1975

Effects of expectations on the evaluation behaviors of speech clinicians

Kenneth Hammer Rue

The University of Montana

Let us know how access to this document benefits you.
Follow this and additional works at: https://scholarworks.umt.edu/etd

Recommended Citation
Rue, Kenneth Hammer, "Effects of expectations on the evaluation behaviors of speech clinicians" (1975). Graduate Student Theses, Dissertations, & Professional Papers. 7751.
https://scholarworks.umt.edu/etd/7751

This Thesis is brought to you for free and open access by the Graduate School at ScholarWorks at University of Montana. It has been accepted for inclusion in Graduate Student Theses, Dissertations, & Professional Papers by an authorized administrator of ScholarWorks at University of Montana. For more information, please contact scholarworks@mso.umt.edu.
THE EFFECTS OF EXPECTATIONS
ON THE EVALUATION BEHAVIORS OF SPEECH CLINICIANS

by
Kenneth Hammer Rue
B.S., Moorhead State College, 1972

Presented in partial fulfillment of the requirements for the degree of
Master of Arts
UNIVERSITY OF MONTANA
1975

Approved by:
Chairman, Board of Examiners
Dean, Graduate School

Date Dec 10, 1975
ABSTRACT

Rue, Kenneth H. and Disorders Communication Sciences

The Effects of Expectations on the Evaluative Behaviors of Speech Clinicians

Director: Dr. Evan Jordan

This study investigates the effects of expectations on speech clinicians' behaviors in the evaluation of nasality in tape recorded voice samples. Nineteen female graduate students in Speech Pathology and Audiology were chosen as subject (raters) according to criteria that provided trained sources of expectations. They rated their perceived nasality in two sequences of 40 backward-played, audio-taped voice samples of normal third and fourth grade boys. In the first sequence, 20 of the samples were randomly assigned the label "cleft palate" with the remaining 20 labeled "normal." A second sequence consisted of the same 40 samples in a re-randomized order with the sample labels reversed from those of the first sequence.

A 2X2X19 analysis of variance and a Scheffe test indicated that speech clinicians can be biased by the expectation effect in their evaluation of voice samples. The analyses indicated that within each of the two sequences and in some comparisons across both sequences the clinicians, as a group, rated samples labeled "cleft palate" more nasal than samples labeled, "normal." The analysis of variance and the Scheffe test produced no supporting data for the second experimental hypothesis: rating differences according to label would decrease from the first to the second sequence. Instead, the analyses showed that the label differences were greater in the second sequence than the first. A trend analysis indicated a linear increase in the scale values regardless of sample label throughout the experiment. This linear trend may have been the result of operation of an extraneous variable(s): loss of reality based referents and consequently resulted in enhancement of the label effect and implied directions to find nasality. Differences in the performance of raters were analyzed for significance in the analysis of variance in order to isolate this source of variability from the analysis of the effects of labels and sequences. These results revealed a significant difference in the performance of individual raters. Additional analyses showed that most raters were affected by labels, some more than others, and a few raters were influenced very little or not at all by the expectation effect.
ACKNOWLEDGMENTS

Many people associated with the University of Montana, Moorhead State University, Missoula Public Schools and Great Falls Public Schools were responsible for the origin and completion of this thesis. To appropriately acknowledge all individuals involved would require additional volumes, however, several key people must be mentioned. Thesis committee members Miss Cynthia Harrell, Dr. Kelly Lyndes, and Dr. Duane Pederson allowed me to begin and complete the project. Dr. Evan Jordan, chairman of the committee, provided the time, support and direction needed throughout the project. Mrs. Darlyne Behan, the 3rd and 4th grade (1973) boys of Lewis and Clark School, and the administration of Missoula Public Schools all made possible the voice samples which were crucial for the study. Special acknowledgment is deserved by: Dr. George Camp, Great Falls Public Schools, for the statistical help and psychological support at crucial times; my wife, Lu, and daughter, Jenni, for their remarkable patience and for making it all worthwhile. Finally, I would like to thank my golf pro.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABSTRACT</td>
<td>ii</td>
</tr>
<tr>
<td>ACKNOWLEDGMENTS</td>
<td>iii</td>
</tr>
<tr>
<td>LIST OF TABLES</td>
<td>vi</td>
</tr>
<tr>
<td>LIST OF FIGURES</td>
<td>vii</td>
</tr>
<tr>
<td>Chapter I. INTRODUCTION</td>
<td>1</td>
</tr>
<tr>
<td>Expectation Effects</td>
<td>2</td>
</tr>
<tr>
<td>Summary of Expectation Effects</td>
<td>5</td>
</tr>
<tr>
<td>Rating Nasality</td>
<td>6</td>
</tr>
<tr>
<td>Summary</td>
<td>8</td>
</tr>
<tr>
<td>Chapter II. PROCEDURE</td>
<td>9</td>
</tr>
<tr>
<td>Experimental Hypotheses</td>
<td>9</td>
</tr>
<tr>
<td>Null Hypotheses</td>
<td>9</td>
</tr>
<tr>
<td>Raters</td>
<td>10</td>
</tr>
<tr>
<td>Voice Sample Selection</td>
<td>10</td>
</tr>
<tr>
<td>Voice Sample Recording</td>
<td>11</td>
</tr>
<tr>
<td>Disguise Techniques</td>
<td>12</td>
</tr>
<tr>
<td>Training-Practice Session</td>
<td>12</td>
</tr>
<tr>
<td>Master Tape</td>
<td>12</td>
</tr>
<tr>
<td>Presentation of Samples to Raters</td>
<td>13</td>
</tr>
<tr>
<td>Score Sheets</td>
<td>14</td>
</tr>
<tr>
<td>Scaling</td>
<td>14</td>
</tr>
<tr>
<td>Chapter</td>
<td>Page</td>
</tr>
<tr>
<td>-------------------------------------------</td>
<td>------</td>
</tr>
<tr>
<td>III. RESULTS</td>
<td>16</td>
</tr>
<tr>
<td>Effect of Labels</td>
<td>16</td>
</tr>
<tr>
<td>Effect of Sequences</td>
<td>19</td>
</tr>
<tr>
<td>Effect of Raters</td>
<td>22</td>
</tr>
<tr>
<td>Summary</td>
<td>22</td>
</tr>
<tr>
<td>IV. DISCUSSION</td>
<td>26</td>
</tr>
<tr>
<td>V. SUMMARY AND CONCLUSIONS</td>
<td>33</td>
</tr>
<tr>
<td>LITERATURE CITED</td>
<td>37</td>
</tr>
<tr>
<td>APPENDIX I</td>
<td>39</td>
</tr>
<tr>
<td>APPENDIX II</td>
<td>40</td>
</tr>
<tr>
<td>APPENDIX III</td>
<td>41</td>
</tr>
<tr>
<td>APPENDIX IV</td>
<td>42</td>
</tr>
<tr>
<td>APPENDIX V</td>
<td>43</td>
</tr>
</tbody>
</table>
### LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>I.</td>
<td>Summary of Analysis of Variance for the Two Experimental Conditions: Labels and Sequences</td>
<td>16</td>
</tr>
<tr>
<td>II.</td>
<td>Scheffe Test of Multiple Comparisons of the Sums of Scores in Four Conditions: Effects of Two Labels in Two Sequences</td>
<td>18</td>
</tr>
<tr>
<td>III.</td>
<td>Trend Analysis Grouping</td>
<td>21</td>
</tr>
<tr>
<td>IV.</td>
<td>Mean Scale Scores for the Five Trend Analysis Groups</td>
<td>21</td>
</tr>
<tr>
<td>V.</td>
<td>Summary of Trend Analysis</td>
<td>21</td>
</tr>
<tr>
<td>VI.</td>
<td>Summary of Sign Test Data for Individual Raters</td>
<td>24</td>
</tr>
</tbody>
</table>
LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>I. Graph of LXS Interaction</td>
<td>20</td>
</tr>
<tr>
<td>II. Graph of LXR Interaction</td>
<td>23</td>
</tr>
</tbody>
</table>

