of error selections. The fact that undergraduate students, that is, those with the least training and experience, performed poorer is in agreement with the findings of Lass et al. (1973).

Errors on articulation samples are low for all three experimental groups. These results are in agreement with the findings of Morrison (1955) in that relatively naive and

more expert judges rate articulation defectiveness in similar fashion. As will be seen in subsequent discussion, this applies to sensitivity values as well.

Voice problem samples were difficult to determine on Presentation I for all three experimental groups. They were most difficult for working professionsls. Given case history information on the second presentation, all three group's accuracy increased. This may be explained by the fact that the case history statements provided cues sufficient enough to suggest the appropriate category. For example, the following negative case history all but implies the category:

> A classroom teacher from last year expressed concern over how this youngster sounded, however, she indicated reluctance to make any referral since "the mother sounds exactly the same."

These kinds of statements, which were writtento closely approximate school record summaries, appear to have been of most benefit to the professionals from the schools as they showed the greatest improvement as a group across presentations. Wilson (1971), in discussing voice problem cases and the public school clinician, described his experience upon beginning employment as director of a large school district's program in speech:

> Most of the speech clinicians who came to work in the District (St. Louis, County, Mo.) seemed to have minimal preparation in the diagnosis and modification of voice deviations (p. 14).

He rationalized the cause of the problem by discussing training practises:

> Minimal time was spent on diagnosis of voice deviations and even less time on therapeutic procedures. Very often, the therapeutic techniques that were taught involved re-habilitation of the laryngectomized patient and were of little practical value in the public school setting (p. 14).

Knepflar, in Hutchinson et al. (1979), is most direct in providing a rationale for poor performance on judgments of voice samples:

> I believe that voice problems constitute the most over-looked area in the diagnosis of communication disorders and that most training programs for speech pathologists are weaker in the area of voice than any other aspect of the field of communication disorders (p. 206).

An additional rationale has been suggested by Filter (1974):

> Perhaps one of the reasons is that the beginning clinician does not have an approach to voice therapy with which he is comfortable (p. 149).

It is apparent that these authors have been concerned with the level of expertise among speech pathologists dealing with voice problems. It appears from their comments, however, that their concern is directed exclusively at a singular area of the problem; emphasis on voice disorders during the initial training experience. Based on results of the present study it is suggested that this concern needs to be distributed across the totality of professional training and experience. In the present study it was the graduate students near the end of their academic preparation who were most accurate in identifying the voice problem samples and the

professional speech pathologists who were least accurate in the task. These results warrant further attention.

There are several possible explanations for the differences in accuracy performance. One factor which may have affected the performance of the public school speech pathologists in the present study is the length of time since any had participated in formal coursework related to voice disorders. The sample of working professionals in the study had considerable experience, many reporting nine or more years of work experience past their Master's degrees. This longevity may not typify public school speech pathologists as a whole. It also seems likely that when the subjects in the present study were in training less was known or taught concerning identification and remediation of voice problems in children. This does not, however, make the problem less important. On the contrary, it strongly suggest the need for ongoing scrutiny of skills across all areas of speech and language problems by practicing professionals and directed formal study to maintain competency in dealing with these problems. This responsibility for training belongs to both the individual professional and to the employers whom they serve.

Results of the present study suggest that speech pathologists may rely on internal referents for making various qualitative judgments of voice and that there is a need, as diagnosticians, to periodically re-evaluate and re-establish this system of referents. Whether using methods for describing

problems such as the equal-appearing interval system employed in the present study and advocated by Wilson or some other alternative system, it appears critical that some methods be identified, applied and consistently revitalized throughout a professional career.

Considering the difficulty evidenced by public school speech pathologists in the present study it comes as no surprise that fundamental information such as the incidence of voice problems among school children vary considerably. The result is that until a system similar to that advocated is devised, the exact incidence of problem children in the schools will remain unknown and more than likely children who need services of speech pathologists will go unseen.

In addition to the problems of voice pathology, these children also have other problems as outlined by Wilson (1979): higher incidence of conductive hearing loss, otolaryngeal problems, tendency toward agressive behavior and pathological family characteristics. This suggestion of multiple problems amplifies the need to accurately identify children with voice problems in the schools.

In examining group accuracy values between presentations, it appears as though groups were highly consistent; and this lack of "difference" would suggest no effect (bias) of case history information. These results are misleading. Beasley and Manning (1973), in explaining the results of their study, indicated that

> ...investigations of biasing effects upon speech pathologists ordinarily have involved

> group data, and found that, as a group, speech pathologists are not easily biased in a particular direction. However, mean data do not consider the possible bias associated with individual experimenters, and the designs used to date have not adequately lent themselves to such analyses. Thus, what may appear to be random error may, in fact, be subject-based experimenter bias (p. 99).

This appears to be precisely what occurred in the present study. Closer examination of individual accuracies between presentations indicated that for normal samples, voice samples and articulation samples respectively, the undergraduate students had ten, ten and thirteen of their twenty-five members who were accurate in categorical judgments on both presentations. Similar values existed for graduate students (11, 10 and 14) and working professionals (10, 11 and 10). The fact is that less than half of each group were accurate on presentations with the working professionals the least accurate overall. These data suggest that case history information affected accuracy of judgments. Finding susceptibility to biasing statements agrees with the results of Lass et al. (1975) and Wilson and Gasek (1975). In addition, Wilson and Gasek (1975) also found working professionals more subject to biasing conditions than students.

Sensitivity

The issue of sensitivity has been addressed along several dimensions: the proportion of subject groups who changed judgments across presentations, the magnitude of judgment changes across presentations, the directionality of changes as a function of case history type, and the comparison of

average judgments of each group with those of the expert judges. Each of these dimensions will be considered separately.

As noted in the discussion of reliability, all three groups were adept at categorical judgments: however, the present issue related to the ability of each subject to rate severity of speech production problems on a seven-point continuum. In further analyzing responses of subjects to the programmed reliability measures with regard to their severity ratings, it is noteworthy that rating behavior is highly variable within groups. For undergraduate students thirty-four percent of the subjects rated samples designated for replication in an identical fashion. This figure is consistent for graduate students (38%) and working professionals (41%). In other words, approximately sixty-two percent of judgments on samples having speech problems repeated during Presentation I were assigned different severity values by the experimental subjects. This relatively large percentage raises questions concerning the use of equal-appearing interval scales for rating speech behavior. Although previous authors (Morrison, 1955; Sherman and Morrison, 1955; Sherman and Moodie, 1957; and Wilson, 1979) have advocated the use of this form of scaling procedure, it may be necessary to develop guidelines for demonstrable, measurable competencies in scaling as part of the process of training of speech pathologists. Likewise, there would appear to be strong evidence to suggest the need to program for ongoing

maintainence of these competencies once a student leaves academics and enters the work environment as previously discussed. Prather (1960) had suggested the use of an alternative scaling strategy: direct-magnitude estimation which would, among other things, provide ratio-type data. Although she had discussed the fact that differences between equal-appearing intervals and direct-magnitude estimation may not be important, results of the present study suggest that perhaps the limitations she identified may, in fact, be of considerable importance. These so-called inherent weaknesses include an end effect, the failure to remove observer bias and the limitation of interval-type data. Attempts were made to control several of these variables in the selection of samples that expert judges rated consistently between themselves and for whom sensitivity ratings were in the middle of the scaling range. Again, however, the variability in subject's scaling behavior raises questions concerning the nature of the task. Perhaps consideration should be given to training speech pathologists in the use of direct-magnitude estimation strategies for measuring various aspects of speech production.

At the same time perhaps it is not the dependent variable which should be considered exclusively. Sherman and Morrison (1955) indicated that absolute values of severity measures of defective articulation are not necessarily comparable from one individual to another. The point being that depending on the amount of shift between groups, variables like

experience might assume greater responsibility for differences. Likewise, written information might help stabilize the scoring (higher agreement among subjects in a particular group), and in this sense perhaps the term "bias" as presently used should be re-examined to determine whether it is as totally undesirable as is typically suggested. Meitus et al. (1973) were proponents of this notion.

In addition to the possibility of application of alternative psychophysical measuring strategies, the presumed skill level or competency level of the experimental subjects should be questioned. It may be that scaling levels of defectiveness is neither a part of clinical training activity nor professional practice. Since the percentages of subjects in each group providing similar scaled judgments were consistent across groups, it can be assumed that training and/or experience are not directly related to scaling behavior. Since the percentages of subjects presenting similar ratings is low, it must be assumed that other strategies are used by speech parthologists for determining the degree of severity for persons having articulation or voice problems. It would seem appropriate to identify these alternative strategies and explore differences that would exist between groups as a function of training and experience using these approaches.

In reviewing the results, it is not surprising that there was a large percentage of subjects who changed ratings from Presentation I to Presentation II. These results suggest an effect due to the introduction of case history statements.

However, given the same rate of change on measures of reliability confounds the issue. It is remarkable that the dependent variable was as tenuous as evidenced given the presumed nature of training speech pathologists. Based on the present findings, conclusions concerning the scaling of severity through the use of equal-appearing interval scales must be evaluated with caution. These findings tend to support the results of Bradford, Brooks and Shelton (1964) who had reported low levels of reliability among judges of voice (nasality) problems. This caution is further underscored as results of scaling for articulation disorders were as inconsistent as for voice problems.

