Examiner Bias and Intelligence Test Scoring: a Study of the Effect of Speech-Handicapped Responses on the Scoring of the WISC-R Comprehension Subtest

Julia B. Peyser
Andrews University

Follow this and additional works at: https://digitalcommons.andrews.edu/dissertations

Part of the Disability and Equity in Education Commons, and the Speech and Hearing Science Commons

Recommended Citation
https://digitalcommons.andrews.edu/dissertations/630

This Dissertation is brought to you for free and open access by the Graduate Research at Digital Commons @ Andrews University. It has been accepted for inclusion in Dissertations by an authorized administrator of Digital Commons @ Andrews University. For more information, please contact repository@andrews.edu.
Thank you for your interest in the

Andrews University Digital Library
of Dissertations and Theses.

Please honor the copyright of this document by not duplicating or distributing additional copies in any form without the author’s express written permission. Thanks for your cooperation.
INFORMATION TO USERS

This reproduction was made from a copy of a document sent to us for microfilming. While the most advanced technology has been used to photograph and reproduce this document, the quality of the reproduction is heavily dependent upon the quality of the material submitted.

The following explanation of techniques is provided to help clarify markings or notations which may appear on this reproduction.

1. The sign or "target" for pages apparently lacking from the document photographed is "Missing Page(s)". If it was possible to obtain the missing page(s) or section, they are spliced into the film along with adjacent pages. This may have necessitated cutting through an image and duplicating adjacent pages to assure complete continuity.

2. When an image on the film is obliterated with a round black mark, it is an indication of either blurred copy because of movement during exposure, duplicate copy, or copyrighted materials that should not have been filmed. For blurred pages, a good image of the page can be found in the adjacent frame. If copyrighted materials were deleted, a target note will appear listing the pages in the adjacent frame.

3. When a map, drawing or chart, etc., is part of the material being photographed, a definite method of "sectioning" the material has been followed. It is customary to begin filming at the upper left hand corner of a large sheet and to continue from left to right in equal sections with small overlaps. If necessary, sectioning is continued again—beginning below the first row and continuing on until complete.

4. For illustrations that cannot be satisfactorily reproduced by xerographic means, photographic prints can be purchased at additional cost and inserted into your xerographic copy. These prints are available upon request from the Dissertations Customer Services Department.

5. Some pages in any document may have indistinct print. In all cases the best available copy has been filmed.

University Microfilms International
300 N. Zeeb Road
Ann Arbor, MI 48106

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
Peyser, Julia B.

EXAMINER BIAS AND INTELLIGENCE TEST SCORING: A STUDY OF THE EFFECT OF SPEECH-HANDICAPPED RESPONSES ON THE SCORING OF THE WISC-R COMPREHENSION SUBTEST

Andrews University

University Microfilms International

Copyright 1984 by Peyser, Julia B.

All Rights Reserved
PLEASE NOTE:

In all cases this material has been filmed in the best possible way from the available copy. Problems encountered with this document have been identified here with a check mark ✓.

1. Glossy photographs or pages
2. Colored illustrations, paper or print
3. Photographs with dark background
4. Illustrations are poor copy ✓
5. Pages with black marks, not original copy ✓
6. Print shows through as there is text on both sides of page
7. Indistinct, broken or small print on several pages ✓
8. Print exceeds margin requirements
9. Tightly bound copy with print lost in spine
10. Computer printout pages with indistinct print
11. Page(s) lacking when material received, and not available from school or author.
12. Page(s) seem to be missing in numbering only as text follows.
13. Two pages numbered. Text follows.
14. Curling and wrinkled pages
15. Other

University Microfilms International

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
Andrews University
School of Graduate Studies

EXAMINER BIAS AND INTELLIGENCE TEST SCORING: A STUDY
OF THE EFFECT OF SPEECH-HANDICAPPED RESPONSES ON THE
SCORING OF THE WISC-R COMPREHENSION SUBTEST

A Dissertation
Presented in Partial Fulfillment
of the Requirements for the Degree
Doctor of Education

by
Julia B. Peyser
September 1984

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
EXAMINER BIAS AND INTELLIGENCE TEST SCORING: A STUDY OF
THE EFFECT OF SPEECH-HANDICAPPED RESPONSES ON THE
SCORING OF THE WISC-R COMPREHENSION SUBTEST

A dissertation
presented in partial fulfillment
of the requirements for the degree
Doctor of Education

by
Julia B. Peyser

APPROVAL BY THE COMMITTEE:

Chairman: Wilfred C. A. Dutcher
Committee Member: Robert A. Williams
Committee Member: Roy E. Hartbauer
Committee Member: Bernard M. Lall
External Examiner: Leonard E. James

Dean, School of Education

Sept. 21, 1984
Date approved

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

EXAMINER BIAS AND INTELLIGENCE TEST SCORING: A STUDY OF THE EFFECT OF SPEECH-HANDICAPPED RESPONSES ON THE SCORING OF THE WISC-R COMPREHENSION SUBTEST

by

Julia B. Peyser

Chairman: Wilfred G. A. Futcher
ABSTRACT OF GRADUATE STUDENT RESEARCH

Dissertation

Andrews University
School of Education

Title: EXAMINER BIAS AND INTELLIGENCE TEST SCORING: A STUDY OF THE EFFECT OF SPEECH-HANDICAPPED RESPONSES ON THE SCORING OF THE WISC-R COMPREHENSION SUBTEST

Name of researcher: Julia B. Peyser

Name and degree of faculty adviser: Wilfred G. A. Futcher, Ph.D.

Date completed: September 1984

Problem

One of the concerns in education today is bias in intelligence tests. Examiner bias and the expectancy effect is largely ignored. An unexplored variable in the research is an examinee's speech handicap and the effect it may produce in the scoring of an intelligence test.

Method

Identical responses to the Comprehension subtest of the Wechsler Intelligence Scale for Children-Revised were audiotaped on separate cassettes by two children, one with and one with no speech handicap. Sixty-four examiners, thirty-two practicing psychologists...
and thirty-two graduate students of psychology, were randomly assigned to score the responses on one of the tapes. Sixteen of each group of examiners scored the responses by the child with the speech handicap; sixteen of each group scored the responses of the child with no speech handicap. A two-way analysis of variance and the test of simple effects were utilized to analyze the obtained scores.

Results

1. There was no significant difference between overall means of scores awarded to the speech-handicapped and non-speech-handicapped responses.

2. There was a significant difference between the scores awarded by the practicing psychologists and the graduate psychology students to the speech-handicapped responses. The psychologists gave the lowest mean score to these responses, the students the highest.

3. There was no significant interaction between type of examiner and type of response.

4. There was a surprisingly wide range of scores overall—a range of two standard deviations.

5. The range of scores awarded by the psychologists was greater than that awarded by the graduate students.

6. Each group awarded a wider range of scores to the speech-handicapped responses than to the non-speech-handicapped responses.

7. The graduate students awarded higher mean scores than did the psychologists to each type of response.
Conclusions

The results corroborate the body of research which has found widely different IQ's obtained by different examiners and that the examiners' experience level is not a significant factor in obtaining more reliable or consistent results in the scoring of test responses. The wider range of scores awarded by each group of examiners to the speech-handicapped responses, as well as the significant difference in their scoring of these responses, exemplify the wide variety of effects a speech handicap has on examiners, ranging from one that is negative to one of compensation, i.e., a negative to a positive "halo effect."
# TABLE OF CONTENTS

LIST OF TABLES ......................................................... v

ACKNOWLEDGMENTS ................................................... vi

Chapter

I. THE PROBLEM ................................................... 1

   Theoretical Bases ........................................... 1
   Problem Situation and Need for Study ............. 3
   Purpose of Study ........................................ 6
   Definition of Terms ..................................... 6
   Scope and Delimitations of Study ................ 8
   Hypotheses ................................................. 9
   Importance of Study .................................... 9

II. REVIEW OF RELATED LITERATURE ...................... 10

   Overview .................................................. 10
   Test Bias .................................................. 11
   Experimenter or Examiner Effect .................. 18
   Expectancy or Halo Effect and Self-Fulfilling Prophecy ................. 30
   Expectancy or Halo Effect and Spoken Language .... 41
   Summary ................................................ 65

III. METHODOLOGY ............................................... 68

   Type of Research ....................................... 68
   Selection of Subjects ................................ 68
   Instrumentation and Test Reliability ............ 69
   Procedures ............................................. 72
   Hypotheses and Statistical Analysis ........... 75
   Summary ................................................ 76

IV. RESULTS ...................................................... 77

   Respondents ............................................. 77
   Basic Data ............................................. 79
   Testing the Hypotheses ................................ 84
   Analysis of Simple Effects ......................... 86
   Test of Simple Effects ............................... 88

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
V. SUMMARY, CONCLUSIONS, IMPLICATIONS AND RECOMMENDATIONS 90

Summary ................................................................. 90
Overview of Related Literature ................................. 91
Test bias .................................................................. 91
Experimenter or examiner effect ............................... 92
Expectancy or "halo effect" and "self-
fulfilling prophecy" ................................................ 93
Expectancy or "halo effect" and spoken
language .................................................................. 94
Methodology ............................................................ 95
Selection of subjects .................................................. 96
Instrumentation ....................................................... 96
Procedures .............................................................. 97
Statistical analysis .................................................... 98
Summary and Discussion of Findings ....................... 98
Range of scores ....................................................... 98
Means of scores ...................................................... 100
Testing the hypotheses ............................................ 101
Conclusions ........................................................... 106
Implications ............................................................ 108
Recommendations ................................................... 113

APPENDIX A: Letter to Participants ......................... 116
APPENDIX B: List of Procedures ............................... 118
APPENDIX C: Letter of Appreciation ....................... 120
APPENDIX D: Responses to Comprehension Subtest Questions 122
APPENDIX E: Raw and Scaled Scores of Each Respondent 125
BIBLIOGRAPHY ...................................................... 128
VITA ........................................................................ 134
LIST OF TABLES