vii

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
Chapter 1

INTRODUCTION

According to Dicke, et al. (1970): "The three main sources of variation and error in the psychiatric diagnostic interview are the interviewer, the person being interviewed, the process involved in the interview." Certainly these three components may be used to represent the whole of any particular communicative, evaluative situation. The labeling of the triad varies from one discipline to another: Rhetoric is concerned with the speaker, message, and audience (listener); education is comprised of the teacher, method, and student; and speech pathology and audiology feature the clinician (or audiologist), procedures, and client. With the exception of rhetorical research concerning the nature of audiences and individual listeners, research dealing with the three main evaluative components has focused on the processes (methods) and the speaker (evaluatee). This is particularly true of clinical evaluation in speech pathology. Although reliability studies of clinical judgments have been conducted, the primary emphasis in the field has been investigation of clinical methodology and the normal and abnormal communication processes of the speaker. As Seigel (1966) reminds us: "What we mean by pathological or disordered speech rests ultimately on an evaluation. This evaluation is mediated by a listener, and the study of speech pathology invites the investigation of listening as well as speaking." Other disciplines, especially education, have begun to look closely at the evaluator's performance as influenced by expectations. It was the purpose of this study to determine if expectations influence the evaluation behaviors of speech clinicians.
Expectation Effects

Evaluator expectation research may be defined by a three part process: (1) sources of expectation, (2) performance of the evaluator, and (3) performance of the evaluatee. The expectation effect, also known as the self-fulfilling prophecy, was originally defined by Merton (1957): "...confident error generates its own spurious confirmation..." Rosenthal (1969) has refined the definition by hypothesizing the existence of two types of expectation effects. The first occurs as expectations bias the performance of the evaluator. Rosenthal described this effect as occurring in "the eye, the hand, and the brain" of the evaluator; only the behavior of the evaluator is altered. Secondly, in its turn, the altered behavior of the evaluator alters the behavior of the evaluatee. An interaction between the expectation-biased evaluator and the evaluatee influences the evaluatee's performance, or as stated by Rosenthal, "this interaction actually can alter the responses or data that are obtained." So that in any particular evaluative situation these hypothesized effects may be seen as functioning as a sequential process: Expectations (from some source) influence the evaluator, the evaluator's performance is biased because of the expectations, and the evaluator's biased behavior subsequently alters the behavior of the evaluatee.

Rosenthal (1968) initially examined the expectation process using teachers and students of an elementary school in San Francisco, California. At the beginning of the school year eighteen classrooms of children were given an I.Q. test, and twenty percent of the children from each class were randomly selected for the experimental group (designated as "intellectual bloomers"). The teachers were told that "the intellectual bloomers had scored high on a test for intellectual blooming and would
show remarkable gains in intellectual development" during the school year. At the end of the school year the children's I.Q.'s were remeasured, as well as the teacher's attitudes toward the children. The results supported the Rosenthal-hypothesized process of expectation effects. The false expectations generated by the experimenter were reflected in the performance of teachers as evaluators of the students. The teachers evaluated the intellectual bloomers as "more appealing, better adjusted, and more affectionate, with less need for social approval" than the other children. The expectation-influenced teachers also altered the performance of the students (evaluatees): The intellectual bloomers averaged gains of four more verbal I.Q. points than the other children. Other studies (Rosenthal, 1968, and 1969; Miller, 1972) in various educational settings and experimental situations have produced significant results, expectations of the evaluator affecting the performance of the evaluatee, and, in turn, the performance of the evaluatee.

Rosenthal research has been the subject of some controversy as to the confirmation of the final stage of the expectation sequence, the altered performance of the evaluatee. Several researchers (Barber, et al., 1969; Claiborn, 1969; Fleming and Antonen, 1971) have attempted to replicate several Rosenthal studies and have not found significant results in the performance of the evaluatees. However, Rosenthal (1969) questioned the procedures and analysis of these attempted replications: He noted that Barber, et al. used evaluators whose number, sex ratio, socioeconomic status, and education differed from those used by Rosenthal, and that Barber used analysis of variance computations with a small N(13) whereas Rosenthal used non-parametric analysis. Miller
(1972) concluded that Barber has shown "that the expectancy effect is difficult to replicate" in the performance of the evaluatee, but that "there are data to indicate that the phenomenon is indeed obtainable." Confirmation of the final stage of the expectation process is important to the implications of the present study, even though this study investigated only the effect of expectations on the performance of the evaluator.

There is, apparently, no controversy concerning the replicability of evaluator-biasing experiments. Even Fleming and Antonen (1971), although failing to find significant differences in the performance of evaluatees, did find that the evaluator's (teachers) attitudes did change significantly and positively toward the intellectual bloomers. Additional research in education (Cahen, 1966; Brophy and Good, 1970; and Meichenbaum, et al., 1969) has supported the Rosenthal evidence of the effects of expectations on evaluator performance.

Rosenthal and others have investigated possible sources of expectations in evaluators. Two general sources have been identified: 1) "Natural" occurring attitudes, learned informally through participation in a given life situation. 2) Attitudes learned formally through academic training and/or professional experience. Investigating the first area, several researchers (Rosenthal, 1969; Dusek, 1972; and Entwisle, et al., 1972) found several differential "natural" occurring attitudes associated with the sex, status, race, or age-group of the evaluators. Rosenthal (1969) found that adult evaluators spend more time in data collection from subjects of the opposite sex. Several investigators (Rosenthal, 1969; and Barber et al., 1969)
studied the second source and found that evaluators were influenced differentially by expectations learned from their professional training and/or experiences. Palardy (1969) found that teachers who believed boys were less successful readers than girls produced students accordingly, and teachers who believed that both sexes were equally successful produced students with no significant differences between the sexes in reading achievement. Finally, attitudes from these two sources are "triggered" by specific cues or information (IQ scores, sex, etc.) presented to the evaluator at the immediate time of the evaluation and resulting in expectation-influenced evaluations.