It may be that subjects in previous studies where equal-appearing interval scales were employed were sufficiently trained in the use of the scaling procedure so as to perform in a highly reliable fashion. A further consideration is the fact that scales of this type may be regarded as highly subjective. In this regard Wilson and Gasek (1975) found their professional and student populations biased when employing subjective measures. Beasley and Manning (1973) had previously cautioned that the more subjective the task, the more susceptible to bias evaluators become.

Regarding the magnitude of changes in rating of severity, significant differences existed between the ratings on Presentation I and Presentation II for undergraduate students on articulation samples given neutral case histories. Graduate students rated articulation and voice samples differently

given negative case history information and voice problems differently given neutral information. These were significant at the .05 level of confidence. Professionals rated articulation samples differently given neutral information.

Given the previous discussion regarding sensitivity differences between samples designed to measure reliability on Presentation I, t-tests for correlated means were performed for the reliability pairs. Results indicated significant differences on both articulation problem samples and voice problem samples for undergraduate students. Working professionals demonstrated significant differences in rating articulation samples from the reliability measurement sequence. The following are proposed rationales for this behavior:

1. Undergraduate students were affected by demand characteristics on reliability measures for Presentation I. These students were highly variable when in conditions without cues, too variable to conclude bias as an exclusive explanation.

2. Graduate students were not as variable given the same listening task and no cues. On Presentation II there were significant differences as a function of case history information, and it may be concluded that there was bias among this group.

3. Working professionals were significantly different between trials of the same speaker on Presentation I. These professionals may be accustomed to evaluating individuals in the presence of more extensive information than was provided. Their increased success on voice samples for Presentation II (given case history information) would tend to support this rationale.

Given this tentative explanation of rater's behavior on Presentation I, there continues to be evidence of bias as a

function of case history for all three groups. This is particularly evident in the case of graduate students. This group did not differ significantly on judgments of severity on Presentation I, and yet three of the six conditions reported in Table 17 contain significant results for this group. Regardless of the direction of the change of ratings, it is apparent that this population reacted to case history statements.

Working professionals demonstrate variance between presentations in rating articulation problems even though it is presumed that they are most familiar with this disorder category as general descriptive information identify caseloads as being composed of as much as 80% articulation cases (Bingham, 1961), although those proportions have shifted in recent years (Van Hattum, 1976). Apparently, cues other than those provided in the present design assist working professionals in the process of diagnosis of articulation problems. Again, the subjective nature of the task may have been somewhat foreign to some of the working professionals who have been in "the field" for a number of years.

Results of statistical analyses concerning the relationship between the direction of change and case history type yielded several significant conditions. These included neutral case history information and undergraduate students and positive case history information and graduate students. In the case of the undergraduate students the precipitating cause is more than likely the demand characteristics of the

experimental situation. This rationale is consistent with the conclusion of Lass et al. (1973). Specifically, the students interpreted information which was of a non-specific type to suggest better functioning in the samples judges. In this instance background information which did not provide cues to speech behavior had a biasing effect on their judgments.

The second significant condition was with graduate students and positive case history information. This population reacted in the opposite direction, giving more severe ratings to the speech samples of all types. It appears that graduate students are actively resistive to case history information, perhaps to the point of biasing themselves totally in a negative direction. This would appear, in part, to coincide with the rationale of Beasely and Manning (1973) that graduate students are more resistent to induced bias.

The graduate students in the present study behaved similarly to those in the Lass et al. experiment in that given biasing information focusing on speech problems, they were biased but less so than their undergraduate counterparts. In this study graduate students also performed more accurately than working professionals.

Other case history and subject conditions also did not reach levels of significant difference. It may be concluded, therefore, that case history information which is directed at speech problems does induce bias. This is in agreement with Lass et al. (1973). However, the notion of directionality,

that is, more negative information would induce more severe ratings, has not been conclusively demonstrated. Neutral information caused significant positive changes among undergraduate students; however, this may have been due to the demand characteristics of the experimental situation.

The final issue in the discussion of sensitivity concerns the comparison of ratings of the expert judges and the experimental groups. Two facts are clear: 1) experimental groups are in close agreement with one another, and 2) these values are generally more severe than those of the expert judges. The strategy of using expert judges to determine a "standard" from which to formulate experimental conditions is not new. Wertz and Mead (1975) report that 24 speech clinicians participating in a rating task using a seven-point scale rated samples of voice problems on an average of 3.79, whereas their panel of three Ph.D. "experts" rated the same samples at 4.0. For articulation cases the judges rated 4.33 and clinicians 4.0. Differences between the Wertz and Mead study and the present investigation are that the subjects in the Wertz and Mead project knew the category to be judged, and their results indicated the experts gave the more severe ratings. In the present study the opposite is true in all but one case. The relatively severe ratings by all experimental groups may be the result of expectations on the part of subjects regarding identification of "problems" (demand characteristics) and/or random factors.

Related Issues

There are several issues related to results of the present study which warrant further discussion. A twofold concern relates to the relatively low accuracy of experimental subjects in identifying voice problems. On one hand is the performance of the subject groups, particularly the working professionals; and on the other is the issue of the use of equal-appearing interval scales for judgments of voice characteristics.

Subjects in the present study appear to have problems similar to those found in the Lass et al. study, demonstrating considerable difficulty in accurately identifying voice problems. In both instances few cues were given under certain conditions and judgments were made primarily from information presented auditorily. Perhaps this was not sufficient for all levels of judgment involved in the experimental task. It would appear sufficient, however, as the typical instructional mode in the training of speech pathologists involves the use of tape recorded samples of vocal pathology to teach voice disorders. Personal experience has shown that many instructors utilize commercially available taped materials (e.g., Aronson; "Psychogenic Voice Disorders"; Wilson and Rice, "A Programmed Approach to Voice Therapy") or their own collection of voice samples (Erickson, 1972; Deal, 1978) for instructional purposes. Apparently this teaching method has some validity as the graduate students in the present study were most effective in accurately

identifying voice problems. Graduate students were also most likely to have had the more recent formal training in voice disorders as undergraduate curricula do not typically involve extensive instruction in this subject area and the sample of working professionals had been employed for time periods which suggested formal coursework in the area had occurred years earlier. The point is that the suggestion of Lass et al. may not totally explain some of the differences in group performance. It is proposed that the number of parameters under investigation is not solely responsible for the problems in judgment, but rather, in the case of the working professionals, it may be the latency between the time of formal training in voice disorders and the present experimental task. This proposal suggests that working professionals are less familiar with voice disorders in children than their graduate and undergraduate student counterparts. This may be due to:

1. Training differences as a function of time and general development of information within the field of speech pathology concerning voice problems in children.

2. Work patterns and conditions which emphasize involvement with populations other than voice problem children.

3. Gradual diminishing of internal referents necessary to make accurate qualitative judgments, presuming these skills were once part of each professional's clinical repitoire.

4. Since the level of accurate judgments for voice problem samples improved considerably on Presentation II, it may be that working professionals have been conditioned to rely on cues other than the actual speech production characteristics demonstrated in order to make accurate judgments.

In analyzing the first proposal it is understandable that changes would come about within a professional discipline over time and only through an ongoing concerted effort would it be possible to remain abreast of research and clinical innovations across the wide variety of areas speech pathologists find themselves dealing. At the same time it may be that the profession as a whole has grossly neglected to apportion the appropriate amount of concern to childhood voice disorders as they may deserve.

Certainly work environments within the category "public schools" vary considerably as do primary responsibilities. However, if the data of Wilson (1979) regarding incidence of childhood voice problems are accurate, it is conceivable that most speech pathologists in the schools will encounter voice cases and that need be prepared to recognize them and program for them.

To the third point, it is proposed that the internal referents which individual clinicians employ to make judgments of normal/non-normal need to be re-evaluated and perhaps re-trained periodically. Since judgments of voice production are qualitative in nature, it is imperative that provisions be made throughout one's professional career to assure that the bases for making qualitative judgment are in tact. This would seem to be even more critical in the case of those professionals who do not see youngsters with these sorts of problems on a regular basis.

The final point addresses the fundamental purpose of the present experiment and was alluded to previously by Beasley and Manning (1973):

> ...speech pathologists should be cautioned to base their diagnoses upon their evaluations, and to minimize possible biasing pre-information. This is particularly important in settings where time and/or administrative policy simply does not permit the speech pathologists to administer a battery of objective speech and language measures. An example of such a setting is the public schools, where caseloads are typically large and time for evaluations short. The speech pathologist is subject to influence by other credible, respected professionals, such as teachers, nurses and social workers regarding the client's level of functioning (p. 100).

Speech pathologists need to consider that most of these "credible others" have been shown to be very poor judges and referral sources for voice problems (Diehl and Stinnett, 1959; Swack and Swack, 1967; Wertz and Mead, 1975). Furthermore, as voice behavior often reflects components of the total personality and psychological well-being of the child it is important that the speech pathologist be able to identify and treat voice problems in children. Wilson (1971) notes several reasons for concern in addition to the presenting voice problem:

1. These children have higher incidence of conductive hearing loss.
2. There are often more otolaryngological problems.
3. Voice problem children have tendencies toward aggressive behavior.
4. Often voice problems are found in conjunction with pathological family characteristics.

Given the results of the present study and those of Lass and his colleagues, it may be that voice disorders cannot be evaluated as effectively as other speech production problems when employing the equal-appearing interval scaling technique. It may be that the qualitative parameters of voice production would be more effectively measured through other psychophysical means.

One potential alternative scaling procedure discussed in the literature is the method of direct-magnitude estimation. Prather (1960) concluded that this method was useful in scaling articulation proficiency. This method has the advantage of providing ratio-type data which is statistically more powerful than equal-appearing intervals can provide. However, it is also more complex to perform and almost impossible to use with only an auditory stimulus and hence may not be a more suitable method since speech pathology training programs frequently rely on tape recorded stimulus materials for training purposes.