1. Range of Raw Scores and Scaled Scores Within Each Group .............................................. 80
2. Number of Individuals Giving Each Raw Score and Each Scaled Score .............................. 81
3. Percentage of Scaled Scores in Each Intelligence Classification ........................................... 82
4. Means of Raw Scores ............................................................................................................ 83
5. Means of Scaled Scores ........................................................................................................... 83
6. Analysis of Variance of Overall Raw Scores by Rater and Child ....................................... 85
7. Analysis of Variance of Overall Scaled Scores by Rater and Child ....................................... 85
8. Mean and Variance of Each Cell .......................................................................................... 87
9. Simple Effects--Raw Scores ................................................................................................ 88
10. Simple Effects--Scaled Scores ............................................................................................. 89

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

The cooperation and assistance of many are needed for an idea to become a research reality. The special contribution of each person who became involved in this study will be long remembered and appreciated—the friendliness of everyone on the Andrews University campus; the helpfulness of the University editor, Joyce Jones, and my typist, Pat Saliba; the cheerful assistance of the departmental secretary, Betty McCormick, and that of a friend, Ila Bigford, both of whom helped collate, distribute, and collect the packets; the cooperation of the psychologists, graduate students, and children who participated so willingly in the research; the encouragement of the first chairman of my committee, Dr. Conrad Reichert; the invaluable assistance of Dr. Peter Blitchington, the second chairman of my committee who followed the study through to the proposal stage; the steady support and suggestions of the committee members, Dr. Roy Hartbauer, Dr. Bernard Lall, and Dr. Robert Williams; the helpful suggestions offered by the external member of the committee, Dr. Leonard James of Indiana University at South Bend; the friendly support of Dr. George Akers, the Dean of the School of Education; and the insights, patience, and unfailing sense of humor of Dr. Wilfred Futcher, the committee chairman who guided the study through to its completion.

I wish to express a special note of thanks and appreciation
to my family. Their confidence and encouragement have been a constant source of support.
CHAPTER I

THE PROBLEM

It has long been hypothesized that human behavior is affected by others' expectations and situational variables. Only in comparatively recent times has the concept been gaining strong support from scientific research. More sophisticated experimental designs are being used to observe human behavior in an attempt to assess and understand the dynamics of its interactions. In his book, *Experimenter Effects in Behavioral Research* (1976), Dr. Robert Rosenthal, of Harvard University, summarizes the extensive research indicating that expectancy effects do exist. Their interpretation, however, is as complex as the interactions themselves.

Theoretical Bases

"Self-fulfilling prophecy" is one conceptualization of an expectancy effect that has become a sine qua non of daily life. It is the influence or effect, intentional or unintentional, that expectations can have on one's own or another's behavior. Whether positive or negative, recipients of these expectations accept the evaluations and unconsciously condition their behavior to "fulfill" the prophesied expectations. Literature attests to the effects teacher expectations may have on student behavior, even on intellectual performance (Rosenthal, 1968; 1976, p. 130).

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
Expectancies can be "fulfilled" as well for the recipient by those who have the expectations. This is the "halo effect," another conceptualization of an expectancy effect that has become a basic assumption of human behavior. The judgmental situation is particularly vulnerable to this type of effect. Examiners, testers, teachers, interviewers, friends, and enemies interpret the actions and responses of those they are assessing according to their own perceptions of them. The more positively someone is viewed by an evaluator, the brighter the "halo" is perceived, the more positive will be the evaluation or rating, regardless of the reality of the responses and vice versa.

The attributes of those involved in the examiner-examinee interaction may directly or indirectly affect not only the subject's responses but the examiner's interpretation of them. Race, sex, age, socio-economic status, warmth of personality, to name a few variables, affect not only the interaction of questions and responses, perceptions and expectations, but the rating or scoring of the responses as well. The "halo effect" can find its way into the scoring to produce a "self-fulfilling prophecy" for the respondent. It can be very subtle, found in apparently neutral and objective situations, where participants who consider themselves unbiased are not aware that the effect is occurring. It has even entered the sacrosanct halls of scientific investigation. Experimenters have to be wary of unconscious bias in the direction of their expectations as well as of personal blind spots.

Cline, Atzet, and Holmes (1972) discuss the issue of "cue validity," which cues "... contribute most to accurate judgment and
appraisal of another" (p. 8) as "... accurately perceiving and judging another is a complex task which is affected by many variables. Cues can be covert or overt, verbal or nonverbal" (p. 17). In their research over the years, they found visual image to have a strong effect on ratings of responses. A negatively perceived visual image can inaccurately cancel out the content of a verbal message, reducing accuracy of judgment. The results of their research indicate that for accurate rating of responses "... what most interviewees say in response to interview questions is far more important, as cues, than what they look like, what the voice sounds like, how they act or move (without the sound) all put together" (p. 14).

If what is said is the most important cue, what happens to the validity of this cue when what is said is colored by how it is said? Various dialect patterns may produce stereotypical judgments about the speaker's socioeconomic status and even intelligence. The less standard the pattern, the lower the level generally estimated. Today much attention is being focused on the apparently negatively biasing effect that the nonstandard English of Black dialect has on expectations, perceptions, and consequent evaluations of the performance of those who use it.

**Problem Situation and Need for Study**

Intelligence testing is coming under careful scrutiny today. It is an area that, like scientific research, has long had an aura of neutrality and objectivity. The validity of the content of the IQ test itself is being strongly questioned, even disputed in court cases, as to cultural bias. The issue of expectation effects has only
tentatively entered the domain of the psychologists themselves. The dynamics of the examiner-examinee interaction in psychological testing and its effects on test results is receiving more and more attention. The research so far is inconclusive, with some results suggesting that there are effects, others that there are none. A number of attributes and factors have been utilized as variables. It would appear, however, that no studies have investigated speech handicap as a variable in this interaction.

In a recent court case in New York, a twenty-seven-year-old man was awarded a half-million dollars in damages for having been erroneously diagnosed and educated as mentally retarded. A school psychologist had "... misinterpreted a severe speech defect as a mental problem" (NASSP Newsletter, 1979, p. 4).

In his textbook, Human Development in Western Culture (1978), Harold W. Bernard makes the statement: "Statistical studies verify a positive relationship between language development and intelligence, but then intelligence tests are largely verbal (Ornstein, 1976)" (p. 241). This is, however, a misrepresentation of the issue Ornstein is addressing in his 1976 Phi Delta Kappan article. The relationship being discussed is not between general language development and intelligence, as Bernard implies, but rather the language development of members of minority groups, specifically Black, and whether it is the cause of an "... intellectual deficit among blacks" (p. 404) found in the results of IQ tests, or whether the "deficit" is due to cultural bias within the tests themselves. Ornstein does not believe the IQ tests are culturally or linguistically biased, but that the problem which needs to be addressed lies in the mode of the language
development of the minority group. Perhaps Bernard agrees with him but proceeds to overgeneralize to language development in general.

Bernard also states: "Speech defective children have slightly below average intelligence scores . . ." (p. 133) but cites no validating research for this statement. Since speech is the vehicle through which language content is expressed, it is conceivable that the two may be confused or viewed as the same. Is "speech defect" being confused with dialectal or language differences? Can it be a cue for a stereotypical judgment? Or, if Bernard's statement is a reality, do the lower intelligence test scores of "speech-defective children" indicate lower mental ability or lowered scoring of the verbal subtests by an examiner due to a halo effect possibly engendered by statements such as those by Bernard?

Does the speech defect negatively bias the examiner's perception and judgment of the examinee and his language ability; produce a lowered expectation in the examiner; cancel out the actual language content of the oral response; and, consequently, result in a lowered interpretation and scoring of the test responses? This could indeed account for the below-average IQ scores of children with speech defects.

In fifteen years of experience as a school speech therapist, elementary-school principal, and member of multidisciplinary teams evaluating children with learning problems, this researcher has heard a number of professional and lay people imply or state that a child with a speech handicap probably has a lower IQ level. Both state and federal governments have become concerned enough about the possible confusion of abilities to include statements in recent rules relevant
to the problem. The Michigan rules for Special Education, for example, include the following:

R340.1721a Evaluation procedure.
Rule 21a ........................................................................................................
(3) When evaluating a person suspected of being handicapped, the public agency shall assure that tests and other evaluation materials used by members of the multidisciplinary team comply with all of the following: .................
(d) are reflective of the person's aptitude or achievement or whatever other factors the test purports to measure rather than reflecting the person's impaired sensory, manual, or speaking skills (emphasis supplied), unless this is what the test is intended to measure (Michigan Special Education Rules, 1980)

Purpose of Study

The specific delimited purpose of the study was to determine if examiners score identical responses to an intelligence-test subtest differently depending on whether the responding examinee has a speech handicap or not. This purpose addressed the issue of whether a speech handicap is a significant variable that must be taken into account in an evaluative situation, producing a "halo effect" in the interpretation and scoring of oral responses to an intelligence test.

Definition of Terms

The major terms utilized in the present study are defined as follows:

1. Articulation errors--"... a speech condition in which the speaker omits or distorts sounds or substitutes one sound for another" (Good, 1973, p. 188) and/or, in this research, inserts sounds
2. Examiner--"one who administers examinations or tests"
Examinee--"a person examined or questioned by means of a test, interview, or questionnaire" (Good, 1973, p. 223)
3. Expectancy effect--a condition where one considers an event as likely to happen--although sometimes used in a narrow sense to indicate the extent to which an expectation is a determinant of another's behavior, it is also utilized generically, a classification of the results of expectations in general; in this latter sense, experimenter effect, halo effect, and self-fulfilling prophecy are types of expectancy effects.