Speech clinicians are exposed to both sources of expectations. Our obvious differences in sex, age, etc. may result in differential "natural" attitudes in clinicians. Attitudes are generated in training and experience to expect certain signs when evaluating an individual with particular anomaly (with cleft palate children we look for facial scars, hypernasality, nasal emission, poor articulation, etc.). Attitudes from these two potential sources can be triggered by specific information about each client, such as labels (cleft palate, aphasia, etc.) and descriptions (case histories). Certainly if these conditions result in expectation-biased evaluations then a modification of our evaluation behaviors is in order.

**Summary of Expectation Effect**

The purpose of this study was to examine speech clinicians as evaluators in the diagnostic process. Other disciplines have provided evidence that evaluators may produce biased evaluations because of expectation effects. Expectation effects have been found by Rosenthal.
to function as a process whereby expectations of the evaluator altered by evaluator expectations, was less conclusive than the findings of measurable expectations of the evaluator; however, the present study tested only the hypothesis that expectations alter the behavior of the evaluator. Two sources of expectations may be conceptualized: 1) Those attitudes which occur naturally in all evaluators because of differences in sex, age, etc. 2) Those attitudes learned from training or previous evaluation experiences. Finally, the effect of these sources is triggered by specific information (labels, description, etc.) about the evaluatee at the time of evaluation. It seems probable that expectations could bias the evaluation performance of a speech clinician. Speech clinicians bring sources of expectation, such as sex differences and educational/professional experiences, to the evaluative situation and are exposed to specific information, such as labeling and description, at the time of the evaluation.

In investigating the possibility of expectations affecting the evaluative performances of speech clinicians, the following specific question was chosen for this study: Does a speech clinician expect more nasality in the voice of a speaker labeled "cleft palate", and, consequently, rate the speaker as more nasal in voice quality than when the same speaker is labeled "normal?"

Such a question called for an examination of studies of the nasality rating behavior of speech clinicians.

Rating Nasality

Several investigators found that speech clinicians-in-training at the graduate level can rate nasality reliably in audio tape-recorded voice samples with minimal practice (Spriesterbach and Powers, 1959;
Lintz and Sherman, 1961; and Counihan and Cullinan, 1970 and 1972). Counihan and Cullinan (1972) presented two samples of the most nasal voice quality and two samples of the least nasal voices to be rated, followed by ten practice items prior to the experimental ratings. The experimental ratings revealed intrajudge correlation coefficients of from .88 to .95 and a mean correlation coefficient of .92. Counihan and Cullinan (1970) found that the type of voice sample in audio-only presentation was a function of reliability ratings of judges; connected speech was rated more reliably than isolated vowels or syllables. Interjudge reliability ratings were about the same for connected speech played forward or backward (correlation coefficients of .96 and .94, respectively). Two investigators (Sherman, 1954; and Spriesterbach, 1955) felt backward-play eliminated contextual cues (articulation abilities, etc.) of the voice samples which may influence the rating of nasality.

When additional rating cues (visual) were provided, clinicians' ratings of nasality were unreliable. Bradford, Brooks, and Shelton (1964) attempted to typify clinical procedures by providing individual evaluation sessions for each of nine cleft palate children. Eight judges (4 graduate students and 4 post-graduate clinicians) spent five minutes in face-to-face evaluation of each child. The intrajudge correlation coefficients ranged from .14 to .33 with interjudge ratings ranging from .41 to .67.

Two types of scaling methods have been used in the rating of nasality: equal-appearing intervals and direct magnitude estimation (Counihan and Cullinan, 1970). However, Stevens (1966) noted that the unconstrained method of direct magnitude estimation offers the most
accurate means of measuring psychophysical judgments. According to Stevens, "The accumulated evidence suggests, however, that whenever possible the observer is best left free to choose his own modulus."

Summary

There is considerable evidence to indicate that speech clinicians may be biased by naturally occurring and training/experience expectations in the evaluation of persons with speech disorders. One specific instance of the effect could be the evaluation of nasality in the voice of a speaker labeled as having a cleft palate as opposed to the rating of a speaker labeled as being normal. Investigators found that clinicians can rate nasality in audio recordings of connected-speech played backwards, reliably, with minimal practice. Direct magnitude estimation is an accurate means of rating nasality as a psychophysical judgment. Therefore, differential scale values of nasality of voice samples rated under two different labels (cleft palate and normal) should provide a measurement of the effects of expectations on clinicians' psychophysical judgments.
Chapter II

PROCEDURE

In general, the procedure used was similar to the methodology used in several investigations of the rating of nasality in voice samples (Spriesterback and Powers, 1959; Lintz and Sherman, 1961; and Counihan and Cullinan, 1972). Adaptations of procedure in the present study were designed to trigger expectation effects.

Nineteen female raters scaled the perceived nasality in backward-played, audio-taped, seven to twelve second voice samples from forty, normal third and fourth-grade boys. Each sample was introduced for rating as that of a cleft palate child in one sequence and that of a normal child in another sequence. The label presentation of the two sequences was exactly counterbalanced, half of the samples were labeled "cleft palate" first and half were labeled "normal" in the first sequence. The labels were reversed for each sample for the second sequence.

**Experimental Hypotheses:**

1) Speech clinicians will perceive more nasality in backward-played, audio-tape recordings of speakers labeled as "cleft palate" than when the same speakers are labeled as "normal."

2) The labels "cleft palate" and "normal" will have a greater influence on the scaled nasality of the forty experimental voice samples when presented for the first time than when presented for the second time (second sequence).

**Null Hypotheses:**

There will be no significant difference in the scaled values of nasality of voice samples labeled as "cleft palate" and the scaled values of nasality of the same voice samples labeled as "normal" when rated by speech clinicians.
Raters

Nineteen female graduate students in the University of Montana Department of Speech Pathology and Audiology were selected from the clinicians possessing the following qualifications:

1) The completion of SPA 432, Organics I (cleft palate) course.

2) The completion of at least one course in the diagnostic and appraisal of speech disorders.

3) The completion of at least one quarter of clinical practicum.

All female raters were used to control a possible "natural" source of expectations, sex of the rater; no other "natural" sources were identified or controlled. The three qualifying criteria were chosen to yield raters with educational and experiential sources of expectations about the general level of nasality in the voices of cleft palate speakers and the relative absence of nasality in the voices of normal speakers.

Voice Sample Selection

A group of forty third- and fourth-grade boys were selected on the bases of the following criteria as judged by the experimenter prior to each recording:

1) All of the boys had normal speech mechanisms.

2) None of the boys had a denasal voice quality; the range of nasal voice quality extended from normal to one or two voices with mildly clinically significant amounts of nasality.

3) All of the boys had normal speaking rates when reading the stimulus material.