At the same time Wilson (1979) continues to advocate the use of seven-point, equal-appearing interval scales. He discusses strategies for their implimentation:

> This can be done through judging types and severity of voice deviations and correlating the ratings with those of other speech pathologists (Wilson and Rice, 1977). Reliability or consistency in rating can be determined by comparing the results of periodic ratings of the same samples. When the ratings of several judges are pooled into one rating, either the mean or median values on the equal-appearing interval scales can be used (p. 66).

The results of the present study do not lend support to either the continued use of equal-appearing interval scales nor to the abandonment of such a notion when evaluating speech samples. Results, particularly with regard to rating of severity, indicate that some method needs to be determined which can be used universally for describing in a quantifiable fashion, the degree or magnitude of involvement of the client. It is premature to suggest that equal-appearing interval scales do not have a place in voice evaluations. Perhaps with the continued application and work of researchers like Wilson, a methodological system will be developed which will be both reliable and functional. It can be stated that based on the findings of the present study, the method of equal-appearing intervals is a relatively easy system to use, requiring little, if any, training.

Implications for Training

The present study employed a repeated measures research design with potentially biasing information being presented prior to the second presentation for each stimulus. This strategy has implications for training sensitivity to potential bias for individuals at all levels of professional preparation and/or practice. In the case of students in training, an exercise of this type might be incorporated into early discussion concerning diagnosis and appraisal of speech and language. The format would allow for identification of relative skill levels in accuracy and sensitivity of judgments

in addition to underscoring the need for objectivity in clinical performance.

In the case of working professionals, it has been demonstrated that need exists for both examination of procedural policies and potential bias as well as training in identification of voice problems. An exercise similar to the one employed in the present study might form the basis for workshops for professionals. Given a format of this sort, persons could address the issue of objectivity in a more or less non-threatening fashion and then discuss various employment demand characteristics. In this manner professional practices and the notion of objectivity could be placed in perspective. Workshops could be given by school districts and/or other employing agencies as part of inservice training activities for professional staff.

The experiences of Lass et al. (1975) suggest that the number of parameters under investigation at any one time should be minimized. In the present study there were three category choices. This number appeared reasonable for categorical judgments, however, general confusion concerning voice disorders suggests that a more rigorous training might first be concentrated on singular disorder areas with binary choice decisions forming the first level of demand. Once questions of normal/non-normal can be answered at a high criterion level for several disorder categories, the process of integrating several categories of problems could be considered. It is proposed that a systemmatic approach to

training various disorder characteristics incorporate notions of potential bias. Further, it appears that this kind of systemmatic training would be worthwhile at all levels of experience.

Another implication for training concerns the issue of scaling various speech behaviors. Results of the present study underscore the need for continued evaluation of scaling procedures as a means of objective measurement for selected aspects of behavior. Wilson (1979) has suggested comparing scaled values for voice disorders and arriving at collective judgments using equal-appearing intervals. This "referent building" among student or professional groups appears to be a worthwhile proposal. Data collected over time regarding these kinds of activities might well be used to shed additional light on the issue of the validity of equal-appearing interval scaling and voice problems. In this regard, if elaborate systems of scoring such as those used in the administration of the Porch Index of Communcative Abilities (Porch, 1967) for aphasic behaviors can be developed and rigidly promoted on a national basis, it would seem possible to develop similar objectives for scaling procedures for articulation and voice. Results of the present study indicate this is particularly necessary for voice problems in children.

A final implication is directed at professional organizations and employers who assume responsibility for identifying needs of members or employees and have as stated goals

improved professional practice. One such professional group is the American Speech-Language and Hearing Association. Given that seventeen of the twenty-five professional subjects in the present study hold Certificates of Clinical Competence from this organization, it is proposed that this body, among others, be made aware of the results of this study and that it seek to develop program activities directed at further development of skills among its members. Likewise, these results have similar implications for public school systems which also need to examine both the work environment of speech pathologists and the skill level of these employees across disorder areas and assist in the ongoing development of professional skills.

Conclusions

The following conclusions have been drawn from the results of the present study:

1. Demand characteristics, that is, the influences of the experimental situation itself, confound the examination of bias.

2. Experimental subject groups appear to be able to identify articulation problems accurately given only auditory information.

3. The accuracy of identification of voice problems was not performed well and is particularly alarming as the highest rate of error was found among professional speech pathologists. Accuracy did increase given case history statements, however, even in the presence of this information the working professionals continued to demonstrate the highest error rate.

4. "Voice problems" was the most frequently used description on selections that were in error thus underscoring the notion that subject groups had serious confusion concerning voice problems.

5. A high rate of change was found for ratings of severity between presentations. Re-evaluation of consistency of rating behavior using the reliability measures of the first presentation raised questions concerning the validity of equal-appearing interval scaling with all subject groups. As a result, conclusions regarding the use of this form of scaling as the appropriate psychophysical method for rating speech behaviors must be guarded.

6. The type of case history information did not consistently influence the direction of change of ratings of severity. Thus, while the strategy suggested by Lass et al. (1975) of using potentially biasing information directed at particular speech problems was successful in inducing bias, there was no definitive correspondence between the type of information (positive, negative, neutral) and the direction of any change.

7. Subject groups collectively varied considerably in the magnitude of ratings of severity from expert judges, presumably as a result of demand characteristics of the experimental task.

Suggestions for Further Research

Based on results and conclusion of the present study, it is recommended that consideration be given for research in the following general areas:

1. Continue examination of scaling procedures which might be applicable to speech behaviors. This would include the method of equal-appearing intervals as well as any other psychophysical scaling method which might prove to be reliable and efficient.

2. Examine further the issue of bias across groups that vary with experience to determine whether various work settings are more disposed to conditions of potential bias and whether various speech or language disorders, by their nature influence clinical pre-determination.

3. Explore methods for systemmatically examining subjectivity and objectivity of students in training and professional practitioners.

APPENDICES

APPENDIX A

JUDGE'S PROTOCOL

Raters:

Thank you again for your willingness to participate in this project. Your task will be to listen to a series of short speech samples and to rate each individual child's speech production characteristics. Each judgment will be scored on an individual response sheet. There are several judgments to be made. Upon completion of a single sample please record:

1. Whether the speech production characteristics were normal for an elementary school child.
2. If they were not, please rate the degree of severity on the 7 point equal-appearing interval scale provided.

Subjects with speech problems have been selected who demonstrate a primary problem of either articulation or voice. Note that the scales progress in degree of severity from left to right in a range from minimal difficulty to severe involvement. In addition to the rating, please respond to the two questions relating to sample adequacy.

INSTRUCTIONS

1. Please be seated and make yourself comfortable.

2. Put a headset on and the investigator will play a short speech segment to allow you to adjust your individual volume control to a comfortable listening level. Indicate when you are ready to begin.

3. Speech samples will be played one at a time. Each will be preceeded by the carrier phrase: "Speaker number ___". Please see that the given sample coincides with eht number given in the upper right hand corner of your response sheet.

4. Following the completion of each individual sample, the recorder will be stopped and sufficient time given for you to respond to the items listed.

5. Following the scoring of the last sample, there are several general format and personal description questions which need to be completed. There is also space for comments.

YOUR HELP IS GREATLY APPRECIATED. ANY COMMENTS OR CRITICISMS WILL LIKEWISE BE HELPFUL. PLEASE DO NOT CONFER OR COMPARE NOTES WHILE SCORING. ARE THERE ANY QUESTIONS?

Speaker Number _____

Please rate the speech production characteristics of the individual speaker on the scales given below.

	YES	NO
Normal Speech Production	☐	☐

If no, rate the degree of severity of the primary speech production problem (one category).

	Mild				Moderate		
Articulation	☐	☐	☐	☐	☐	☐	☐
Voice	☐	☐	☐	☐	☐	☐	☐

	YES	NO
1. Was the quality of the sample adequate for making judgments?	☐	☐
2. In your opinion did this speaker evidence behaviors so unusual that they would be easily recognized on a second presentation thirty minutes later?	☐	☐

GENERAL QUESTIONS

	YES	NO
1. Were the samples		
adequate?	☐	
too short?	☐	
too long?	☐	
2. Was the overall task fatiquing?	☐	☐
3. Do you think the concept of equal-appearing intervals needs to be explained to subjects in the following groups:		
Undergraduate students	☐	☐
Graduate students	☐	☐
Working Speech Pathologists	☐	☐
4. Were the instructions clear?	☐	☐

5. In general, what length of response time do you feel is necessary to complete the scaling task.

0 to 5 sec.	5 to 10 sec.	10 to 20 sec.	20 sec. or more
☐	☐	☐	☐

PERSONAL INFORMATION

1. Highest academic degree:

2. Length of time you have worked in Speech Pathology (past master's, in years).

3. Please list several facts relevant to your experience as a clinician, diagnostician, supervisor or instructor with regard to childhood articulation and voice cases (e.g., taught diagnostics - 10 years; clinical supervisor - 7 years; special professional interest - voice disorders).