4. Experimenter effect--"...results the experimenter obtains from his subjects may be determined in part by what he expects to obtain" (Rosenthal, 1976, p. 127)

5. Halo effect--"...tendency for an irrelevant feature of a unit of study to influence the relevant feature in a favorable or unfavorable direction. Typically, a strong initial positive or negative impression of a person, group or event tends to influence ratings on all subsequent observations...the more vague and impressionistic the variable to be rated, the more powerful is the effect..." (Isaac, 1977, p. 58)

6. Self-fulfilling prophecy--"...the expectation of the event...changes the behavior of the prophet in such a way as to make the prophesied event more likely" (Rosenthal, 1976, p. 129)

7. Speech handicap--an impairment in one or more of the speech processes--articulation, voice, fluency

8. Test, intelligence--"...a standardized test which measures a wide range of abilities, including verbal, numerical and social competence" (Wolman, 1973, pp. 375)

Test, nonverbal--"Tests which do not utilize verbal
material for presentation or solution of the problem . . . also known as performance tests" (Ibid., p. 382)

Test, verbal--"Any test which is constructed so as to require verbal ability to perform the tasks" (ibid., p. 381).

Scope and Delimitations of Study

Delimitations of the study focused on the examiners, test used, and type of speech handicap.

The subjects of the study, the examiners, were practicing psychologists who administer intelligence tests and graduate psychology students who had the basic training and experience in administering these tests.

The test used was the Comprehension subtest in the Verbal half of the Wechsler Intelligence Scale for Children-Revised, known by the acronym, WISC-R. The Verbal subtests are considered the most vulnerable to individual judgment (Babad et al., 1975; Sattler, 1974; Schroeder & Kleinsasser, 1972). It is in judgmental situations that an expectation effect would be more likely to occur.

The content of the WISC-R, its administration, and its scoring have been highly standardized with explicit instructions for its use (Wechsler, 1974, pp. 53-115). Therefore, the basic training of both groups of examiners in utilizing this test was assumed to be very similar no matter where they received their training.

The speech handicap was one of multi-articulation errors, the substitution, distortion, insertion, and/or omission of sounds, as determined by a standardized test of articulation, the Photo Articulation Test (Pendergast, Dickey, Selmar, and Soder, 1969). The test responses were on cassette audiotapes. There were two sets of
audiotapes, one set containing the responses of an eight-year eleven-month-old girl with the specified speech handicap, the other set containing the responses, identical in content, of another girl of the same age, but with no speech handicap. The use of audiotapes was to substantially decrease or eliminate the confounding cues or variables of physical appearance, visual and verbal interaction, and direct examiner influence; only auditory cues were evident.

**Hypotheses**

The research hypothesis was that there would be a significant difference in the scoring of the two types of responses. The differences would be in the direction of lower mean scores for the speech-handicapped responses. It was also hypothesized that there would be a significant difference between the scoring by the psychologists and the graduate psychology students. These research hypotheses were based upon personal experience and the results of a number of related studies reviewed in chapter 2. The formal hypotheses tested statistically for the study, however, were stated in null and non-directional terms.

**Importance of Study**

The presence of any significant differences would have implications for the education of psychologists in intelligence test scoring. It might indicate a need for greater knowledge in the area of speech handicaps—what they are, and what each means—as well as greater awareness of expectation effects and examiner bias. 

With no studies found on the effect an examinee's speech handicap may have on an examiner, there was already an evident need for research in this area.
CHAPTER II

REVIEW OF RELATED LITERATURE

Overview

In the ten years between 1966 and 1976, Rosenthal (1976) found that over three hundred studies had been done on interpersonal effects both in and out of the laboratory, most concluding that experimenter expectations and biases do exist and affect results obtained. From the very beginning of an experiment, the variables selected by the experimenter, the choice of design, the statistical tests used, the analyses and interpretation of data are all affected, as are the behavior and responses of the subjects themselves.

Many of the studies deal with the effects of the experimenter or examiner on the actual performance of the examinees. Survey research, for example, finds that the race, sex, and perceived socio-economic status of an interviewer affect the content of the responses to questions asked.

Much has been written on the interaction of examiner and examinee attributes in experimental research, as well as linguistic cultural bias within tests and general attitudes towards minorities using nonstandard English. Comparatively little, however, is available on the expectations and biases of the psychologist in the testing situation.

10
Only that literature which was more pertinent to the goal and purpose of the present research was reviewed. Since research studies often involve a number of variables within each experiment, they were not easily separated into discrete categories. Some may therefore be cited in more than one section. They have been divided as much as possible into four main sections: Test Bias, Experimenter or Examiner Effect, Expectancy or Halo Effect and Self-Fulfilling Prophecy, Expectancy or Halo Effect and Spoken Language.

**Test Bias**

There has been much controversy over whether or not IQ tests are culturally biased. Ornstein (1976), in referring to a number of articles on the issue, concludes that they are, of necessity, "culturally laden," reflecting what is needed to succeed in school, but not unfair or biased. Rather "... the scores reflect the unfairness of social and environmental conditions. ... The remedy is not the elimination of the test ... but the elimination of learning inequalities and social inequalities" (p. 404).

Hanley and Barclay (1979) raise the question of the representativeness of the standardization sample of the WISC-R, the revised edition of the WISC. They point out that the percentage of nonwhites in the sample does not match those of the population. It should have been almost double the number used. Although it was thought WISC-R results would have fewer differences in the performance of Whites and Blacks, their study found the contrary. Groups from both races scored lower on the WISC-R than the WISC, but Black subjects scored significantly lower, with a difference of 6.83 points. They warn of the
potential danger of using the WISC-R rather than the WISC with young Black children.

After New York City schools eliminated the IQ from pupils' records and replaced it with achievement tests, David Wechsler, the author of the Wechsler tests, wrote that "Opposition is generally focused not on the way that IQ's are computed, but, more pointedly, on the way they are interpreted and utilized. One interpretation that has caused understandable concern is the notion that a person is 'born' with an IQ which remains immutable" (Wechsler, 1966, p. 64).

An increasing number of court cases attest to the inappropriate placement of minority children in special education classes due to linguistic and cultural bias factors within placement and intelligence tests. Most tests are verbal--written and oral. Hispanic, Black, Mexican, and Chinese Americans--anyone, that is, with a linguistic background different from the linguistic style of the population used to norm the test, norms against which he is being measured--may be penalized for linguistic differences.

Federal district court judges in two cases, "Larry P," in California and "PASE" in Illinois, gave conflicting opinions on the use of standardized IQ testing for assigning Black children to special education classes. Their use was ruled invalid in California where Judge Peckham believed not only that the tests had not been validated for Black populations but that there was "discriminatory intent" by school officials. He stated that the assumption of intellectual inferiority of large numbers of Black children... without validation is unacceptable... invidious when legitimated by ostensibly
neutral scientific IQ scores (emphasis supplied) . . ." (APA Monitor, 1979).

A diametrically opposite ruling was handed down in Illinois where Judge Grady believed that most of the IQ test items (Wechsler and Stanford-Binet tests) were racially neutral, that the use of non-standard English should not affect a child's score, and that " . . . the examiner who knows the milieu of the child can correct for cultural bias by asking the questions in a sensitive and intelligent way . . ." (APA Monitor, 1980). There may be a problem, however, with this assumption.

Bersoff (1981), counsel for the American Psychological Association, in discussing the impact of these two cases, believes that the most lasting results will be the condemnation of EMR (Educable Mentally Retarded) programs which can have such " . . . powerful adverse consequences . . . (especially) if the child is not genuinely retarded" (p. 6). He points out, as did Wechsler in 1966, that the question of the users of the tests is crucial. There may be flaws in the IQ tests themselves but it was also the " . . . failure of psychologists to question their role, to scrutinize the psychometric soundness of their instruments, and to test the validity of their interpretations (emphasis supplied) that resulted in misclassification and miseducation and in the injuries that flowed therefrom to significant numbers of children of all racial and cultural backgrounds" (p. 8).

In January 1984, Judge Peckham's decision was upheld by the U.S. Court of Appeals for the Ninth Circuit, thus prohibiting the use of IQ tests to evaluate Black students for placement in special
education classes not only in California but in the school districts of Arizona, Idaho, Montana, Nevada, Oregon, and Washington State (Education Week, 1984, p. 5).

In Maryland, two psychologists and their school board were sued by parents of a child who graduated high school reading at a third-grade level. They claimed he had been improperly placed in classes for the mentally retarded. A private physician diagnosed the condition as "... dyslexia, a neurological disorder, not related to intelligence, that impairs a person's ability to read" (Education Week, 1983, p. 3). The Maryland Court of Appeals ruled in favor of the defendants but by a close 4-3 decision. The majority decision was based on the fact that there was not malicious intent and therefore the defendants were immune from "education malpractice" suits. The dissenting opinion, however, said medical, not educational, malpractice was involved and therefore recognizable in Maryland courts. The parents intend an appeal to the U.S. Supreme Court.

Socioeconomic factors are often alluded to in the cases, but they are not generally a central issue. Judge Grady in the Illinois case pointed out that most of the children in the EMR classes came from the city's poverty areas, implying that this as well as race was somehow causally involved.

The problem is international. IQ testing is being reviewed and questioned in France (Brasey, 1982; Tort, 1974). Michel Tort takes it beyond the question of interpretation. He believes strongly that neither the IQ test nor the psychologist is scientifically neutral. To the contrary, he sees the process being used ideologically to reinforce and justify the domination of the bourgeoisie over
the lower class. He views the testing situation not as an experimental but a social interaction, the significance of which is not the same for children from different social classes.

Although he expounds from an ideological bias of his own, he nevertheless poses an interesting thought that gives one pause. Given the correlation of IQ with socio-professional groups, i.e., the higher the socio-professional class, the higher the IQ scores, and given the larger numbers of children produced by the unskilled labor class, one should, he states, expect in the future a lowering of the national IQ with a debilitating effect on society as a whole. He quotes an American scientist, William Shockley, winner of the 1956 Nobel Prize, as having actually suggested that those with lower intelligence, IQ scores under 100, be encouraged to undergo vasectomies to avoid this future effect (p. 25).

This extension of the use of IQ tests may seem stretched to the point of absurdity, but it is nonetheless thought-provoking as to the many moral and ethical issues that surround intelligence testing.