1Two of the raters were senior undergraduates meeting all other qualifications.
This educational level was chosen because the children could be expected to read the stimulus material without difficulty and still fall within the ages when physical rehabilitative work is carried out with cleft palate children. Only males were used in order to control a likely sex effect on ratings: a vocal intensity interaction with speaker sex to produce differential scale values of perceived nasality (Counihan and Cullinan, 1972). Each boy was given a visual examination of the hard and soft palate and other visible oral structures by the experimenter to determine normality of each speaker's speech mechanism. Vocal nasality (lack of it) was controlled to prevent reverse effects found by Rosenthal (1969) when the evaluator is presented with bogus information which is "too-obvious" (in this case, a child with denasal voice quality labeled "cleft palate"). Speaking rate was a selection criterion because it enabled general equating of the voice samples with respect to duration.

**Voice Sample Recording**

The voice sample stimulus material read by each boy was:

"I have just twenty-nine cents. But who needs a million dollars: I like people and it's a wonderful day!" (New Reading Skill Builder, 1966).

This material was selected to conform, approximately, to the usual relative frequency of occurrence of English phonemes (French, Carter, and Koenig, 1929), and to yield voice samples of, approximately, 10 seconds.

Each boy read the material for familiarization. He then read it for recording until he read the material correctly in about 10 seconds. The recording was produced on a Uher Model 1107 (frequency response: 20 to 20k Hz) tape recorder. The microphone was placed approximately eight
inches from the boys' mouths and the experimenter maintained an adequate VU level from sample to sample and during each sample recording.

**Disguise Techniques**

Backward-play presentation to the raters was utilized to aid in disguising the contextual cues (articulation skills, etc.) of each sample and to conceal the fact that each sample was presented twice. Additional disguise features were the number of samples (40) and the random ordering of each of the two presentations (sequences) of the forty samples.

**Training-Practice Session**

The session followed the procedure outlined by Counihan and Cullinan (1972). Four samples representative of the experimental range of hypernasality of the 40 voice samples (as judged by a faculty member and E) were scaled by the raters. An additional 10 practice samples were randomly selected from the 40 experimental samples and were rated following the range samples. The purpose of the session was to insure reliability of ratings by familiarization of the range of judgments and the scaling task.

**Master Tape**

The master audio tape was approximately 33 minutes long and contained the following recorded segments:

1) Instructions for the range scaling task (See Appendix I).

2) Four voice samples representative of the nasality range of the experimental samples (each sample separated from one another by approximately 10 seconds of silence).

3) Instructions for ten additional practice-rating samples (See Appendix II).
4) Ten voice samples were randomly selected from the forty experimental samples; each sample was separated by about 10 seconds with sample number spoken by E about three seconds before each sample.

5) Instructions (containing intended "trigger" information) for the rating of the experimental samples (See Appendix III).

6) Eighty voice samples for experimental judgments. Forty samples were randomized and re-recorded backwards, then the same forty samples were rerecorded and once again recorded backwards on the master tape following the first forty. Each sample was separated by about ten seconds of silence except for a recorded number and label (voice of E) about three seconds before each sample in the first presentation of the forty samples. The same procedure was followed for the taping of the second presentation (sequence) of the same, rerandomized samples, except that each sample's label was changed. That is, a sample labeled as "cleft palate" in the first sequence was labeled as "normal" in the second sequence and vice versa.

7) A final statement by E (See Appendix IV).

The instructions to the raters were structured in an attempt to create an expectation effect. A definition and acoustic referents for the term "hyponasality" (used in the directions to the raters) were not provided by E in order to use the raters' educationally derived definitions and referents of hyponasality.

Presentation of Samples to Raters

The nineteen raters listened, individually, to the master tape presented in a sound treated room (audiology testing suite). The master tape was played from an adjoining room by E and fed to the raters' room at an intensity level of 65dB SPL free field. The level was a comfortable listening level but below intensity levels likely to produce differential ratings favoring the perception of nasality in open vowels (Counihan and Cullinan, 1972).

The instructions for the range scaling task were replayed for
one rater when she indicated that she did not understand the task. Four other raters asked if they should score the four "range" samples on the score sheet (following the range scaling instructions) and the experimenter responded with "yes" in all four instances. These were the only questions by the raters and the only responses given by the experimenter during the experiment.

**Score Sheets**

Each rater's score sheet was composed of four 8½" x 11" pages stapled together in the upper left corner. The numbers 1 through 94 were typed and double-spaced in vertical rows one inch from the left margins (requiring 3½ pages). The letters c and n were typed to the left of each experimental sample number (15 through 94), c corresponding to the samples labeled as "cleft palate" on the master tape and n corresponding to those samples labeled as "normal." The score sheets contained no other markings except for the title Score Sheet typed at the top of the first page of each score sheet (see sample in Appendix V).

**Scaling**

Each rater expressed her rating of the nasality of each sample by drawing a horizontal line next to the sample number on the score sheets (See Appendix I for line drawing instructions). When all raters had completed the experiment, each line was measured in millimeters. Then each line value was converted to a fraction with the denominator being the range value for the particular rater and the numerator being the length of each line minus the particular rater's smallest line value. Therefore, each rater's scale values were expressed as some fraction of
the total range used by the rater; the values ranged from 0 to 1 for each rater. Each fraction was converted to a decimal value (ratio) rounded to the third place (thousandths). For example, rater #2 had a longest line of 39 millimeters and a shortest of 2 millimeters, equaling a range of 37 millimeters. She drew a 6 millimeter line for the first experimental sample; the rater's shortest line value (2) was subtracted from 6 yielding 4. This value was placed over the range value of 37 to produce the fractional value of \( \frac{4}{37} \) to represent the rater's judgment of the first experimental sample. Converting the ratio, \( \frac{4}{37} \), to a decimal produced the value of .108; the same procedure was followed for all experimental sample ratings (80) for all raters (19). These ratio values were then used as the scores for the statistical analysis.
Chapter III

RESULTS

Effect of Labels

The scores were analyzed using a 2X2X19 analysis of variance (Edwards, 1965). The results of this analysis are summarized in Table I. As Table I reveals, all main effects (Labels, Sequences, and Raters) and two-factor interactions were significant at the .01 probability level; the three-factor interaction was not significant.

TABLE I

<table>
<thead>
<tr>
<th>Source of Variance</th>
<th>Sum of Squares</th>
<th>df</th>
<th>Mean Squares</th>
<th>F</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Labels</td>
<td>4.491</td>
<td>1</td>
<td>4.491</td>
<td>106.93</td>
<td>.01</td>
</tr>
<tr>
<td>Sequences</td>
<td>3.728</td>
<td>1</td>
<td>3.728</td>
<td>88.76</td>
<td>.01</td>
</tr>
<tr>
<td>Raters</td>
<td>8.503</td>
<td>18</td>
<td>.472</td>
<td>11.02</td>
<td>.01</td>
</tr>
<tr>
<td>LXS</td>
<td>.666</td>
<td>1</td>
<td>.666</td>
<td>15.86</td>
<td>.01</td>
</tr>
<tr>
<td>LXR</td>
<td>2.376</td>
<td>18</td>
<td>.132</td>
<td>3.14</td>
<td>.01</td>
</tr>
<tr>
<td>SXR</td>
<td>5.983</td>
<td>18</td>
<td>.332</td>
<td>7.90</td>
<td>.01</td>
</tr>
<tr>
<td>LXSXR</td>
<td>.500</td>
<td>18</td>
<td>.028</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Error Within</td>
<td>61.186</td>
<td>1444</td>
<td>.042</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Total                  |                | 1519|