APPENDIX B

RANDOMIZATION AND ORDERING PROCEDURES

FILE RANORD.BAS PAGE 1

```
1000 DIM L(84),S$(1)
1010 S$(0)=' '
1020 S$(1)='  '
1030 RANDOMIZE
1040 PRINT 'RANORD: A PROGRAM TO PRINT LISTS OF NUMBERS';
1050 PRINT ' IN RANDOM ORDERS'
1060 GO TO 1080
1070 PRINT 'NUMBER MUST BE >1 AND <86: RE';
1080 PRINT 'ENTER NUMBER OF ELEMENTS IN LIST';
1090 INPUT N
1100 IF N<2 THEN 1070
1110 IF N>85 THEN 1070
1120 GO TO 1140
1130 PRINT 'NUMBER MUST BE > 1: RE';
1140 PRINT 'ENTER NUMBER OF LISTS DESIRED';
1150 INPUT M
1160 IF M<1 THEN 1130
1170 FOR J=1 TO M
1180 PRINT 'LIST #';J
1190 REM SET UP THE LIST
1200 FOR I=0 TO N-1
1210 L(I)=I+1
1220 NEXT I
1230 FOR I=N TO 2 STEP -1
1240 K=INT(RND(X)*I)
1250 T=L(K)
1260 L(K)=L(I-1)
1270 L(I-1)=T
1280 NEXT I
1290 A$=''
1300 FOR I=0 TO N-1
1310 B$=STR$(L(I))
1320 A$=A$&S$(2-LEN(B$))&B$
1330 NEXT I
1340 PRINT A$
1350 NEXT J
1360 END
```

Appendix C

CASE HISTORY STATEMENTS

NORMAL SPEAKERS:

(positive) This child has been described as an excellent student who has a great many activities outside the classroom including sports, scouting, etc. Teachers report the child's family is active as a group in many of these interest areas.

(positive) This child has been described by several individuals as precoscious...having used complete sentences before age 2. The classroom teacher has likewise verified the excellent language skills.

(negative) This child has been described as immature for her age. She is physically a small child, the youngest of six children by five years.

(negative) Teachers report this child is having considerable problems in school. This report coincides with a similar observation from last year's records. In addition, the speech therapist from the reporting school indicated this child was considered for her caseload last Fall.

(neutral) This child is considered an average performer in school. The child is one of seven children who range in age from 3 to 16 years.

(neutral) This child comes from a family who has moved quite frequently. The father is an army officer and as a result the children have seen a great deal of the world, even at their young ages.

VOICE PROBLEM SPEAKERS:

(positive) This child was seen at a famous cleft palate clinic for possible velo-pharyngeal insufficiency. Although the staff speech pathologist was not in the office the day the child was seen, the rest of the staff reported evidence of apparent normal functioning.

CASE HISTORY STATEMENTS (cont.)

VOICE PROBLEM SPEAKERS (cont.):

(positive) Prior to the recent family move persistent laryngitis, secondary to allergy problems was diagnosed by an allergist. A regiment of medication has been administered for the past six weeks for allergy symptoms.

(negative) This child is reported as highly active and excitable. In addition the classroom teacher notes this child "constantly yells while at play".

(negative) A classroom teacher from last year expressed concern over how this youngster sounded however she indicated reluctance to make any referral since "the mother sounds exactly the same".

(neutral) Upon recommendation of last year's teacher the family took this child to an ENT for an examination. The report has not yet been received and the mother did not know the results of the visit.

(neutral) Although the change in cities and schools was seen by the family as a potential problem, the mother reports that this youngster and the other two family children seem to be adjusting adequately.

ARTICULATION PROBLEM SPEAKERS:

(positive) This child has made considerable improvement of articulation skills following major reconstructive surgery, the result of a severe accident.

(positive) This child has shown increasing adaptation to school and is reported as performing adequately in the classroom. The mother indicated that the child was dismissed from speech therapy last year.

(negative) This child's mother reports that their previous school system did not have the services of a speech pathologist. In looking through records it was noted there are no former teacher reports yet either.

CASE HISTORY STATEMENTS (cont.)

ARTICULATION PROBLEM SPEAKERS (cont.):

(negative) Last year's speech clinician reports spending a tremendous amount of time working with this child and the family. Since that time the mother has filed for divorce and moved out of the home with the child.

(neutral) This child is reported as being physically well developed and an excellent young athelete. In the previous school situation the child was considered to be one of the most popular young people at the school by both teachers and students.

(neutral) This child has well developed reading skills, although performs at an average level overall in school. Likewise the youngster is described as physically well coordinated.

APPENDIX D

SAMPLE RESPONSE PACKED INCLUDING INSTRUCTIONS

APPENDIX D.

INSTRUCTIONS: (to be read aloud)

Thank you for your willingness to participate in this study. I want to provide you with an overview of the task and to go through the instructions with you. In order to give the exact instructions to all groups, I will read them to you.

> For today's activity we will assume you are a public school speech pathologist. In a few minutes you will be asked to listen to short segments of children's speech. The voices will be those of children from a school you have responsibility for...all of the samples are of elementary school children. You will be asked to make several judgments concerning what you hear:
>
> 1. Is the speech normal or not
> 2. If it is not, is the primary problem one of articulation or voice and where would it appear on a seven-point equal-appearing interval scale with one being least severe and seven the most severe?
> 3. Lastly, once the primary speech production problem has been identified, what are your diagnostic/prognostic impressions of the child? These will be developed in four short questions which also employ a seven-point equal-appearing interval scale.

All responses will remain totally anonymous. My interest is in seeing how persons with your level of training and background respond to this kind of task. I would ask that you listen carefully and do the best job possible.

```
1000 DIM L(94),S$(1)
1010 S$(0)=' '
1020 S$(1)='  '
1030 RANDOMIZE
1040 PRINT 'RANORD: A PROGRAM TO PRINT LISTS OF NUMBERS';
1050 PRINT ' IN RANDOM ORDERS'
1060 GO TO 1080
1070 PRINT 'NUMBER MUST BE >1 AND <86: RE';
1080 PRINT 'ENTER NUMBER OF ELEMENTS IN LIST';
1090 INPUT N
1100 IF N<2 THEN 1070
1110 IF N>85 THEN 1070
1120 GO TO 1140
1130 PRINT 'NUMBER MUST BE > 1: RE';
1140 PRINT 'ENTER NUMBER OF LISTS DESIRED';
1150 INPUT M
1160 IF M<1 THEN 1130
1170 FOR J=1 TO M
1180 PRINT 'LIST #';J
1190 REM SET UP THE LIST
1200 FOR I=0 TO N-1
1210 L(I)=I+1
1220 NEXT I
1230 FOR I=N TO 2 STEP -1
1240 K=INT(RND(X)*I)
1250 T=L(K)
1260 L(K)=L(I-1)
1270 L(I-1)=T
1280 NEXT I
1290 A$=''
1300 FOR I=0 TO N-1
1310 B$=STR$(L(I))
1320 A$=A$&S$(2-LEN(B$))&B$
1330 NEXT I
1340 PRINT A$
1350 NEXT J
1360 END
```

APPENDIX C

CASE HISTORY STATEMENTS

Appendix C

CASE HISTORY STATEMENTS

NORMAL SPEAKERS:

(positive) This child has been described as an excellent student who has a great many activities outside the classroom including sports, scouting, etc. Teachers report the child's family is active as a group in many of these interest areas.

(positive) This child has been described by several individuals as precoscious...having used complete sentences before age 2. The classroom teacher has likewise verified the excellent language skills.

(negative) This child has been described as immature for her age. She is physically a small child, the youngest of six children by five years.

(negative) Teachers report this child is having considerable problems in school. This report coincides with a similar observation from last year's records. In addition, the speech therapist from the reporting school indicated this child was considered for her caseload last Fall.

(neutral) This child is considered an average performer in school. The child is one of seven children who range in age from 3 to 16 years.

(neutral) This child comes from a family who has moved quite frequently. The father is an army officer and as a result the children have seen a great deal of the world, even at their young ages.

VOICE PROBLEM SPEAKERS:

(positive) This child was seen at a famous cleft palate clinic for possible velo-pharyngeal insufficiency. Although the staff speech pathologist was not in the office the day the child was seen, the rest of the staff reported evidence of apparent normal functioning.

CASE HISTORY STATEMENTS (cont.)

VOICE PROBLEM SPEAKERS (cont.):

(positive) Prior to the recent family move persistent laryngitis, secondary to allergy problems was diagnosed by an allergist. A regiment of medication has been administered for the past six weeks for allergy symptoms.

(negative) This child is reported as highly active and excitable. In addition the classroom teacher notes this child "constantly yells while at play".

(negative) A classroom teacher from last year expressed concern over how this youngster sounded however she indicated reluctance to make any referral since "the mother sounds exactly the same".

(neutral) Upon recommendation of last year's teacher the family took this child to an ENT for an examination. The report has not yet been received and the mother did not know the results of the visit.

(neutral) Although the change in cities and schools was seen by the family as a potential problem, the mother reports that this youngster and the other two family children seem to be adjusting adequately.

ARTICULATION PROBLEM SPEAKERS:

(positive) This child has made considerable improvement of articulation skills following major reconstructive surgery, the result of a severe accident.

(positive) This child has shown increasing adaptation to school and is reported as performing adequately in the classroom. The mother indicated that the child was dismissed from speech therapy last year.

(negative) This child's mother reports that their previous school system did not have the services of a speech pathologist. In looking through records it was noted there are no former teacher reports yet either.

CASE HISTORY STATEMENTS (cont.)

ARTICULATION PROBLEM SPEAKERS (cont.):

(negative) Last year's speech clinician reports spending a tremendous amount of time working with this child and the family. Since that time the mother has filed for divorce and moved out of the home with the child.