Although never used as extensively as in the U.S., Brasey indicates that intelligence testing is being utilized less and less in France. One children's counselor calls the use of the Wechsler for children a "bete noire," literally, a "black beast," a bugbear, an object of fear or aversion. Some professional recruitment groups no longer use the IQ test for screening applicants. The test loses its significance in the Orient and Africa where, he notes, greater value is placed on wisdom and mastery of one's "self."

Richard Shore, a psychologist in the office of the Assistant
Secretary for Policy, Evaluation, and Research in the U.S. Department of Labor, goes so far as to call psychologists "servants of power." In stress management, workers are helped to cope and adapt to conditions at work that produce stress, using such techniques as muscle relaxation, biofeedback, and meditation. He wonders why there are not suggestions to try to change the work conditions that are so stressful and promote better health in that manner. He notes the participation of industrial psychologists in management campaigns to counter labor unions (APA Monitor, 1979, p. 2).

Various education articles deal with the need for reading and speech specialists as well to be aware of the potential bias in materials and tests that they use with children having nonstandard dialects of English. They point out how important it is to monitor the content validity of tests used, to determine the cause of incorrect responses, and to be able to justify assessments made given the problem of linguistic and cultural bias in tests. Specific tests in these areas are discussed, with dialect forms of tests or alternate scoring procedures for dialect speakers considered preferable (Condon, 1979; Harber, 1976; Wolfram, 1976, 1979; Wolfram & Christian, 1979; Wolfram, et al., 1979).

Wolfram (1979) points out the need to determine if the dialect used is a "language difference" or a "language disorder." If the speaker's dialect conforms to the language patterns of his/her community environment, then it is "different," not a "disorder." He prefers the use of "language difference" as it is a more neutral term than "dialect" which often has a negative connotation. Clinicians need to know the different forms involved, the sounds and structures,
what they are and where they occur. Another aspect of the problem is the way in which language responses are obtained during a testing situation. The form of the questions used by the examiner often differs from the natural use of language and so may be more difficult for the dialect speaker to comprehend. These admonitions are deemed appropriate for those who administer any test.

It is important to note that it is not suggested that these children need no help or should not be taught standard English, but rather that they should not be officially assessed as having inferior intelligence or disabilities requiring special placement. It becomes essential to determine who may have a true handicap and who may only need special approaches in a regular class setting to facilitate their learning standard English, a prerequisite to operating successfully in the job market.

Underlying the discussion of test bias is the issue which few appear to consider, that of the expectations and even possible bias of the examiner administering the tests and interpreting the results. Judge Peckham saw "discriminatory intent" on the part of school officials, calling IQ scores "ostensibly" neutral. Suggestions are made to correct for test bias or to use alternate scoring procedures for dialect speakers. They avoid, ignore, or are not cognizant of the fact that expectations effects and examiner bias occur and may be part of the issue.

To compound the problem, IQ scores are often accepted at face value by others not having Judge Peckham's or even Michel Tort's healthy skepticism of their neutrality or scientific objectivity. One cannot assume, as Judge Grady did, that extenuating factors will
be taken into account by anyone making qualitative judgments based on test data. Even professional teachers may accept an IQ score at face value, limiting their expectations for a student.

This brings into sharp relief the importance of understanding the role of the psychologist, his/her attributes, personality, expectations, bias, and influence in any examiner-examinee interaction.

**Experimenter or Examiner Effect**

The various factors present in any dyadic situation weave a complex web of interactions, each strand dependent on the other, making the likelihood of attributing causality of results to any one of them difficult indeed. Even in experimental research, the complexities of the interaction are such that confounding variables are ever present, complicating the analysis and interpretation of data.

Situational factors present are the experimenter's experience, whether prior to or gained during the experiment, the physical locale of the interaction, and even the personality and behavior of the "principal investigator," the person for whom an experimenter may be gathering data, which transforms a dyad into a triadic situation.

Modeling effects sometimes occur, even in psychological testing, when a subject may model his behavior after the examiner. The behavior of subjects and experimenters, their expectations of each other, differ according to their biosocial and psychosocial attributes. There has been much investigation into the interaction of the sex of each participant, producing differing results, as well as race, age, appearance, religion, anxiety level, need for approval, personality, intelligence, and perceived status.
In an early study with nondifferentiated variables, Cattell (1937) reported on results of the testing of 1300 pupils, using the Stanford-Binet Intelligence Scale (S-B), during the first nine years of the Harvard Growth Study. Marked differences were found between IQ's obtained by different examiners testing the same children at various intervals in the early and late 1920s. Most of the differences fell into a normal curve pattern but there were some unusually large variations at the extremes. Four cases varied over 40 IQ points. It was not a randomized selection of pupils, having a large number of superior pupils. Nonetheless, caution was urged in interpreting results of a single intelligence test. The main point of the study was to indicate how a child's score can fluctuate. The examiner variable was inferred but not fully explored as a major factor.

"It is significant that a problem which perplexed some of the most influential scientists of Germany in 1904 was resolved at that time, yet should contaminate psychological investigations of the present day, that is, the experimenter's influence on his subjects" (Kintz, Delprato, Mettee, Persons, & Schappe, 1965, p. 223). They refer to the case of Clever Hans, the horse which performed mathematical feats that perplexed many until intensive study determined that, although each questioner made an effort not to give any cue at all to the horse, "unconscious" cues were given, unbeknownst to the examiner but picked up by the horse.

Kintz et al. (1965) provide a historical review of studies with the experimenter as an independent variable. From experiments with rats to verbal conditioning to projective testing, the experimenter was found to be a significant variable in reporting data, the
handling of animals, the effects of his/her personality (warm or cold), and the effect of pre-experiment information on how well a group of rats would perform. In working with animals, cues are transmitted inadvertently by tactile and kinaesthetic means. With human subjects, cues are transmitted verbally and/or visually. "But verbally implies not only the words, but also the inflectional and dynamic processes of speaking" (p. 228). Tone, manner, gesture, and facial expression, all nonverbal cues, also play a part.

In the area of testing, including IQ tests, they believe it is not sufficient to know the "score"; one must also know the personality of the examiner and his/her expectations.

Sattler and Theye (1967) consider Kintz et al. unduly pessimistic. In their own review of the research up to 1966, they focus on the factors affecting results specifically in individual intelligence testing. They classify the variables involved as procedural, situational, and interpersonal, with some overlapping of areas.

Since the examiners in the present study did not administer the test involved, procedural variables are not of major concern. It is, however, interesting to note that studies on this factor "...do not appear to strongly confirm the assumption that modifying standard procedures seriously affects the overall test results" (p. 349). This puts into question the warnings in test manuals against changing any procedures.

The same appears to be true of the situational variables they reviewed in studies manipulating various testing conditions, including distraction, discouragement, anxiety, positive and negative test administration, and the use of incentives. "Generalizations concerning
the effects of situational variables on test performance must be tentative" (p. 351). Of twenty studies reviewed, findings were significant in five, mixed in three, and nonsignificant in twelve. It seems that discouragement may affect children's performance on tests. Short-term memory, as measured by digit-span subtests, may be affected by all three factors, procedural, situational, and interpersonal.

In evaluating the studies on the interpersonal variables, they conclude that the experimenter-subject relationship is important, except for the factors of race and training. "Familiarity, understanding, warmth, preference, and adjustment all play a role in altering the subject's test performance, but not always in the same direction" (p. 355).

Subject as well as experimenter variables play a role in the interaction. Attitude, level of anxiety, and warmth of subject's personality affected test scores as well, but Barbe (1965) concluded that scores were perhaps influenced "... even more by the examiner's reaction to the subject than vice versa" (p. 147).

In the studies on race as a factor, however, Sattler and Theye found that experimental procedures were not always adequate, with problems in design, sampling, and statistics. Subjects were not always randomly assigned to the experimenters. Samples were nonrepresentative, with some studies on the influence of race not even including both races (Black and White) as either examiners or subjects. They believe any study on experimenter effects should have at least two levels of the experimenter variable to be considered adequate.

They also found "... little evidence that trained experimenters differ from less trained experimenters with respect to obtaining
reliable results, being influenced by the subject, or scoring errors. Some confidence can be had in the results obtained by less qualified experimenters" (p. 355). In studies using a variety of IQ tests, the S-B, Wechsler-Bellevue Intelligence Scales (W-B), and the Wechsler Adult Intelligence Scale (WAIS), and with no specific experimenter parameter involved, they found results evenly divided, with five showing significant differences and five no significant differences in scores. In general, Sattler and Theye are far more cautious than Kintz et al. in their interpretation of results of the various studies.

Schwarz and Flanigan (1971) did a study similar to Cattell's but on a much smaller scale and with the examiner as the variable studied, a variable they believe has been ignored. Twenty-one children were divided into three groups. Each child was given three IQ tests, the WISC, WPPSI (Wechsler Preschool and Primary Scale of Intelligence), and S-B Form L-M, by three different examiners over thirty-day intervals. One of the concerns of the study was the special class placement of children. They found significant differences in the scores attributable to the examiners. Nine children tested by one examiner would have been eligible for special class placement, five tested by a second examiner, and only three tested by the third. Six "failed" with the first and "passed" with the third. These children "... should loom large in the eyes of those charged with special educational programs" (p. 265). It is not only the effect of different examiners on subjects but their evaluations of test responses and scoring as well that is crucial.

Experience as a factor in the evaluation of test responses
has been studied by various experimenters. Responses to subtest items are scored by examiners with different levels of experience. The scores are then analyzed and the amount of agreement and disagreement among examiners is determined.

Plumb and Charles (1955) had two groups of judges, twenty-four psychologists and a group of students (number unspecified), score 254 responses selected from those given by a male reformatory population to the Comprehension subtest of the W-B Form I. They found very low interscorer agreement, with very similar variations in scoring within each group. The more experienced psychologists failed to agree on 31 to 74 percent of the responses; the students on 14 to 62 percent of the responses. Not only were ambiguous responses difficult to score but experience did not appear to be of any help in their scoring.