A Scheffé test for multiple comparisons described by Edwards (1965) was used to analyze the simple effects of Labels and Sequences,
as shown in Table II. For the purposes of this analysis, the two sequences of the experimental design were separated into four segments; "cleft palate"-first sequence (cpl), "normal"-first sequence (nl), "cleft palate"-second sequence (cp2), and "normal"-second sequence (n2). In single segment comparisons of the four segments, three of four possible comparisons between label conditions showed that the samples labeled "cleft palate" were rated significantly more nasal than the samples labeled "normal." As shown by the comparison of 1 vs. 2, cpl was rated significantly more nasal than nl; in comparison 3 vs. 4, cp2 was significantly greater than n2; and, finally, the 2 vs. 3 comparison showed cp2 as significantly greater than n2. In multi-segment comparisons, the doubled value of cpl was significantly greater than nl+n2, and the doubled value of cp2 was significantly greater than n1+n2.

The Scheffe results indicate that samples labeled "cleft palate" had a significantly higher mean nasality rating than samples labeled "normal" within sequences (different samples, i.e. half of the experimental samples vs. the remaining half of the experimental samples) and in all but two comparisons across sequences (same samples, i.e. first and second presentations of samples with labels reversed in the second presentation). The segment cpl included the same backward-played samples as n2, and nl contained the same samples as cp2. Comparing cpl to nl is comparing the ratings of randomly labeled samples, and comparing cp2 to n2 is comparing a repeated presentation of the first sequence with the sample labels reversed.

Samples labeled "cleft palate" in the first sequence (cpl) were not rated significantly more nasal than samples labeled "normal" in the
### TABLE II

**SCHEFFE TEST OF MULTIPLE COMPARISONS OF THE SUMS OF SCORES IN FOUR CONDITIONS. EFFECTS OF TWO LABELS IN TWO SEQUENCES.**

<table>
<thead>
<tr>
<th>Comparisons</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>( \sum_{i=2}^4 )</th>
<th>( D )</th>
<th>( D^2 )</th>
<th>Sum of Squares</th>
<th>( F )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 vs. 2</td>
<td>1</td>
<td>-1</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>33.35</td>
<td>1112.75</td>
<td>1.464</td>
<td>34.857</td>
</tr>
<tr>
<td>1 vs. 3</td>
<td>1</td>
<td>0</td>
<td>-1</td>
<td>0</td>
<td>2</td>
<td>-45.59</td>
<td>2079.17</td>
<td>2.735</td>
<td>65.119</td>
</tr>
<tr>
<td>1 vs. 4</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>-1</td>
<td>2</td>
<td>3.66</td>
<td>13.46</td>
<td>.017</td>
<td></td>
</tr>
<tr>
<td>2 vs. 3</td>
<td>0</td>
<td>1</td>
<td>-1</td>
<td>0</td>
<td>2</td>
<td>-78.95</td>
<td>6234.05</td>
<td>8.202</td>
<td>195.285</td>
</tr>
<tr>
<td>2 vs. 4</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>-1</td>
<td>2</td>
<td>-29.68</td>
<td>881.43</td>
<td>1.159</td>
<td>27.595</td>
</tr>
<tr>
<td>3 vs. 4</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>-1</td>
<td>2</td>
<td>49.26</td>
<td>2427.23</td>
<td>3.193</td>
<td>76.023</td>
</tr>
<tr>
<td>1 vs. 2+4</td>
<td>2</td>
<td>-1</td>
<td>0</td>
<td>-1</td>
<td>6</td>
<td>37.02</td>
<td>1370.99</td>
<td>.601</td>
<td>14.309</td>
</tr>
<tr>
<td>2 vs. 1+3</td>
<td>-1</td>
<td>2</td>
<td>-1</td>
<td>0</td>
<td>6</td>
<td>-112.31</td>
<td>12614.43</td>
<td>5.532</td>
<td>131.714</td>
</tr>
<tr>
<td>3 vs. 2+4</td>
<td>0</td>
<td>-1</td>
<td>2</td>
<td>-1</td>
<td>6</td>
<td>128.22</td>
<td>16441.13</td>
<td>7.211</td>
<td>171.690</td>
</tr>
<tr>
<td>4 vs. 1+3</td>
<td>-1</td>
<td>0</td>
<td>-1</td>
<td>2</td>
<td>6</td>
<td>-52.93</td>
<td>2802.22</td>
<td>1.229</td>
<td>29.261</td>
</tr>
<tr>
<td>1+2 vs. 3+4</td>
<td>1</td>
<td>1</td>
<td>-1</td>
<td>-1</td>
<td>4</td>
<td>-75.28</td>
<td>5668.13</td>
<td>3.729</td>
<td>88.785</td>
</tr>
<tr>
<td>1+3 vs. 2+4</td>
<td>1</td>
<td>-1</td>
<td>1</td>
<td>-1</td>
<td>4</td>
<td>82.62</td>
<td>6826.89</td>
<td>4.491</td>
<td>106.928</td>
</tr>
<tr>
<td>1+4 vs. 2+3</td>
<td>1</td>
<td>-1</td>
<td>-1</td>
<td>1</td>
<td>4</td>
<td>-15.90</td>
<td>253.09</td>
<td>.166</td>
<td></td>
</tr>
</tbody>
</table>

All \( F \) listed (greater than \( F^* = 7.80 \)) significant at \( P .05 \).

\*\( \sum_{cpl} \): sum of all "cleft palate" samples in first sequence.

\*\( \sum_{n1} \): sum of all "normal" samples in first sequence.

\*\( \sum_{cpl2} \): sum of all "cleft palate" samples in the second sequence.

\*\( \sum_{n2} \): sum of all "normal" samples in the second sequence.
second sequence, and the mean rating of both label conditions were greater in the second sequence than their respective labels in the first sequence. Thus, a trend effect is suggested and will be analyzed in considering the effect of sequences.

**Effect of Sequence**

The second experimental hypothesis was that the effect of labels (expectation effect) would be less in the second sequence than in the first sequence. Two statistical findings concerning sequence differences are revealed by the analysis of variance, the Scheffe test, and visual examination of the interaction figures: 1) The difference between the mean scale values of samples labeled "cleft palate" and the mean scale values of samples labeled "normal" was not less, but rather greater in the second sequence than the first. Therefore, the second experimental hypothesis must be rejected since label (expectation) effects did not decrease over prolonged use.