(neutral) This child is reported as being physically well developed and an excellent young athelete. In the previous school situation the child was considered to be one of the most popular young people at the school by both teachers and students.

(neutral) This child has well developed reading skills, although performs at an average level overall in school. Likewise the youngster is described as physically well coordinated.

APPENDIX D

SAMPLE RESPONSE PACKED INCLUDING INSTRUCTIONS

APPENDIX D.

INSTRUCTIONS: (to be read aloud)

Thank you for your willingness to participate in this study. I want to provide you with an overview of the task and to go through the instructions with you. In order to give the exact instructions to all groups, I will read them to you.

> For today's activity we will assume you are a public school speech pathologist. In a few minutes you will be asked to listen to short segments of children's speech. The voices will be those of children from a school you have responsibility for...all of the samples are of elementary school children. You will be asked to make several judgments concerning what you hear:
>
> 1. Is the speech normal or not
> 2. If it is not, is the primary problem one of articulation or voice and where would it appear on a seven-point equal-appearing interval scale with one being least severe and seven the most severe?
> 3. Lastly, once the primary speech production problem has been identified, what are your diagnostic/prognostic impressions of the child? These will be developed in four short questions which also employ a seven-point equal-appearing interval scale.

All responses will remain totally anonymous. My interest is in seeing how persons with your level of training and background respond to this kind of task. I would ask that you listen carefully and do the best job possible.

INSTRUCTIONS (cont.)

Here are your response packets (distribute). Please read the introduction section. Note the purpose and description.

Next, please fill in the general information section. Pencils are available for all responding. The last item in the general information section, you'll note, refers to known hearing loss. What is implied is that once we adjust the headphones for volume, if you have a hearing loss and there isn't sufficient intensity or if there is too much distortion for you to make adequate judgments, you will be excused from participation. The intent is to have good judgments and the limits of the equipment must be recognized.

After filling out page one completely and reading the introduction section, pull off the back sheet of this packet. This is the consent and release form. Please read it carefully, sign it and I will collect them.

I will be playing a tape for you which contains samples of children's speech. Each child is responding to the Sounds in Sentences sub-test of the Goldman-Fristoe Sound Test of Articulation...the story of Jack and Ricky. Please turn to the second sheet of the packet and we'll read through the script for that sub-test. (Read aloud) The purpose for reading this is so that you can listen to the production of each child rather than being concerned with what is being said.

INSTRUCTIONS (cont.)

Turn to the first response sheet, the third page in your packet. At the very top of the response sheet is the question of whether or not the production is normal. Please indicate yes or no. If you judge the speech production characteristics to be abnormal, determine the degree of severity of the primary problem and circle that designated number on the equal-appearing interval scale. Note the two areas of speech production problems are articulation and voice.

Following that decision, the four questions on the lower half of the sheet refer to diagnostic/prognostic impressions of the child and again you should circle the best number according to an equal-appearing interval scale.

The number in the upper right hand corner should correspond to the number of the sample indicated on the tape. If not, please bring it to my attention immediately.

The tape samples are forty-five to sixty seconds long. You can proceed to make judgments at any time during the sample or following it. Since people will vary in response time I will control the tape as is necessary. When everyone has completed judgments we can proceed to the next sample. If judgments are made as the speaker comes to the end of the passage I will let the tape run to the next sample. There are approximately two seconds between the end of one sample and the beginning of the next if the tape is allowed to run continuously. Each speech sample is preceeded by the phrase: "Speaker number ___".

INSTRUCTIONS (cont.)

Return to the first page of the response packet. Please pencil in the number I give you which will serve to identify this packet with the second one we'll be doing. This is the only form of identification that will be used and again, it is simply to match packets. Please do not write your name on any of the materials.

One comment on scoring. In order to arrive at the best estimate of everyone's judgments I need for you to make each entirely on your own...please do not consult your neighbor. Likewise, I would ask that once we go through a sample and you have marked a score, please leave it. Also, following the exercise this morning/afternoon, I would appreciate it if you wouldn't discuss the task with others in the program who will be participating in order to preserve their naivete.

The final step before going into the experimental task will be to put the headphones on and adjust for appropriate volume. Before that, are there any questions? If not, you can put the headphones on and adjust the volume control found on the blue box in the center of the table. Please find a comfortable volume setting. The beginning of this tape has a portion of the passage "My Grandfather" during which you can adjust things...if that isn't sufficient, let me know. (Begin tape.)

SUBJECT CONSENT AND RELEASE FORM

I, _____ hereby agree to participate in the study being conducted. I understand my task will be to listen to short speech samples and rate the subjects' performance on an equal-appearing interval scale. I understand that throughout the duration of this study, I will remain completely anonymous and have, as my option, the privilege of withdrawing from participation at any time without penalty. I have read this statement and, agreeing to its contents, hereby give my permission for the experimenter to use data collected from me.

_____
Signed

_____
Date

I. INTRODUCTION

This study is concerned with judgements of children's speech samples. As a participant you will be asked to listen to a number of short samples and to rate your diagnostic/prognostic impressions of the youngster on the tape. You will listen to these samples under headphones which you will adjust to a most comfortable listening level. Your judgements remain totally anonymous at all times.

II. GENERAL INFORMATION

Please fill in the general information section, but DO NOT SIGN THIS FORM. The testor will read the criteria for participation. If you do not meet these criteria, please indicate this immediately.

PLEASE INDICATE WHETHER YOU HAVE SUCCESSFULLY COMPLETED THE FOLLOWING COURSES:

ASC (or the equivalent)	#108	Yes	No
	222	Yes	No
	274	Yes	No
	276	Yes	No
	277	Yes	No
	372	Yes	No
	373	Yes	No
Do you have a known hearing loss?		Yes	No

III. CONSENT AND RELEASE FORM

Next, pull off the back sheet of this packet. It is a Consent and Release form. Please read it carefully, sign and date it. Today's date is _____. When you have completed the Consent and Release form please pass it in to the testor.

IV. INSTRUCTIONS

The testor will now read the instructions for the task. Please listen carefully and ask questions if the instructions are not clear.

V. SCRIPT FOR STIMULUS TASK

Please read the script and look up when you've finished.

I. INTRODUCTION

This study is concerned with judgments of children's speech samples. As a participant you will be asked to listen to a number of short samples and to rate your diagnostic/prognostic impressions of the youngster on the tape. You will listen to these samples under headphones which you will adjust to a most comfortable listening level. Your judgments remain totally anonymous at all times.

II. GENERAL INFORMATION

Please fill in the general information section, but DO NOT SIGN THIS FORM. The testor will read the criteria for participation. If you do not meet these criteria, please indicate this immediately.

Any Questions?

Please fill in this section now.

Highest academic degree:	BA( )	BS( )	MA( )	MS( )	Other( )
Presently enrolled in graduate training in speech pathology:	YES( )		NO( )		
Successfully completed either 20 semester or 30 term hours at graduate level:	YES( )		NO( )		
Successfully completed minimum of 50% of clinical hours required for degree:	YES( )		NO( )		
Previous work experience:	NONE( )		1 YEAR( )		
Do you have a known hearing problem:	YES( )		NO( )		

III. CONSENT AND RELEASE FORM

Next, pull off the back sheet of this packet. It is a Consent and Release form. Please read it carefully, sign and date it. Today's date is_____. When you have completed the Consent and Release Form please pass it in to the testor.

IV. INSTRUCTIONS

The testor will now read the instructions for the task. Please listen carefully and ask questions if the instructions are not clear.

V. SCRIPT FOR STIMULUS TASK

Please read the script and look up when you've finished.

I. INTRODUCTION

This study is concerned with judgements of children's speech samples. As a participant you will be asked to listen to a number of short samples and to rate your diagnostic/prognostic impressions of the youngster on the tape. You will listen to these samples under headphones which you will adjust to a most comfortable listening level. Your judgements remain totally anonymous at all times.

II. GENERAL INFORMATION

Please fill in the general information section, but DO NOT SIGN THIS FORM. The testor will read the criteria for participation. If you do not meet these criteria, please indicate this immediately.

Please fill in this section now.

Highest academic degree: MA( ) MS( ) PHD( ) OTHER( )

Public School Work Setting: YES( ) NO( )

Years of experience in the schools: 3-5( ) 6-8( ) 9 OR MORE( )

A.S.H.A. Certification: YES( ) NO( )

Do you have a known hearing problem: YES( ) NO( )

III. CONSENT AND RELEASE FORM

Next, pull off the back sheet of this packet. It is a Consent and Release form. Please read it carefully, sign and date it. Today's date is _____. When you have completed the Consent and Release form please pass it in to the testor.

IV. INSTRUCTIONS

The testor will now read the instructions for the task. Please listen carefully and ask questions if the instructions are not clear.

V. SCRIPT FOR STIMULUS TASK

Please read the script and look up when you've finished.

Speaker Number _____

Please rate the speech production characteristics of the individual speaker on the scales given below.

	YES	NO
Normal Speech Production	( )	( )

If no, rate the degree of severity of the primary speech production problem (one category).

Articulation	least severe	1	2	3	4	5	6	7	most severe
Voice	least severe	1	2	3	4	5	6	7	most severe

PLEASE MARK YOUR RESPONSES TO THE FOLLOWING STATEMENTS AT ONE OF THE NUMBERED POINTS ON EACH LINE.