Replicating a study which found low interclinician agreement on the scoring of WAIS Comprehension subtest responses, Schwartz (1966) found similar results, although samples were rather small, five experienced clinical psychologists and four comparatively inexperienced interns, plus one postdoctoral fellow. The experience factor again was negligible, with much similarity between the groups on the various levels of criteria for agreement.

Sattler, Winget, and Roth (1969) did a study using ambiguous WAIS and WISC Comprehension, Similarities, and Vocabulary subtest responses scored by two groups of eight clinical psychologists, each grouping a year apart. The psychologists were told the investigators' purpose was to develop a list of ambiguous test responses. Again, the percentage of interscorer agreement increased only as the
criterion for agreement was lowered, reaching 60 percent when the
criterion was agreement on six out of eight responses, 77 percent on
five out of eight, and 100 percent on four out of eight responses.
They considered these differences in scoring an indicator of test
unreliability. The sample was small, however, and so they do not,
or cannot, examine other factors which may have been involved, such
as sex or personality of the examiners. It is interesting to note
they focus on the reliability of the test and not that of the judgment
of the psychologists in scoring responses.

A 1972 study by Auffrey and Robertson used three levels of
experience. The sample consisted of thirty-six examiners, twelve in
each level of experience as determined mainly by the number of
Wechsler tests administered: experts--at least two years of experi­
ence and administration of at least fifty WAIS and fifty WISC tests;
interns--graduate students trained in Wechsler testing with at least
ten but not more than fifty Wechsler test administrations; and
novices--members of a graduate class in Wechsler tests with at least
one WISC and one WAIS administration. Each scored one WAIS and one
WISC written protocol with ambiguous responses and with a randomly
assigned biasing case history, either pessimistic or optimistic in
tone. Different case histories had no significant effect on the scor­
ing. There was, however, significant scoring variability within the
three experience levels. In the WAIS and WISC, the subtest scoring
range was higher for the novice than the expert group. Differences
between the expert and intern groups were not significant.

Bersoff (1981) and Wechsler (1966), mentioned earlier,
indicate that it is not only the administration and scoring of tests,
but the interpretation of the results that can have important effects. When estimating an examinee's intellectual potential from IQ and other test scores, psychologists commonly "...speculate upon a child's 'effective intelligence'...in the context of his age, sex, race, and socio-economic status" (Nalven, Hofmann, & Bierbryer, 1969, p. 271).

To study the extent of influence of these factors on their judgments of "true IQ" Nalven et al. sent a questionnaire randomly assigned to 650 psychologists selected at random from the 1967 APA Directory. Each questionnaire gave the same WISC pattern of subtest scores and full scale IQ of 91, but the background information given about the child obtaining the score was different, one of sixteen category combinations: age 8 or 14; Male or Female; Caucasian or Negro; Lower or Middle Socioeconomic Status (SES).

With 65 percent (approximately 422) of the questionnaires returned, the first twenty from each category were analyzed. There was a wide variation of mean "true IQ" estimates from the 320 responses used, ranging from 91.1 to 104.7. Race and socioeconomic status were found to be significantly related to the ratings, but age and sex were not. In general, those responding considered the IQ scores of 91 an underestimate of the "true IQ." The highest estimates were given to the lower class and Black categories. The psychologists evidently "...assume that lower class and Negro children's obtained WISC IQ scores represent significant underestimates of their true intellectual capacities" (p. 274). One questions, however, why the researchers used only 320 of the approximately 422 responses received from the psychologists. Those who returned the questionnaires first
may not be a truly representative sample. In addition, seeing the category stated on a questionnaire may have alerted the respondent to the variable involved and caused him/her to overcompensate for any possible bias, overt or covert. There is some question, too, about the variables surrounding the assessment of an already obtained score. They may be quite different from those found in the actual administration and/or scoring of a test.

As pointed out by Rosenthal (1976), research since 1966 has become more numerous and specific, with more complex designs in the study of the experimenter/examiner effect and interaction with subjects. In a 2 x 2 x 2 x 2 factorial design, Samuel (1977) studied the influence of the variables of atmosphere, race, and sex of tester, race and SES of subject, and tester expectation on the WISC performance of 208 Black and 208 White female secondary-school students under sixteen years of age.

Atmosphere (stressful or gamelike), expectations (pre-test information on how a subject would perform), and SES had significant effects, with high SES Blacks obtaining mean IQ's equal to or greater than low SES Whites under motivating conditions. In general, subjects obtained higher IQ's with a female tester, but more so when the tester was Black, causing a significant interaction between race and sex of tester. Something in the style or personality may have biased the results and also affected the main effect results on the sex factor. There was a total of eight examiners, four Black (two male and two female) and four White (two male, two female), all graduate students of psychology. This small sample of examiners makes any conclusion about race and sex of examiner rather tenuous.
Hanley and Barclay (1979), cited earlier, report studies indicating that children aged six to eight of both sexes score higher on intelligence measures administered by female examiners. However, those reviewed on the effects of the examiner's race and the effects of subjects' socioeconomic class have produced mixed results, some showing an effect, others not.

Both sex of examiner and sex of examinee were found to be significant in Gillingham (1970), with female examiners producing higher mean WISC scores and male examinees obtaining higher mean WISC scores. The study involved eight examiners each testing eight of sixty-four students. Expectancy effects were not significant.

Goetz (1977), concerned with the larger numbers of males in classes for exceptional students, both retarded and gifted, examined whether the sex of the child being tested affects the scoring process. Eighty practicing school and clinical psychologists, forty male and forty female, scored hypothetical WISC-R protocols of high and low IQ male and female examinees. Results showed no sex bias in scoring, with neither significant differences between the scoring by male and female psychologists, nor in the scores of the male and female protocols in either the high or low IQ condition.

The differing results obtained by Goetz may be due to a number of factors: the sample of examiners was much larger; the interaction was with written protocols rather than face-to-face with examinees; the examiners were practicing psychologists who may have had more experience and whose scoring may or may not be more reliable.

Sattler and Gwynne (1982) provide an extensive review of the literature on race-of-examiner effects. In evaluating the research
published between 1936 and 1980, they found no significant relationship between race of examiner and examinee's scores in twenty-three of twenty-seven studies utilizing a wide variety of tests, grade levels, and geographic areas. Five additional studies found various combinations of scores of Black and White children with Black and White examiners. They present the twenty-seven studies in charts clearly summarizing the results. They are particularly critical of two studies frequently cited in the research literature supporting the race-of-examiner effect, considering them inappropriate in design, with an inadequate sample of examiners, and a nonrepresentative sample of examinees.

In addition, they discount claims that Black children do not clearly understand White examiners because of dialect differences, with a consequent lowering of scores. Research studies they believe have generally not supported this view. Rather they find that Black children may be bidialectal, understanding both standard spoken English as well as Black dialect. They conclude that the popularly held belief that Black examinees perform better with Black than with White examiners is a myth.

A study by Quay (1974) reviews the contrasting views on whether or not use of Black dialect poses a problem in effective communication between children who use it and Black or White middle-class teachers and psychologists who use standard English. She observes that most discussions revolve around speech production and that "... studies of children's speech production, taken by themselves, do not permit generalizations about linguistic competence" (p. 463). The procedure used in some of the studies to assess Black children's
use of standard and nonstandard English was sentence imitation. Quay believes that this procedure gives no indication of language comprehension nor does it clarify how a child's bidialectalism might affect performance in intellectual assessments.

Some research indicates that Black children understand both standard and nonstandard English equally well but that they cannot produce standard English as well. Quay's research investigated the effect of the administration of the Stanford Binet in both standard and nonstandard dialects by one Black female examiner to 104 third- and sixth-grade "disadvantaged" Black children, boys and girls, in Philadelphia schools in ghetto areas. Fifty-two subjects from each grade were randomly selected by sex for one of the treatment conditions (standard or nonstandard English test administration).

Results indicated that the language condition had no significant effect on IQ scores at either age level or for either sex. An interesting finding was on the sentence memory task, where performance was poor in both dialect conditions. This brings into question the use of sentence imitation procedures to assess the language of Black children "... since the ability to imitate appears to be one in which they are particularly handicapped, regardless of whether their own language patterns are used" (p. 468).

Warmth of examiner has been found by some to be a crucial factor in obtaining more positive results, fostering better performance on intelligence tests. The better performance, however, may be making the examiner "warmer," or the subject who performs better may rate an examiner as "warmer" (Rosenthal, 1976, pp. 79-86).

American high-school students in Lebanon achieved higher...
WISC-R scores when examiners used positive as opposed to neutral non-verbal treatment in a study by Saish (1980). Previously Saish and Payne (1977) found that positive verbal statements had a significant effect on the WISC scores of educable mentally retarded students.

Conclusions on experimenter/examiner effect are difficult and tenuous at best, relevant to the particular experimental conditions at most, and suspect in generalizability to any large numbers of the population as a whole. Human beings are highly variable individuals, subject to change, even during an experimental situation. Therefore "... for the control of the effects of experimenter attributes, as for the control of the other effects discussed ... we must rely heavily on the process of replication" (Rosenthal, 1976, p. 111).

Biosocial and psychosocial attributes, however, are comparatively constant in experimental situations. One is either male or female, Black or White. They affect responses but are not necessary as a function of the conditions since they are constant. An attribute that is quite dependent on the specific experimental situation is the experimenter's expectancy of how the subject will respond. This is the research perhaps the most relevant to the present study.

Expectancy or Halo Effect and Self-Fulfilling Prophecy

Since any occurrence of expectancy effects in an experiment or other interaction would not be known by the public, other scientists, or perhaps even the investigator himself, "The consequences of such an expectancy bias can be quite serious... The sex of the experimenter does not change as a function of the subject's treatment condition in an experiment. The experimenter's expectancy of
how the subject will respond does . . ." (Rosenthal, 1976, p. 128).