2) The mean scale values of both label conditions were greater in the second sequence than the mean scale values of the same label conditions in the first sequence. That is, samples labeled "cleft palate" in the second sequence were rated as significantly more nasal than samples labeled "cleft palate" in the first sequence and the same result occurred for samples labeled "normal." These results and visual examination of the LXS graph indicated the possibility of a continuous increase in scale values from first to last ratings (across labels) during the course of the experiment.
In order to test for a linear increase in the scale values according to their order of rating a trend analysis was performed across labels and sequences. The 80 scores of the 19 raters were divided into five equal groups of 16 as they were rated according to order of appearance in the experiment. Although labels were disregarded in forming the groups, all five groups had approximately equal numbers of samples labeled "cleft palate" and, similarly, equal numbers of samples labeled "normal" (Table III). As Table IV indicates, the trend analysis shows a significant linear trend, indicating that throughout both sequences the scale values became progressively higher. That is, the increase in perceived nasality was not just greater in the later sequence, but the increase was significantly close to uniform throughout the experiment. The ordered variable causing the linear trend in this study was extraneous to the experimental variables, such as change of scaling referents. Finally, the trend analysis used in this
### TABLE III

**TREND ANALYSIS GROUPING**

<table>
<thead>
<tr>
<th>Group</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of samples labeled &quot;cleft palate&quot;</td>
<td>9</td>
<td>7</td>
<td>8</td>
<td>7</td>
<td>9</td>
</tr>
<tr>
<td>Number of samples labeled &quot;normal&quot;</td>
<td>7</td>
<td>9</td>
<td>8</td>
<td>9</td>
<td>7</td>
</tr>
<tr>
<td>Totals</td>
<td>16</td>
<td>16</td>
<td>16</td>
<td>16</td>
<td>16</td>
</tr>
</tbody>
</table>

### TABLE IV

**MEAN RATIO SCORES FOR THE FIVE TREND ANALYSIS GROUPS**

<table>
<thead>
<tr>
<th>Group</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean Scale Scores</td>
<td>.317</td>
<td>.343</td>
<td>.406</td>
<td>.383</td>
<td>.462</td>
</tr>
</tbody>
</table>

### TABLE V

**SUMMARY OF TREND ANALYSIS**

<table>
<thead>
<tr>
<th>Source of Variance</th>
<th>Sum of Squares</th>
<th>df</th>
<th>Mean Square</th>
<th>F</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linear Regression</td>
<td>3.304</td>
<td>1</td>
<td>3.304</td>
<td>60.07</td>
<td>.05</td>
</tr>
<tr>
<td>Deviations</td>
<td>1.232</td>
<td>3</td>
<td>.410</td>
<td>7.49</td>
<td>.05</td>
</tr>
<tr>
<td>Quadratic Component (curvature)</td>
<td>.013</td>
<td>1</td>
<td>.013</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Within</td>
<td>82.897</td>
<td>1515</td>
<td>.055</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>1520</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
study is not designed for post hoc analysis and the results, therefore, must be viewed with appropriate caution (Edwards, 1965).

**Effect of Raters**

The third main effect of the factorial analysis of variance, raters, was also responsible for significant variance of the experiment. Although there was no hypothesis regarding differences in individual raters, their performances were compared in order to determine if the significant effect of labels was merely the result of a few extreme rater means. First, visual inspection of Figure II indicated that every rater's mean ratio value for all samples labeled "cleft palate" was greater than her respective mean for the samples labeled "normal." Secondly, as seen in Table VI, a sign test for each individual rater was computed comparing the paired ratings of the same samples in the two different label conditions and disregarding the fact that the comparisons were across the two experimental sequences. The results revealed that six of the raters were highly influenced ($Z > .01$) by labels, another nine raters had $Z$ scores ranging from .09 to .21 probability levels, and only four raters produced $Z$ scores at or near the .50 probability level (.30 to .50). Finally, a sign test using each rater as one sign (plus or minus, depending on the most frequent sign in her sample comparisons). The results showed that all but one rater had more plus than minus ratings in the predicted direction ($P > .001$).

**Summary**

A 2X2X19 factorial analysis of variance and Scheffé test indicated that the speech clinicians were influenced in their evaluations of voice samples by expectation effects. These analyses indicate that within each
Figure II

LXR INTERACTION FIGURE

- 40 samples labeled "cleft palate"

---- same 40 samples labeled "normal"
TABLE VI
SIGN TEST FOR PAIRED COMPARISONS OF SAME SAMPLES UNDER TWO LABEL CONDITIONS, DISREGARDING SEQUENCES, FOR INDIVIDUAL RATERS.

<table>
<thead>
<tr>
<th>Rater</th>
<th>Z Score</th>
<th>Probability of Occurrence of Z</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3.30</td>
<td>.0005</td>
</tr>
<tr>
<td>2</td>
<td>2.83</td>
<td>.0023</td>
</tr>
<tr>
<td>3</td>
<td>.81</td>
<td>.2090</td>
</tr>
<tr>
<td>4</td>
<td>1.11</td>
<td>.1335</td>
</tr>
<tr>
<td>5</td>
<td>4.93</td>
<td>.00001</td>
</tr>
<tr>
<td>6</td>
<td>.17</td>
<td>.4325</td>
</tr>
<tr>
<td>7</td>
<td>.16</td>
<td>.4364</td>
</tr>
<tr>
<td>8</td>
<td>.80</td>
<td>.2119</td>
</tr>
<tr>
<td>9</td>
<td>1.00</td>
<td>.1587</td>
</tr>
<tr>
<td>10</td>
<td>1.36</td>
<td>.0869</td>
</tr>
<tr>
<td>11</td>
<td>1.00</td>
<td>.1587</td>
</tr>
<tr>
<td>12</td>
<td>1.00</td>
<td>.1587</td>
</tr>
<tr>
<td>13</td>
<td>.50</td>
<td>.3085</td>
</tr>
<tr>
<td>14</td>
<td>.00</td>
<td>.50</td>
</tr>
<tr>
<td>15</td>
<td>4.60</td>
<td>.00001</td>
</tr>
<tr>
<td>16</td>
<td>1.00</td>
<td>.1587</td>
</tr>
<tr>
<td>17</td>
<td>4.70</td>
<td>.00001</td>
</tr>
<tr>
<td>18</td>
<td>2.44</td>
<td>.0073</td>
</tr>
<tr>
<td>19</td>
<td>1.12</td>
<td>.1314</td>
</tr>
</tbody>
</table>
of the two sequences clinicians rated samples labeled "cleft palate" more nasal than samples labeled "normal."

The analysis of variance and Scheffé test produced no support for the second experimental hypothesis that the rating differences according to label would decrease in the second sequence. Instead, the analyses showed that the label differences were greater in the second sequence than in the first.

A trend analysis indicated a linear increase in the scale values regardless of sample label throughout the experiment. This linear trend was the result of an extraneous variable, such as change in raters' referents.

Differences in the performance of raters were analyzed for significance in the analysis of variance in order to isolate this source of variability from the analysis of the effects of labels and sequences. The results showed a significant difference in the performance of raters.
Chapter IV

DISCUSSION

As stated previously, the effect of labels (expectation effect) significantly biased, on the average, the judgments of the raters. This finding agrees with those of other investigators (Rosenthal, 1968 and 1969; Fleming and Antonen, 1971; Cahen, 1966; Brophy and Good, 1970; and Meichenbaum, et al., 1969). These investigations were in the areas of education and psychology and demonstrated the effects of expectations on the behavior of teachers and psychologists in various evaluative situations. The present study demonstrates that speech clinicians can be biased in an evaluative situation by the expectation effect, generated by their training.