This child is in need of speech services:	1 strongly agree	2	3 agree	4	5 disagree	6	7 strongly disagree
If therapy is recommended, the prognosis for the first year of therapy would be:	1 very good	2	3 good	4	5 poor	6	7 very poor
If therapy is not recommended, the prognosis for improvement in speech during the year would be:	1 very good	2	3 good	4	5 poor	6	7 very poor
I would expect this child to be a difficult case to work with.	1 strongly agree	2	3 agree	4	5 disagree	6	7 strongly disagree

JACK AND RICKY

Jack and Ricky should be in school.
Instead they are going fishing.
Ricky is in such a rush that he drops his glasses,
and gets his shirt caught in the zipper of his jacket.

They fish from the old bridge.
All of a sudden they hear a loud noise.
Oh! Its only the dog chasing a squirrel.

Jack and Ricky catch thirteen fish.
1...2...3...4...5...6...7...8...9...10...11...12...13.
They laugh because they are very, very, very happy.

They think that no one will catch them.
They sneak back and hide under the house.
Oh, no! Jack's mother finds them.

VI. HEADPHONE ADJUSTMENT, FINAL QUESTIONS, ETC.

Turn to response sheet #1 and wait for the tape to begin.

PLEASE RATE THESE SAMPLES ON YOUR OWN!

INSTRUCTIONS (cont.)

Following the presentation of the first tape a short break was announced. During this time an "optional" questionnaire was distributed and people were asked to consider filling it out. The following instructions preceeded the second tape.

This second section is a bit shorter than the first. Once again you are a clinician in the schools and the next group of youngsters are transfers into your building this Fall. Again, you are being asked to make judgments about their primary speech production problem on an equal-appearing interval scale. At the top of the response sheet is a short statement about the child which has been "lifted" from the accompanying school records. Again, please note the number in the upper right hand corner and see that it corresponds to the number of the taped sample. Are there any questions? If not, let's proceed.

Dear (Undergraduate) Students:

While taking a break between tapes I would like to ask for your opinion (anonymous, of course) regarding aspects of training. Since I am involved in a training program in Pennsylvania, I am interested in students' perceptions of their needs. Of particular interest is the area of clinical training. If you would, I would appreciate general comments to the following questions. Again, your responses will be of benefit in planning undergraduate training activities.

1. Based on your experience, do you believe practicum training should be offered on the undergraduate level? Why or why not...

2. In your training was your academic preparation sufficient for your initial clinical experience?

3. Do you feel comfortable with your "mechanical" skills at this point (mechanical implies objective preparation, plan writing, behavioral management, etc.)?

   YES ( ) NO ( )

   If no, which would you like more information about?

GRAD STUDENTS,

It occurred to me yesterday after class that since the task you're involved in required a changing of tapes,etc., that part of the time between might be spent responding to a few general questions regarding the program in ASC. As we have discussed in class, as the graduate student representative to the faculty, I'd like to provide Dr. Deal with our collective impressions of the training program after we leave. In this fashion I believe we can congratulate and reinforce positive aspects of the program and identify and underscore what we believe to be areas for concern in the program. The ultimate goal is to make certain our program continues to grow and maintain a good reputation. After all, they'll be referring to me as "Flahive from MSU" for a long time and I want to have come from the best... I believe a few minutes to give an honest appraisal will help the faculty and staff here in doing just that.

Generally there are three areas I've designated for comments:

academic
clinical
personal

These are in no way exhaustive. The following choice questions are intended to get at general information and to provide stimulus for comments. If you have specific items you'd like to include but that are longer than a line or two, feel free to jot them down and deposit them in my mail box. I'd like to use short statements in the letter I'll draft to Dr. Deal. Please do not sign this or any other data you give me. I'll generate the letter in mid-September and so anyone wanting a copy should sign the list Dave Snyder has with an address and I'll be happy to forward a copy...otherwise we can meet at an ASHA party sometime and I'll be glad to go over what is written!!!

If you do not care to generate anything, please feel free to avoid ...

PROGRAM COMMENTS: Please be brief and sincere

ACADEMIC:

All pre-employment paranoia aside...are you prepared fundamentally to function as a speech pathologist?

What were the strongest and weakest classes you had...but it dosen't do any good unless there's a reason you perceived it that way!!! In other words, how can the best stay good and the weaker get better.

If you were to make improvements in the academic offerings, what would you do? (this includes the two-year issue, keeping or changing staff assignments, etc.)

General Comments:

CLINICAL COMMENTS:

Where you adequately supervised during your practicum experiences, given your perception of the load supervisors have to deal with? What is your perception of their job...are they overworked, is the ratio a good one, etc.?

Were your experiences varied? Did you have exposures which were representative of the disorder groups you will work with someday? What kinds of things would you maintain and change if you were responsible for the clinical training program?

How were the off-campus supervisors...this is not intended to be a name-calling or praising section...general comments about the quality of the off-campus people should be sufficient (unless there is a real need to express +)

How would you rate your clinical skills? (On a seven-point equal-appearing interval scale!!! ...I participated in the study too!)

Personal Comments : This section is to make comments of a general nature and to note the kinds of interactions you've had with the Departmental staff...the secretaries and significant others with whom we all interact during the course of training. As a consumer, how would you rate your treatment? Again, comments of both a positive and negative sort are encouraged...and again, name-calling, etc. is not intended ...without trying to interject anything to influence your comments, I thought this section would allow for feedback to that component of the program which is often not acknowledged...

Dear Professional:

While taking a break between tapes I would like to ask for your opinion (anonymous, of course) regarding training needs or refresher needs you might have relative to speech apthology. With the ongoing generation of information in our profession it sometimes seems difficult to stay on top of everything. If you'd care to reflect on the few questions below, I would be interested in knowing what needs are present, if any, since I am involved in a training program myself. Your responses are totally anonymous and will be collected separate from the response packets. I am simply interested in getting a handle on what public school clinicians see as training needs.

1. Are there areas of professional preparation you would like to have "refresher" information about?

   YES ( ) NO ( )

   If yes, do these areas pertain to present responsibilities?

   YES ( ) NO ( )

   Elaborate on one or two of these.

2. What in your experience, is the best vehicle for receiving this kind of information if one is a working school speech pathologist:

   district or intermediate district in-service ( )
   local speech and hearing groups ( )
   state speech and hearing conventions ( )
   national speech and hearing conventions ( )
   other (specify) ( )

Professional Letter (cont.)

3. Do you have ideas about viable means for post-degree information dissemination?

Appendix D.

This short release form has been used to secure permission from parents of children whose voices were used in the development of the general stimulus tape. The child's name was not secured. The only identifying information asked was the age and sex of the youngster. Each child was given the option of participating in addition to the signed permission. Likewise the child was assured that he/she could withdraw from participation at any time.

SPEECH SAMPLE CONSENT FORM

Michigan State University Speech and Hearing Clinic is hereby authorized to use for educational, scientific, and professional purposes the photographs or audiotapes taken of me or my minor child_____ on_____.

Signed_____

Witnessed by_____

APPENDIX E

RAW DATA AND PERTINENT TABLES AND FIGURES

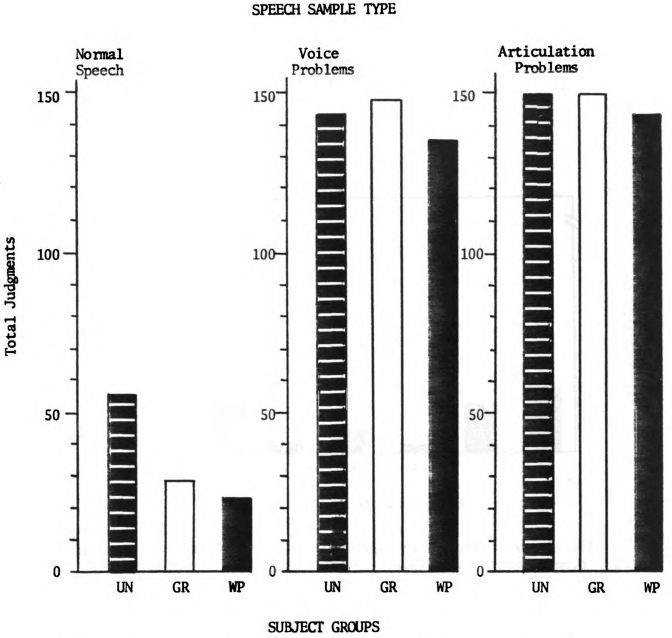

Figure 1. Judgments of non-normal speech behavior by experimental subject groups for each speech sample type on Presentation I.

Key:

UN - Undergraduate students
GR - Graduate students
WP - Working professionals

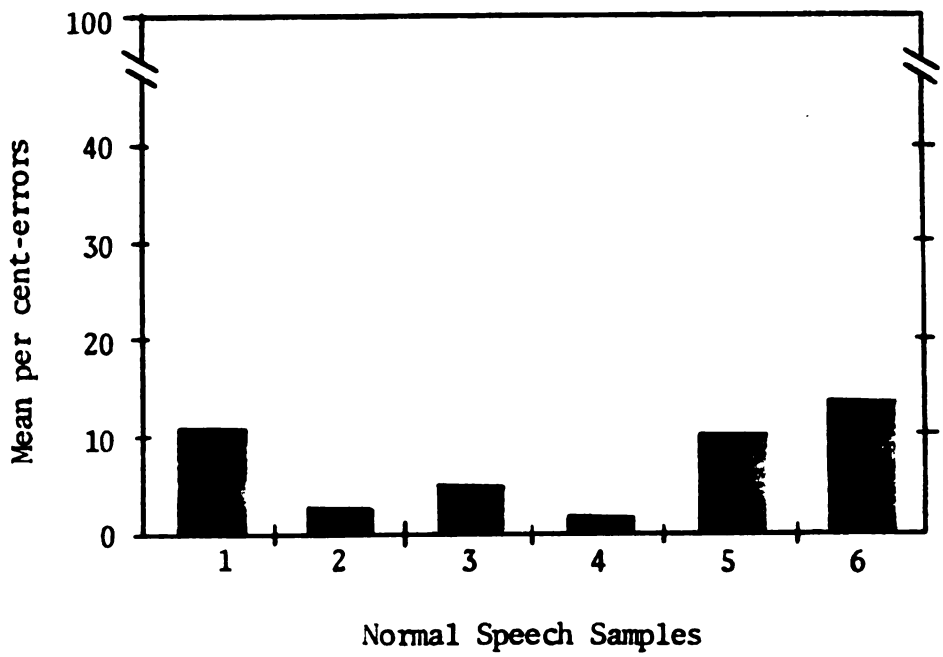

Figure 3. Mean per cent-errors on categorical judgments across all experimental subjects for each individual normal speech sample on Presentation I.