There has been extensive experimental research into expectancy effects in the fields of clinical and experimental psychology, survey research, and the everyday world of work. From the early case of the unconscious cueing of the horse, Clever Hans, performing his arithmetical feats, to influence during Rorschach testing, to children's IQ rising when teachers expect it to, Rosenthal (1968, 1976) presents a significant review of the literature. Self-fulfilling prophecies were found to occur in various settings--industry, hospitals, the medical professions, psychotherapy, athletics, social situations, and experimental psychology. Although some studies showed no effects, Rosenthal concludes that expectations generally do affect results, bringing about self-fulfilling prophecies.

An earlier research experiment, Pygmalion in the Classroom, by Rosenthal and Jacobson (1968), has been alternately accepted with acclaim, rejected as methodologically flawed, and now accepted again. The study centers on whether a teacher's expectations of a pupil's intellectual ability affects the student's behavior in such a way as to become a self-fulfilling prophecy. The expectations arise even before meeting a pupil, from knowledge of standardized test scores, previous grades, reputation, and awareness of race and SES of the family. Results of research studies on teacher expectations were found to be mixed. However, the findings of most writers on teacher expectations in the schools generally "... are in agreement about at least one formulation: children defined as disadvantaged are expected by their teachers to be unable to learn" (p. 53).

The Pygmalion experiment took place in 1964 in a California
public elementary school of 650 students. Most of the children were from a largely lower-class but not extremely poor community. One-sixth of the pupils were Mexican, bilingual, and the only minority group. Names of 20 percent of the student population were randomly selected before the school year began. Their teachers, eighteen in number, were told that these children showed unusual potential for intellectual growth, purportedly based on a nonverbal intelligence test that had actually been administered school-wide the previous spring semester.

Retests were administered during the following school year and two years later. The differences between pre- and post-tests were analyzed according to the variables of age, ability, sex, and minority group status. Overall, the experimental group showed a gain of over 12 IQ points while the control group of children showed a gain of over 8. The lower the grade level, the greater the increase found in the "expectancy" group, with grade one showing a gain of 27.4 IQ points in the "expectancy" group vs. 12 IQ points in the control group.

There were no statistically significant differences among the three tracks (ability groupings) or between the sexes, although girls in the "expectancy" group showed greater gains in reasoning ability over girls in the control group, especially in the early grades, and boys showed greater gain in verbal ability. Minority group children gained some advantage but the difference was not significant.

Generally, then, a self-fulfilling prophecy did occur, primarily in the lower grades. "But perhaps it is only the younger children whose performance is affected by the special things the teacher says to them, the special ways in which she says them, the
way she looks, postures, and touches the children from whom she expects greater intellectual growth" (p. 83). Also, the teachers of the lower grades may have been more effective teachers, and the parents more involved in their children's school work. In addition, the testing done after two years had passed showed that the expectancy advantage had decreased slightly compared to the testing that had been done after one year.

In a replication of the study in two Midwestern schools in middle-class communities, no expectancy advantage was found. Although the boys showed some expectancy advantage, this time in gain in reasoning ability, for the girls, the control group showed greater gains. The Pygmalion effect, then, occurs but is not universally generalizable.

A classic example of the expectancy effect in industry is that known as the "Hawthorne Effect" from a series of industrial experiments in the mid-1920s at the Hawthorne Works of the Western Electric Company in the Chicago area (Isaac, 1977, p. 58; Rosenthal, 1968, pp. 164-166). The performance of workers was studied under a variety of working conditions. In the variable of level of illumination, the control groups, with no change in level of illumination, as well as the experimental groups, with various intensity levels of illumination, "... achieved remarkable and virtually identical gains in efficiency of production" (Rosenthal, p. 165).

The meanings attributed to the results of this series of experiments are varied: variables not primarily considered change agents may bring about change; any new program may bring about change; scientific prestige may make the change effective; expectations for
the effectiveness of the change by the "subjects" as well as the "instituters" may bring about the change as a result of an interpersonal expectation.

The experiment brings into question the classic "treatment" vs. "no-treatment" control group type of experiment used to show change in children's learning due to a curriculum change. The "treatment" may produce gains just from expectations, or by being an innovation, the "Hawthorne effect." Rosenthal proposes a third group, a "Hawthorne control" group, where a different type of curriculum change is introduced that does not include that of the "treatment" or "experimental" group. The change in all three groups can then be compared.

In the current research on what makes an "effective school," one of the key variables is teacher expectations. In Creating Effective Schools (1982), Brookover, Beamer, Efthim, Hathaway, Lezotte, Miller, Passalacqua, and Tornatzky review the literature on teacher expectations and point out the importance of their effect. They find that both quantity and quality of teacher instruction differ according to whether the teacher has high or low expectations for a student. Teachers provide more instruction, more time interacting, and higher quality materials to those for whom they have the higher expectations.

The research reviewed indicates that biasing factors, cues for expectations, include sex of student, race, SES, talk in the teacher's lounge, appearance, test scores and records, and oral language patterns (emphases supplied). Test scores appear to cue a "fixed ability." Negative cues arise from any nonstandard English pattern.

Teachers are predisposed to expectations by biases which arise
from various sources: societal prejudices, their own social experiences and personality traits, educational concepts in the myths and ideologies advanced in textbooks and teacher training institutions, grouping practices in the schools, and beliefs about IQ and cultural deficits. Apparently the entire process is unconscious, subtle, and therefore difficult to determine or terminate. There is often a double standard of evaluation which they illustrate with an example of how the ownership of a white Cadillac may be viewed—a virtue for the “in-group” but status-seeking or being a spendthrift for the minority group.

The amount of influence the expectations have on the students depends on the student’s age, peer group, respect for the teacher, community values, and school learning climate. Lower expectations can affect a student’s self-concept, as well as lower the parents’ expectations for him/her. Low income and/or minority students appear to be more affected than middle-class students by teacher expectations.

The effect of pre-test biasing case histories on the evaluation of children’s speech by speech clinicians was investigated by Meitus, Ringel, House, and Hotchkiss (1973). Thirty graduate speech-clinicians-in-training at Purdue University evaluated the video-taped speech of seven elementary-school-age children on scales of articulation proficiency, prognosis, and anticipated achievement in therapy. Before viewing, however, ten were given case histories with a positive bias on such factors as intelligence, family status, emotional status, medical and attitudinal history, and school success; ten given negatively biased histories; and ten no case histories.
Results indicated close agreement in all three biasing conditions on all three rating scales. This was attributed to various factors: the case histories had no effect on the students; there was some unknown variable present in the histories and/or the video presentations; or the students "... in trying to second-guess the purpose of the experiment, accepted all children into therapy, expressed positive feelings about all children and parents, etc., to please the supervising investigator" (p. 150). Although the researchers were pleased that the students all agreed on the types of articulation errors, they were concerned over the fact that all the children were routinely accepted into therapy and expected to remain in therapy for the same amount of time. There should have been some differential relationship with the severity of the speech problem. The case histories proved useless causing the researchers to question the use of case histories in the training of speech clinicians.

What the researchers fail to consider, however, is the fact that the students' discounting of the case histories may imply that they have high expectations for all prospective clients in speech therapy. This might indeed produce a "Pygmalion effect" with the resultant time in therapy as short for those with severe problems as for those with less severe difficulties.

Case histories, then, or information about a prospective student or examinee may or may not be biasing, producing expectations and an expectancy or self-fulfilling prophecy effect.

Of direct concern to the present study is whether expectancy effects or bias occur in intelligence testing. Most studies in this area are based on the effect of giving biasing information to the
examiners, prior to testing, on the examinees' achievement level, socioeconomic status, and/or intelligence. The intelligence tests generally utilized are the Wechsler scales and the Stanford-Binet, tests considered the most resistant to bias effects, particularly the Performance subtests of the Wechsler scales. Conclusions vary. Samples in many instances are small and the examiners are often graduate students in psychology rather than practicing psychologists.

Schroeder and Kleinsasser (1972) report on research which, utilizing the WISC Performance subtests of Block Design and Picture Arrangement, confirm the commonly held belief that they are resistant to bias. Their own study examines the effects of examiner bias on children's performance on the WISC Verbal subtests where bias effects have been reported previously.

Eighteen students in a graduate course in intellectual assessment were the examiners. Subjects were a homogeneous group of eighteen fourth-grade children, all in the normal range of intelligence as measured by the California Short Form Test of Mental Maturity. To the examiners, however, the children were characterized as either "bright" or "dull."

The total Verbal IQ scores were found to be significantly affected by the expectations of the examiners, with a mean difference of 6 IQ points between the "positive bias" and "negative bias" groups. The authors consider this "... well beyond the 3 IQ points given as the standard error of measurement for age ten and a half on the verbal portion of the WISC" (p. 453). They conclude that test administrators must be very aware of possible effects of examiner bias on verbal subtests.
Investigations by Sneed (1975) and Dangel (1971) produced opposite results, with the examiners apparently not affected by similar biasing information given in pretest referral information. In addition, Sneed found no differences between upper and middle-class socioeconomic status groupings of students. Again, graduate psychology students or those just completing training were utilized as examiners. In Sneed's study, six examiners tested eighteen third-graders; in Dangel's, three tested fifty-four eight- to twelve-year-old educable mentally retarded children. Several studies cited earlier (Auffrey and Robertson, 1972; Gillingham, 1970; Goetz, 1977) also found no significant expectancy effects in their investigations.

Mason (1976), in her study of sixty-one boys from grades three to six, tested by twenty-one graduate students, found that social desirability of appearance and/or behavior did not produce an expectancy effect on the examiners to bias the IQ scores. However, it was determined that the appearance and behavioral descriptions made by the examiners were not distinguishably different from each other. This brings into question whether the appearances of the examinees were discrete enough for the experiment.

The question arises as to how expectancy or halo effects, when they occur, become self-validating, a self-fulfilling prophecy. Is the bias in the administration or the scoring of the test? Is it the manner in which the questions are asked or in the way the responses are scored?

A study by Babad, Mann, and Mar-Hayim (1975) specifically controlled for the administration effect by having eighteen graduate students in educational psychology score a written WISC record of an
Israeli fifth-grader. The record was the same for all; however, a cover sheet indicated it was either that of an underachieving disadvantaged child or of a high-achieving, upper-middle-class child. In all the comparisons of means made, the mean of the high-expectancy group was higher than that of the low-expectancy group. There were significant differences in the Comprehension subtest and the total verbal sum of scores.