It is important to note that raters, as individuals, were differentially affected by labels and sequences. Some raters were highly influenced by labels beyond almost any doubt; a few did not seem to be influenced by labels much at all and rated relatively few samples labeled "cleft palate" as more nasal than the same samples labeled "normal" (raters 6, 7 and 14). Nonetheless, as the label X raters interaction graph indicates, the average nasality rating of samples labeled "cleft palate" was higher for all raters than the average rating of samples labeled "normal." Clearly, some raters are more susceptible than others to labeling effects, and, just as clearly, the effect of labels (expectation effect) over groups of raters are obvious and in the predicted direction.

The present study used a very artificial evaluative situation which was intended to maximize the possibility of producing an expecta-
tion effect. The expectation effect was generated solely through the use of intensive and repeated presentation of triggering devices. The initial triggers were in the instructions for the rating of the experimental samples; this information was false, but, yet, apparently the raters accepted the task as leading to a reasonable, worthy research goal: the study of rating nasality in the voice quality of cleft palate and normal speakers. Secondly, triggers for expectations regarding individual samples were the auditory and visual stimuli (labels) immediately preceding each sample. The auditory labels ("cleft palate" or "normal") were spoken by the experimenter and the subtle visual cues were c or n typed to the left of each numbered scoring space for each sample on the score sheet pages. Apparently this three-part cuing or triggering system was sufficient to evoke strong enough expectations to override the raters' judgments based solely on the physical, acoustic stimuli. The biasing process occurred even though the validity and reliability of the nasality judgments should have been enhanced by the backward-played recordings (Lintz and Sherman, 1961), the short time between sample presentations allowing for relative sample comparisons, and the voice of E preceding each sample which conceivably could have been used as an acoustic referent. The judgment tasks of this study differ from other expectation effect studies in these latter regards. Previous studies in education and psychology used judgment tasks which appear to be more nebulous, such as ratings of a student's classroom attitudes and/or general academic skills over a period of several months.

The second hypothesis dealt with a sequence effect in that the author believed that, with a second experimental presentation of the voice samples, the raters would become suspect of the labeling instruc-
tions/procedures and realize the fallaciousness of the labels. However, the data shows a reverse effect of that which was hypothesized. Even though the raters were hearing the samples for a second time their mean ratio values differences between labels was greater in the second sequence than the first. This result would appear to indicate that the expectation effect had more influence in the second sequence. However, this indication may be partly a function of the trend for increasing nasality values over latter samples.

It is difficult to account for the linear increase in line drawing for all samples; however, Young (1970) described two effects of psycho-physical scaling which bias the evaluation of speech samples. Data indicate that if scaling "anchors" (referents) are not provided several times during the scaling task the raters will lose their original scaling referents and adopt other referents, such as a sample immediately preceding the sample to be judged. Use of referents of relatively higher scale values result in ratings of higher scale values. In the present study the raters could have readily lost the acoustic referents, due to the number of samples and score sheet pages, that they had developed in their training and/or life experience and the range scaling referents given at the outset of the experiment. The raters behaved as if they became increasingly separated from the afore-mentioned referents, and progressively adopted higher scale value referents due to implicit directions to hear nasality and explicit instructions imbibed in the labels "cleft palate" and "normal." That is, implicit directions to hear nasality, reiterated over and over, could have cumulative effects. Thus, the raters perceive progressively more nasality. Instead of getting the predicted semantic satiation, decrease in label effect, the study seemed to produce satiation of the raters real-life voice quality (nasality).
While conclusions about the effect of sequences can only be speculative, the analysis of variance and the Scheffe test data support acceptance of the hypothesis that speech clinicians can be biased by training-induced expectations. In this study, the previous statement must be qualified in that the raters, as a group, rated the voice samples differentially according to labels within the two sequences. While the data is unclear regarding between sequence comparisons, none of the data indicates that, on the average, labels were ineffective in triggering expectation effects in the group of raters.

The immediate implications of this study are that the expectation effect may influence a great deal of the clinical evaluative behavior of speech clinicians. The situation in which the raters made their judgments in this study was vastly different from typical clinical evaluative situations. That is, in a clinical evaluation the clinician spends much more time listening to the client, evaluating nasality in a variety of speech responses, and seeing as well as hearing the client. Nevertheless, the clinicians of this study, some more than others, were biased by labels in the artificial situation and the implications are that expectations may distort our perceptions in real evaluative situations. The clinical judgments of a typical cleft palate-voice evaluation lead to serious decisions about treatment recommendations for surgery, therapy placement, etc. and such judgments should therefore be as free of label effects as possible. Those interested in training clinicians should be interested in identification of those variables which help some students to be more resistant to labeling effects than others. Research investigating the various effects of training procedures on subsequent judgments in the presence of labels is very necessary.
If expectations influence voice evaluations they may also bias our evaluations of articulation, fluency, and language disorders. If our evaluations of these dimensions are biased then our recommendations will be inaccurate.

The sample presentation method used in this study closely resembles that of other voice evaluation studies which did not consider the expectation effect as a variable (Spriesterbach and powers, 1959; Lintz and Sherman, 1961; and Counihan and Cullinan, 1972). We need to know whether, and to what extent, expectation effects have biased the general body or parts of data accumulated to date about the rating of speech or voice samples.

Future research in this area should utilize different designs and analyses to gather data about questions raised by the present study. The analysis problems of sequence could be resolved by using several groups of raters: a control group (rating samples without labels), and groups to rate various numbers of samples. Such a design would allow analysis of comparisons of sample ratings with and without labels, the analysis of the effect of the length of the experiment (number of samples), and, as suggested by Young (1970), the analysis of presenting the initial referents several times during the experiment (to prevent intrarater change of referents). Also, score sheets could be varied from one page, containing referents, practice, and experimental ratings, to several pages of score sheets as in the present study. Random grouping of clinicians in these designs should allow analysis of the apparent extraneous variables which affected the present study. Future designs should closely analyze the differences in individual raters to determine what behaviors of some raters causes them to be only minimally affected by
expectation while others appear to be greatly biased. Such findings could allow analysis of the causes and control of the expectation effect.

Many more aspects of expectation effects could be measured by a variety of experiments. For instance, in the present study only female graduate student clinicians were used as raters. We need to know if the expectation phenomenon is similar for other categories of clinicians: males, experienced clinicians employed in a variety of settings, clinicians trained in a variety of training programs, etc. Numerous experimental conditions need to be tested: audio plus visual presentations of voice samples, varied lengths of sample presentation, varied times between presentations, samples from both sexes, and samples utilizing various speech responses. Finally, all evaluative situations in speech pathology and audiology need to be investigated for the possible influences of the expectation effect.