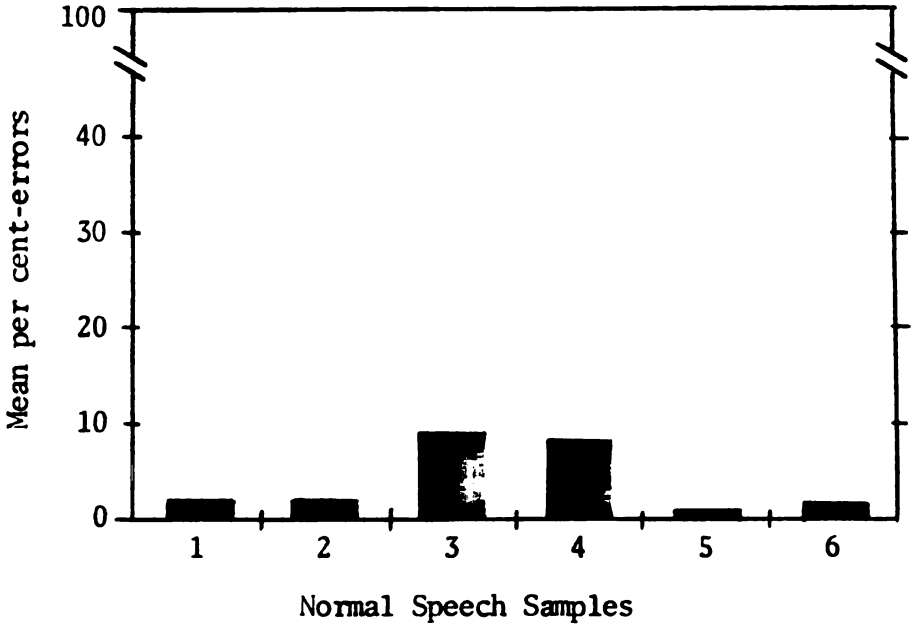

Figure 4. Mean per cent-errors on categorical judgments across all experimental subjects for each individual articulation problem speech sample on Presentation I.

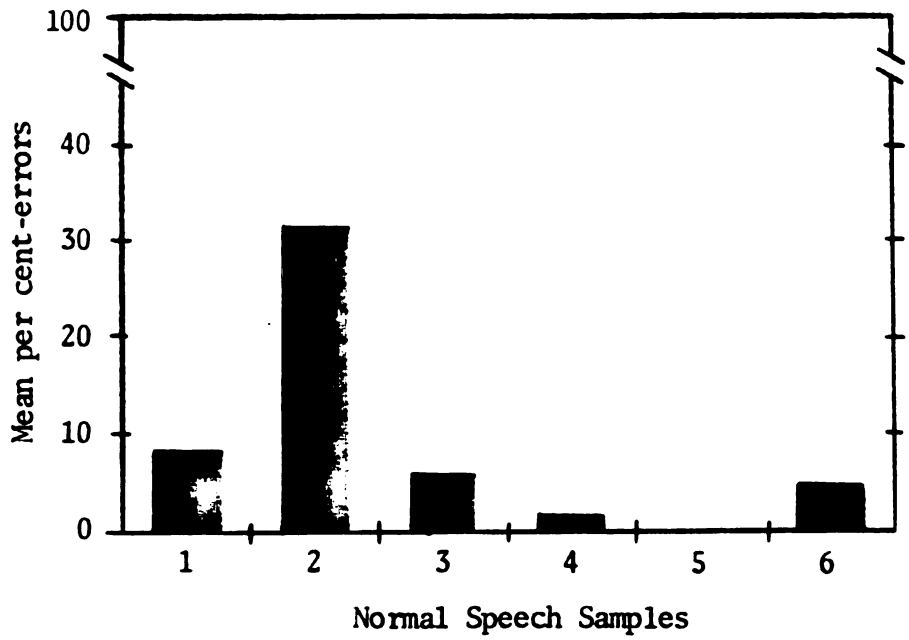

Figure 5. Mean per cent-errors on categorical judgments across all experimental subjects for each individual voice problem speech sample on Presentation I.

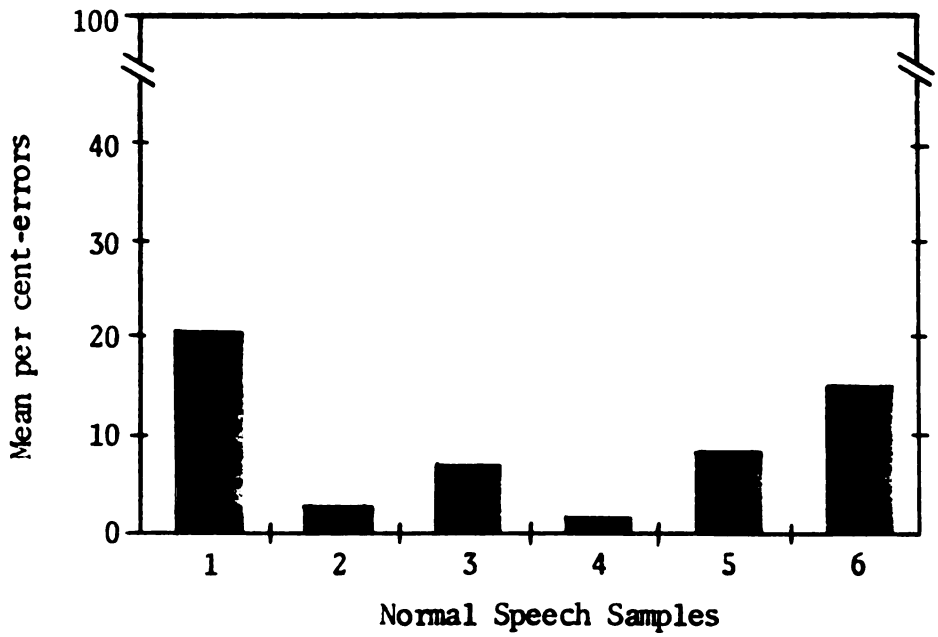

Figure 6. Mean per cent-errors on categorical judgments across all experimental subjects for each individual normal speech sample on Presentation II.

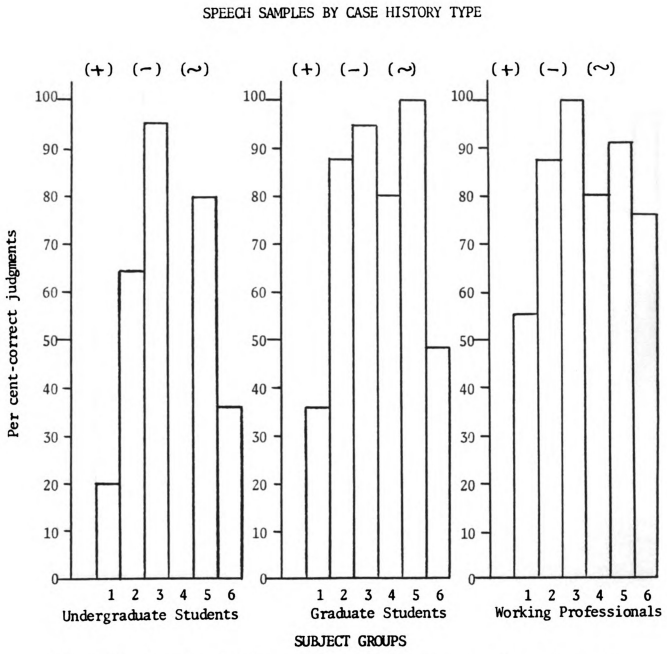

Figure 12. Per cent-correct judgments for the six normal speech samples of Presentation II. Data are grouped according to case history type.

Case History Type Key:

(+) - Positive history
(−) - Negative history
(~) - Neutral history

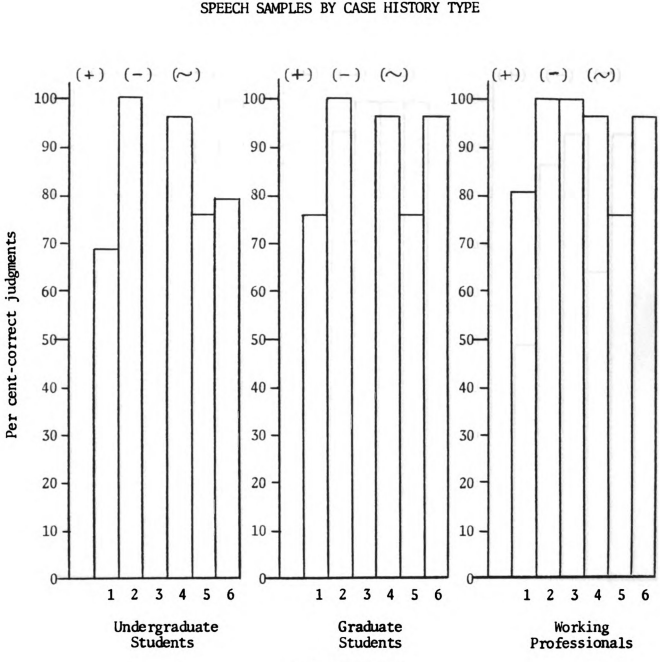

Figure 13. Per cent-correct judgments for the six articulation problem samples of Presentation II. Data are grouped according to case history type.