"Administering an individual intelligence test is somewhat analogous to conducting an experiment. Prior information about the examinee's ability level and classroom performance, and impressions resulting from the information obtained from the responses to early test questions, may lead E to formulate a hypothesis, albeit vague, regarding the examinee's level of intelligence" (Sattler, Hillix, & Neher, 1970, p. 172).

In two experiments using graduate students as examiners, the investigators hypothesized that pre-test information would produce a halo effect on the scoring of test responses. In one, fifteen graduate students in psychology or counseling and guidance scored protocols with written responses to the Comprehension and Vocabulary subtests of the WISC Verbal scale. The students were randomly assigned to one of three intellectual levels inferred from scores the examinee had supposedly received on other Verbal subtests and provided in the instructions to the students. The levels equated to bright, dull, or unspecified level of ability. There was a significant difference in the scoring of Vocabulary but not Comprehension responses, with the bright level scored significantly higher.

In the second experiment, each of eight graduate students in
the same majors scored taped responses to the verbal scale subtests of the WAIS. Two accomplices, female, each taped two scripts. The responses of one script were designed to be considered superior and the other, average. Ambiguous responses were also included. The same male examiner was used on all the tapes. Only the scores for the thirteen ambiguous responses were analyzed. The "superior" examinee was given significantly higher scores than the "average" examinee for the same ambiguous responses.

The researchers indicate that these results support the hypothesis that the halo effect occurs in the scoring of intelligence test responses. Even though the scoring of taped responses is not the same as a standard test administration, it may produce a stronger result since fewer cues are provided: no pretest information on examinee's ability level, SES, or personality; scorer is more wary, being in an experimental situation; there is no face-to-face interaction. "We are hypothesizing that in actual, nonexperimental test administrations, the chances for the operation of halo effect may be increased . . . the extent of the halo effect may depend on the responses, the tests, the method of inducing the halo effect, and the method of presenting the responses" (pp. 175, 176).

Although there was no specific written pre-test information on the ability level of the taped examinees, the responses were "designed" to be superior or average. The content of the nonambiguous responses, then, produced the expectancy or halo effect with the consequent biased scoring of the ambiguous responses. What might be the effect if the responses were exactly the same but the manner in which they were said was different? What might be the result if, for
example, nonstandard English were used in one set of similar responses?

**Expectancy or Halo Effect and Spoken Language**

In Greek mythology, Pygmalion, king of Cyprus as well as a sculptor, carved a statue of a woman with which he fell in love. The goddess Aphrodite brought the statue to life in answer to his prayers. She was called Galatea. Although fulfilled by a third party, it was a type of expectation effect, a wish fulfilled. In the modern play, *Pygmalion*, George Bernard Shaw transposes this theme to the story of a Cockney flower girl who is transformed into a "lady" by a professor who trains her to use the speech, dress, and mannerisms typical of the high socioeconomic class in England. The expectations are fulfilled; society judges and accepts her on first impression as a duchess, a perfect example of the halo effect.

Although the phrase "Pygmalion effect" was used by Rosenthal (1968) to characterize the effect of teachers' expectations on their students, it generally signifies "... the attitudes which language characteristics may elicit in listeners" (Williams, Whitehead, & Miller, 1971, p. 166). Research in this area has involved two aspects of language, the effect of a speaker's tape-recorded language (English or French, for example) on listeners' "... attitudes about the speaker's personality, intelligence, social status and the like" (ibid), or the effect of socially stratified characteristics of a language or dialect, such as pronunciation of a particular sound, on listeners' judgments. The research indicates that these judgments are an essential part of social stereotyping.

In their *Dialogue on Dialects*, Wolfram and Christian (1979)
maintain that everyone speaks some dialectal form of a language, learned according to the region of the country in which the person was raised. Certain speech sounds are articulated differently, or even omitted—the Boston /r/ for example. Within the norms of that community it is perfectly correct. Vocabulary may be different as well. The word for a large sandwich changes from "hero" to "submarine" to "hoagie" depending on the region of the country.

Wolfram believes the term "dialect" should be a neutral one, indicating only a "difference" in pronunciation and vocabulary, a "difference" in the language systems. One is not better than another, merely different. Evaluation of a dialect as "standard," therefore "correct," or "nonstandard," therefore "incorrect," is not linguistically based but socially, determined by what the mainstream of society considers acceptable, usually that used by the educated middle class. This is the pronunciation generally found in the dictionary. Those who consider the speakers of nonstandard dialects as having a "deficit" believe these speakers have not learned the proper use of the language. They view them as having both a social and cognitive handicap.

It is this implication of a cognitive handicap that can have such deleterious effects. Nathanson (1975) cites court cases involving discrimination not only against disadvantaged students in Washington, D.C., schools but adults in the workplace in various parts of the country. Aptitude and placement tests require a command of standard English. Anyone without these verbal skills would be at a disadvantage and "... often assumed to have inferior intelligence ..." (p. 53).
Walter E. Douglas, in his article printed in the Michigan State University Fact Sheet (1979), finds that people who speak English with a French, Spanish, Japanese, or other accent considered "exotic" by Western society are "... listened to with delight. Nobody assumes that an individual who speaks with such accents is not intelligent." He proceeds to discuss the Ann Arbor, Michigan, "Black English Case," as it became known, where teacher attitude was a prime factor.

In 1977, a group of Black parents, four families with fifteen children in King Elementary School, concerned about their children's progress in school, took the Ann Arbor school board to court. U.S. District Court Judge Joiner ordered the school board to "... devise a teaching plan that would take the Black English dialect into account." The plan developed was to provide teachers at the particular school involved with an in-service of twenty hour-long sessions on Black English so that they would be able to recognize and understand it, know the difference between speech patterns and pathology, and, most importantly, not view it as an indication of lower intelligence or a child's inability to learn.

A more detailed description of the issues and the opposing points of view in the case is presented in The Ann Arbor Black English Case (Brown, 1980). Many of the plaintiffs' children had been placed in special education programs for the learning disabled, emotionally impaired, or speech handicapped; others were retained or suspended. None had progressed academically.

The case was based on two theories: the impact of poverty on educational achievement and denial of equal educational opportunity.
Plaintiffs claimed that the Michigan Special Education Code of 1973 was not designed for problems stemming from social, economic, and cultural factors. Since these factors were not addressed in the children's evaluations, and if their problems were so based, then special education programs were a misplacement for these children.

Judge Joiner's decision was two-fold. It held that federal and state authorities did not secure plaintiffs' rights not to be misplaced. This dismissed all claims but one, that based on the provision in the Equal Educational Opportunity Act of 1974 wherein states must take appropriate action to overcome language barriers which prevent equal participation by students in instructional programs. This in turn was based on the 1974 Lau case where the U.S. Supreme Court ruled that the San Francisco School District had violated the rights of Chinese-speaking children by not providing a compensatory language program to bridge the gap between the home language and the school language.

The defense in the Ann Arbor case claimed that the language barrier referred to in the EEO Act applied to children who spoke foreign languages and not to those who spoke nonstandard English.

The plaintiffs had to proceed to identify the language barrier, how it hindered participation in instructional programs, and what action the defendants had failed to take. A team of researchers for the plaintiffs addressed the issues to Judge Joiner's satisfaction in the second trial two years later in 1979. They established that Black English is a legitimate language system; its features create interference with a child's understanding and reading standard English (they must be able to "code switch," according to Dr. Geneva
Smitherman, a consultant to the plaintiffs); the teachers did not appreciate the fact that the children spoke a language acceptable in their peer community; teachers had negative attitudes about the children and their intellectual abilities (emphasis supplied), causing the children to become ashamed of their language and impeding the learning process; and Black English does not have a favorable reputation.

The court-ordered in-service for the teachers on Black English which "concluded" the case has been rated successful according to the president of the Ann Arbor Education Association. Teachers in another Ann Arbor elementary school have requested the same in-service. "The unfortunate results included undesired publicity, incorrect interpretations, and a feeling by King teachers that they were unfairly put on the spot for sins for which they were not necessarily responsible" (MEA Teacher's Voice, 1981).

Nathan Glazer (1981), Harvard University Professor of Education and Social Structure, in deploring the "... remarkable expansion of judicial control and supervision of state institutions ..." (p. 42), discusses Judge Joiner's decision as a case in point, wondering why there was even a need for judicial action. "And after all this effort of sensitization to Black English, will these children read any better, write any better, communicate more effectively, do better at understanding and filling out forms? It is an exercise in futility ..." (p. 52). No one, he points out, really knows how to remedy this kind of problem and any judge involved in this type of litigation should so state.

What Glazer perhaps fails to consider is the factor of teacher expectations, shown by Rosenthal (1976) and Brookover et al. (1982),
to name a few of the researchers in the field, as having a definite
effect on students, their learning and achievement. It is conceivable
that understanding more about Black dialect may bring about a more
positive expectation effect with all that implies. As William Labov
(1970) of Columbia University in New York City points out, "... An
understanding of this nonstandard language is a necessary first step
in understanding one's students and achieving the basic goals of
education" (p. 5). He proceeds to describe the structural differences
between standard English and the Black dialect, the different styles
of speaking, and the various linguistic rules (sound systems, for
example). He discusses the implications for education, applies the
knowledge to the teaching of reading in particular, and offers
suggestions for more accurate testing.

It is not only the nonstandard English of Black dialect that
affects attitudes, however. As Thomas Pietras, the consultant for the
Ann Arbor School District, pointed out in that case, "... whatever
is said about black vernacular English can be said about other variant
forms of the English language ... rural Appalachian, poor southern
white, Polish-American, Italian-American, Hispanic-American, or work­
ing class English--that language identified with laborers who use
double negatives, substitute the accusative case of pronouns for the
nominative, and use singular forms of "be" with plural subjects. ... The term working class, as applied to language, is usually a
pejorative term; it more often than not carries a negative connota­
tion" (Brown, 1980, p. 61).