Two areas of expectation research were not investigated by this study, but the implications of this study are relevant to these areas. First, as mentioned in the first chapter, a second stage of the expectation effect, that evaluatees respond according to evaluators expectations, has been investigated (Rosenthal, 1968 and 1969; Miller, 1972; Barber, et al., 1969; Claiborn, 1969; and Fleming and Antonen, 1971). The confirmation of this hypothesis is controversial, as indicated in the literature by Miller (1972), therefore, more research is needed in this area. The present study has shown that speech clinicians can be biased by expectations, we need to determine whether our clients react to conform to those expectations. Secondly, Rosenthal (1968 and 1969) has extended this concept to the whole of the experimental method in behavioral research. That is, he has produced data, again unreplicated and questioned by other
investigators (Barber, et al., 1969; Claiborn, 1969; and Fleming and Antonen, 1971), indicating experimental subjects produce results which conform to the expectations of experimenters. Confirmation of the Rosenthal research would raise serious questions about the validity of much of the accumulated behavioral research data.
Chapter V

SUMMARY AND CONCLUSIONS

Previous studies in education and psychology have found that evaluators have been biased by expectations about evaluatees. The present study indicated that speech clinicians, when evaluating tape recorded voice samples, can also be biased by their expectations in an evaluative situation.

Nineteen female graduate students in Speech Pathology and Audiology were chosen according to criteria that provided training sources of expectations. They rated their perceived nasality in 40 backward-played, audio-taped voice samples of normal third and fourth grade boys. The raters judged each sample twice, once in each of the two connected sequences. In the first sequence twenty of the samples were randomly assigned the label "cleft palate" with the remaining twenty labeled "normal," then all forty samples were randomly ordered and re-recorded backward on a master tape. The second sequence consisted of the same 40 samples in a rerandomized order with their labels reversed, so that the twenty samples labeled "cleft palate" in the first sequence were labeled "normal" in the second sequence, etc. This second sequence was recorded immediately following the first, in a similar manner, with no discernible separation between the two sequences. Prior to presentation of the two experimental sequences, the raters judged four range scaling samples and ten practice samples. Following the instructions for and presentation of the four range samples and ten practice samples, and immediately preceding the presentation of the experimental samples, fallacious instructions were given which were designed to trigger the
raters' expectations about the labeled samples. The entire task required about 33 minutes.

A 2X2X19 analysis of variance and Scheffe test indicated that speech clinicians can be influenced in their evaluations of voice samples by the expectation effect. These analyses indicate that, within each of the two sequences and in some comparisons across sequences, the clinicians rated samples labeled "cleft palate" as more nasal than samples labeled "normal."

The analysis of variance and Scheffe test produced no support for the second experimental hypothesis that the rating differences according to label would decrease in the second sequence. Instead, the analysis showed that the label differences were greater in the second sequence than the first.

A trend analysis indicated a linear increase in the scale values regardless of sample label throughout the experiment. This linear trend may have been the result of an extraneous variable such as change in referents by the raters as the experiment progressed.

Differences in the performance of raters were analyzed for significance in the factorial analysis of variance in order to isolate this source of variability from the analysis of the effects of labels and sequences. The results reveal a significant difference in the performance of raters; additional analysis indicated that some raters are much more influenced by labeling effects than others.

The results indicate that, as a group, the raters were biased by formally learned expectations about differential nasality characteristics in cleft palate and normal children. The effect of these expectations was produced by the extensive use of "triggering" stimuli in the rating
instructions and labeling procedures.

A linear increase in ratings across sequences and labels occurred and was attributed to an extraneous factor or factors. The extraneous influence appeared to be a loss and change of rating referents due to the nature and length of the experiment.

The implications of this study are that expectations may influence a great deal of speech clinicians' evalutive behaviors. Our evaluations of cleft palate children, as well as clients with other disorders, lead to important recommendations for surgery, therapy placement, etc. If our evaluations are biased then our recommendations may be inappropriate. Also, other voice studies in the area of nasality rating have not considered expectation effects as a variable, therefore, the data may be questionable.

Future research is indicated to examine two areas of incomplete analysis in this study. A design utilizing a control group plus several experimental groups of raters could be used to analyze the sequence findings of this study and the possible influence of extraneous variables in the present study. Secondly, the use of a design to test individual rater differences could determine why some raters are, apparently, only slightly influenced by expectations. Such findings could show us how to control the expectation effect in our evaluative behavior. Finally, the need for investigation of all evaluative situations in speech pathology and audiology is indicated to determine the scope of the expectation effect.

The second stage of expectation effects were not investigated, but the findings of the present study are relevant. The second stage, as hypothesized by Rosenthal, is the reaction of the evaluatees to the
expectations of the evaluators. Confirmation of this hypothesis would indicate the behavior of our clients, as well as experimental subjects, may be influenced by our expectations.
REFERENCES


APPENDIX I

INSTRUCTIONS FOR RANGE SCALING TASK

"In just a few minutes I'm going to ask you to listen to the voices of some children. What you will hear is a short reading sample of their speech played backwards. What I would like you to do is to tell me about the hypernasality or lack of it in the child's voice. The samples will be played backwards so that your judgments will not be influenced by the child's articulation errors or other speech problems.

First you will hear four samples which represent the most hypernasal and the least hypernasal voices of those you will judge later. On the table in front of you are pencils and score sheets. To rate the hypernasality of each sample merely draw a horizontal line on your score sheet beside the number corresponding to the number of the speech sample. You are to indicate the degree of hypernasality you detect by the length of the line you draw; the more hypernasality you detect, the longer the line. You may choose any length line you want to represent the hypernasality you hear; just remember that more hypernasality calls for a longer line, less hypernasality calls for a shorter line. Each speech sample is about 10 seconds long and all samples are followed by at least five seconds of silence. Listen to the entire sample and judge the hypernasality of the sample on a whole—not just on one or two words. There are no right or wrong lengths in the lines you draw. I am interested only in what your judgment of the sample is. Any questions? If not, here are four samples to illustrate the differences in hypernasality you'll be judging throughout the session, and to give you practice in rating."

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
APPENDIX II

INSTRUCTIONS FOR PRACTICE RATINGS

"Here are 10 more practice samples for you to judge. The procedures are the same as when you scored the previous four samples, however, the next 10 practice items are not necessarily representative of differences in nasality you will be judging later. Now, here are ten more practice samples."
APPENDIX III

INSTRUCTIONS FOR EXPERIMENTAL RATINGS

"Now I want you to judge the hypernasality in the voices of some cleft palate children and that of some normal children. Some of the cleft palate children were recorded before they received medical treatment, some before the completion of medical treatment, and all before they received speech therapy. I want to know if you can differentiate between the cleft palate children's level of treatment by judging the hypernasality of their voices."
APPENDIX IV

FINAL STATEMENT

"Thank you for participating in this experiment. Please do not discuss this experiment with anyone until I inform you that all subjects have completed the experiment."
APPENDIX V

SCORE SHEET

1.
2.
3.
4.
5.
6.
7.
8.
9.
10.
11.
12.
13.
14.
c15.
n16.
n17.
n18.
c19.
n20.
c21.
c22.
c23.
c24.
n25.
n26.
n27.