Case History Type Key:

(+) - Positive history
(−) - Negative history
(~) - Neutral history

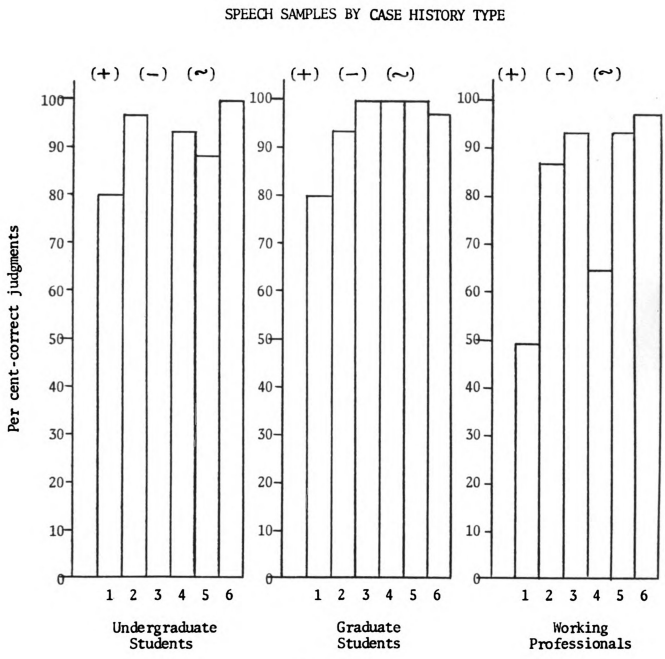

Figure 14. Per cent-correct judgments for the six voice problem samples of Presentation II. Data are grouped according to case history type.

Case History Type Key:
(+) - Positive history
(−) - Negative history
(~) - Neutral history

REFERENCES

REFERENCES

1. Ad Hoc Committee on Ethical Standards, American Psychological Association, Ethical Principles in the Conduct of Research with Human Participants. Washington: American Psychological Association, (1973).

2. Auffrey, J. J. Jr., The physical attractiveness of mentally retarded program candidates as a determinant of evaluation by professionals of varying training and experience. Unpublished doctoral dissertation, Michigan State University, (1975).

3. Barber, T. X. and Silver, M. J., Fact, fiction and the experimenter bias effect. Psych. Bull. Monog. Supp., Vol. 70, No. 6, part 2, 1-29 (1968).

4. Beasley, D. S. and Manning, J. I., Experimenter bias and speech pathologists' evaluation of children's language skills, J. Comm. Dis., 6, 99-101, (1973).

5. Bradford, L. J., Brooks, A. R., and Shelton, R. L. Clinical judgments of hypernasality in cleft palate children. Cleft Palate J., 1, 329-335, (1964).

6. Brunning, J. L. and Kintz, B. L., Computational handbook of statistics. Glenview, IL: Scott, Foresman and Co., (1968).

7. Chamberlin, T. C., The multiple working hypothesis. J. of Geology, 5, 837, (1897).

8. Cook, D. L., The impact of the Hawthorne effect in experimental design in educational research, Cooperative Research Project, #1757, Washington, DC: U.S. Office of Education (1967).

9. Deal, L. V., Personal Communication, (1978).

10. Diehl, C. F. and Stinnett, C. D., Efficiency of teacher referrals in a school speech testing program. J. Speech Hear. Dis., 24, 34-36, (1959).

11. Elashoff, J. and Snow, R. E., Pygmallion Reconsidered. Worthington, OH: Charles A. Jones Publishing Co., (1971).

12. Elliot, L., Hirsh, I. and Simmons, A., Language of young hearing-impaired children. Lang. and Speech, 10, 141-158, (1967).

13. Erickson, R. L., Personal communication, (1972).

14. Filter, M., Propreceptive-tactile-kinesthetic feedback in voice therapy. Lang. Speech, Hear. Ser. Schools, 5, 149-151, (1974).

15. Flahive, M. J. and Magistro, M., "Examiner Bias in a Population of Working Speech Pathologists". Paper presented at the Fall Conference, Michigan Speech and Hearing Association, October, (1974).

16. Friedman, N., Kurland, D. and Rosenthal, R., Experimenter behavior as an unintended determinant of experimental results. J. of Proj. Tech. Person. Asses., 29, 479-490, (1965).

17. Gephart, W. J. and Antonoplos, D. P., The effects of expectancy and other research-biasing factors. Phi Delta Kappan, June, 579-583, (1969).

18. Hutchinson, B. B., Hanson, M. L. and Mecham, M. J., Diagnostic Handbook of Speech Pathology. Baltimore, MD: Williams and Wilkins Co., 206-239, (1979).

19. Johnson, W., Spriestersbach, D. C. and Darley, F. L., Diagnostic Methods in Speech Pathology. New York: Harper and Row Publishers, (1963).

20. Lass, N. J., Browning, K. N. and Brown, D. M., Clinician bias: the effects of pretesting information on the evaluations of speech clinicians, J. Comm. Dis., 8, 105-113, (1975).

21. Linton, M. and Gallo, P., Jr., The Practical Statistician: Simplified handbook of Statistics. Monterey, CA: Brook/Cole Publishing Co., (1975).

22. Meitus, I. J., Ringel, R. L., House, A. S. and Hotchkiss, J. C., Clinician bias in evaluating speech proficiency, Br. J. Dis. Comm., (8)2, 146-151, (1973).

23. Morrison, S., Measuring the severity of articulation defectiveness. J. Speech Hear. Dis., 20, 347-351, (1955).

24. Naremore, R. C. and Hipskind, N. M., Responses to the Language of educable mentally retarded and normal children: stereotypes and judgments. Lang. Speech Hear. Ser. Schools, 10, 27-34, (1979).

25. Noll, J. D., Articulation assessment. In J. E. Fricke (ed.), Speech and the dentofacial complex: the state of the art. Washington, DC: American Speech and Hearing Association, 283-298, (1970).

26. Orne, M. T., On the social psychology of the psychological experiment with particular reference to demand characteristics and their implications. Amer. Psychol., 17, 776-783, (1962).

27. Plutchik, R., Foundations of Experimental Research (2nd Ed.). New York: Harper and Row Publishers Inc., (1974).

28. Porch, B. E., Porch Index of Communicative Ability. Volume 2: (Revised Edition) Administration, Scoring and Interpretation. Palo Alto, CA: Consulting Psychologist, (1973).

29. Prather, E. M., Scaling defectiveness of articulation by direct magnitude-estimation. J. Speech Hear. Res., 3, 380-392, (1960).

30. Rothlishberger, F. J. and Dickson, W. J., Management and the Worker, Cambridge, MA: Harvard University Press, (1939).

31. Rosenthal, R. and Jacobson, L., Teacher expectancies: determinants of pupils' I.Q. gains. Psychological Reports, 19, 115-118, (1966).

32. Schulz, R., Heller, J. C., Gens, G. W. and Lewin, M., Pharyngeal flap surgery and voice quality factors related to success and failure. Cleft Palate J., 10, 166-175, (1973).

33. Sherman, D. and Moodie, C. E., Four psychological scaling methods applied to articulation defectiveness. J. Speech Hear. Dis., 22, 698-706, (1957).

34. Sherman, D. and Morrison, S., Reliability of individual ratings of severity of defective articulation, J. Speech Hear. Dis., 20, 352-358, (1955).

35. Swack, J. W. and Swack, M. J., Efficiency of teacher referral of children with speech deviations. J. Mich. Speech Hear. Assoc., 3, 47-52, (1967).

36. Thurstone, L. L. and Chave, E. J., The measurement of attitude. Chicago: University of Chicago Press, (1929).

37. Van Hattum, R. J., Services of the Speech Clinician in schools: Progress and prospects. Amer. Speech Hear. Assoc., 59-63, (1976).

38. Wells, F. L., A statistical study of literary merit. Archives of Psychol., 16, (1907).

39. Wertz, R. T. and Mead, M. D., Classroom teacher and speech clinician severity ratings of different speech disorders. Lang. Speech Hear. Ser. Schools, 6, 119-124, (1975).

40. Wilson, D. K., Voice Problems of Children (second edition). Baltimore: The Williams and Wilkins Co., (1979).

41. Wilson, F. B., The voice-disordered child: A descriptive approach. Lang. Speech Hear. Ser. Schools, 1, 14-22, (1971).

42. Wilson, F. B. and Rice, M., A programmed approach to voice therapy. Austin, TX: Learning Concepts, (1977).

43. Wilson, W. R. and Gasek, G., The influence of pre-information on the rating of articulation. J. of Comm. Dis., 8, 15-22, (1975).

44. Yount, M. A. and Downs, T. D., Testing the significance of the agreement among observers. J. Speech Hear. Res., 11, 5-17, (1968).