Granger, Matthews, Quay, and Verner (1977) report a number of
research studies indicating that teachers consider the speakers
of nonstandard dialects to be less adequate in their speech than speakers of standard English. Not all the studies controlled for task content, however, as Granger did, with only the patterns of speech varying. Teacher evaluations of the taped recordings of middle and lower socioeconomic status Black and White children showed them to be negatively influenced by lower-class White as well as lower-class Black speech patterns.

Nathanson (1975), cited earlier, is of the opinion that in America those who use nonstandard English are generally looked upon as socially inferior, exemplified by how those in the North view migrants from the South, marked as inferior by their speech. He discusses the Cockney dialect in England as a parallel example.

Nonstandard speech patterns, then, may not only cue lower intelligence, but the perception of socioeconomic class may be implicated in the assessment as well.

Labov (1964, 1966) did a now classic study of the phonological variants in the language behavior of a stratified sample of residents of the Lower East Side of New York City. The variants were shown to be stratified by social class. The social status of the subjects was determined by the indicators of occupation (of the breadwinner), education (of the informant), and family income. It was determined that "... a person's own occupation is more closely correlated with his linguistic behavior ... than any other single social characteristic" (1966, p. 65).

The rankings of sub-groups of the sample on a scale of social stratification correlated with their rankings on their differential use of various sounds. The speech forms used by working class New
Yorkers were not those used by middle- or upper middle-class speakers. Use of the /r/ sound was one variable studied. The traditional New York City speech was /r/-less in final and preconsonantal position (e.g., gua/r/d), a characteristic which had become a prestige pronunciation. Recent years saw the imposition of a new prestige pattern—an "r-pronouncing" dialect.

As one illustration, salespeople in three department stores were interviewed noting their use of the sound of /r/. Those who worked in the store with the highest prestige used the highest (i.e., standard) value of /r/ with emphatic pronunciation; those in the store with lowest prestige, the lowest value, some or no pronunciation of /r/. There were many more Black employees in the latter store which contributed to the results. Another variable studied, the use of /th/, produced the same results for the same reasons.

Although all the salespeople in all the stores would have been considered of the same socioeconomic class on some rating scales, those with the more standard speech patterns working in the highest prestige store indicated that they enjoyed their connection with the prestige of the clients using the store as well as with the store itself, even accepting lower wages to be so affiliated. Use of the more standard pronunciation was evidently considered more prestigious and therefore emulation was desirable. (They probably would not have been hired if their speech had been nonstandard.)

The research studied the socially differentiated usage of five phonological variables, /r/ as in car or morning, /th/ as in thing, /dh/ as in that, /eh/ as in man, and /oh/ as in dog. Results indicated that the "... variant that is used by most New Yorkers
in formal styles is also the variant that is used most often in all styles by speakers who are ranked higher on an objective socio-economic scale" (p. 405). The more standard the speech, the higher the socioeconomic status ranking accorded by the listener.

The study also provided "... evidence of the linguistic self-hatred which marks the average New Yorker" (p. 480) who misperceives his own speech patterns, seeing them in terms of what he aims at rather than the reality. The aim, especially by the lower middle class, is to attain the patterns considered prestige pronunciations.

A study done in London by Bernstein (1973) explores not dialect or "standardness" but the "... entirely different modes of speech found within the middle-class and lower working-class ..." with "... different emphases ... placed on language potential. Once this emphasis ... is placed, then the resulting forms of language use progressively orient the speaker to distinct and different types of relationships to objects and persons, irrespective of the level of measured intelligence (emphasis supplied)" (p. 271). He suggests that the speech mode of the middle class facilitates the organization of experience in abstract concepts, whereas that of the lower working class is oriented to descriptive rather than abstract concepts.

The performance of two socially different groups are compared on two tests, the Raven's Progressive Matrices Test, a nonverbal measure of intelligence, and the Mill-Hill Vocabulary Scale Form I Senior. Knowledge of vocabulary is purported to be a major factor in overall intelligence. The working-class group of sixty-one boys, ages fifteen to eighteen, were all messenger boys, none of whom had
a grammar-school education. The other group, matched for age and sex with the first group, attended one of six major public schools (the term "public school" in England being the equivalent of the American private school).

The mean scores of the public-school group were closely matched at all ranges of both tests. However, for those in the working-class group in the higher ranges of the Raven's test, the mean language scores on the vocabulary test were depressed. Extrapolated for a mean age of sixteen years, there was a difference of 8 to 10 IQ points between the means for the two groups on the Raven's test, but a 23 to 24 IQ point difference for the vocabulary means. Bernstein deduces from this that "... the mode of expression of intelligence, in particular the general factor 'g', may well be a matter of learning: in particular, the early learning of speech forms" (p. 276). Vocabulary scores, then, generally considered an indicator of "g," general intelligence, may merely be a way of differentiating between two modes of using speech, with the mode of the lower working class limiting the "... type of stimuli to which the child learns to respond" (p. 276).

Since most intelligence tests as well as school achievement tests are largely verbal, it is no wonder that the lower working class, i.e., lower socioeconomic group, may not in general fare well on these estimates of ability level, reinforcing the negative attitudes and perception of intelligence already cued by nonstandard pronunciation, articulation, and/or language usage.

In any discussion of the effects of standard and nonstandard dialects, the variable of the listener needs to be considered.
Telephone engineers, principally from the Bell Telephone Laboratories, investigated speech reception for the evaluation of telephone equipment (Licklider & Miller, 1951). They considered the perception of speech a psychological problem and so their main concern was the analysis of intelligibility of speech. Some of their findings relate to the present discussion.

Listening through headphones or loudspeakers, a group of listeners heard an announcer read lists of syllables, words, and sentences. The percentage of items recorded correctly by them was the articulation score. The results showed that the sounds of /v/, /f/, and /th/ account for more than half the phonetic mistakes made in listening to ordinary conversation; consonants are harder to recognize correctly than vowels; two of the factors which affect a speaker's intelligibility are the strength and precision of the consonants and the dialect; communication can be improved more by training the speaker than by developing better microphones.

It is conceivable, then, that one of the factors in the lower ratings given those with nonstandard dialect is the difficulty that the listeners may have in recognizing some of the sounds, the "phonetic mistakes" made by the listener when receiving a message from a speaker who is using a "dialect." This applies to listeners of all levels of intelligence as "... correlations between listening ability and intelligence, memory span, and speaking ability are all quite low" (p. 1068). The speech would have to be severely distorted, however, to obliterate the message completely. In addition to what the speaker says, clues are provided in what he/she might say.

In discussing studies on judgments of personality based upon
hearing a speaker's voice, for example, the researchers found that "Stereotypes are common: Not all the listener's information depends on the talker's choice of words . . . he perceives and evaluates vocal information that does not appear in the results of an articulation test" (p. 1070). What the person is perceived to be, then, may be as important as the message itself. To paraphrase McLuhan, "the person (medium) is the message."

A number of studies have investigated listeners' evaluations of and reactions to spoken language. In general, subjects are given various types of rating scales on which to rank speech samples, usually recorded, on various characteristics, such as socioeconomic status, ethnicity, desirability, and/or personality, including intelligence. Williams (1970) believed that a further step in the research area was needed, ". . . to link whatever language and speech features serve as salient cues in this judgmental process with whatever kinds of evaluations or stereotypes are of interest to us in the behavior of listeners" (p. 473). This was necessary in order to make ". . . inferences about mediational processes--i.e., what goes on covertly in the individual between his perceiving the stimulus and his making a required decision about the scale marking" (p. 474).

Investigating teacher judgments, Williams explored the usefulness of twenty-two scales for predicting social status and ethnicity evaluations of tapes of forty Black and White children by thirty-three Black and White teachers. The thirty-three teachers, mostly female, were from inner-city Chicago schools attending a summer institute in speech and language. It was an attempt to determine which speech
characteristics cue the psychological perception of social status and ethnicity.

The tapes used were those of fifth- and sixth-grade children from the taped samples of the Detroit dialect study by the Institute of Research on Poverty at the University of Wisconsin. Black and White children were matched by sex and socioeconomic status. Their SES was based on the education, occupation and residence of the family. The ethnicity-standardness characteristics measured included deviations in verb construction, pronominal apposition, and the articulation of various phonemes--final /s/, /z/, /t/, /d/, /ing/ sounds and medial and final /n/, /m/, /th/, as in thin, and /th/ as in thin.

A factor analysis of the scales resulted in the emergence of two major factors, "confidence-eagerness" and "ethnicity-standardness." These two dimensions included a total of twelve scales. "Confidence-eagerness" included: child reticent/eager to speak; sounds confident/unsure; uses language effectively/ineffectively. The "nonstandardness-ethnicity" scale included: language shows standard American style/marked ethnic style; pronunciation is nonstandard/standard; child's speech indicates poor educational background/a good background. The author warns that this model may not be generalizable beyond their own samples.

Results indicated stereotyping behavior on the part of the White teachers as a group, with some individual exceptions. As an example, of nine children rated as high status and White by the White teachers, six were actually Black children. For White teachers, "... sounding white is equated with high status ... speech cues
may elicit some type of general personality, cultural, or ethnic stereotype, and most of a teacher's judgments draw from this stereotype rather than from the continuous and detailed variety of input cues . . . and are part of the dynamics of teacher-pupil behaviors" (p. 486). Williams hastens to add that this may not indicate "pur­
pose" bias on the part of the White teachers, but rather "... a reflection of a stereotype of pupil language which more strongly equates standardness with linguistic effectiveness and social status than would be found with the Negro teachers" (p. 487). This may be true but does not address the other part of the bias, equating these elements with being White. The assumption is made, which is open to question, that Black teachers, because of their own language experiences both in childhood and education are "sensitized" to Black and White styles of speech and better able to determine level of effect­
iveness and status in each style.

Naremore's study (1971) expanded on the Williams' research, focusing on the individual teacher differences rather than the scales and the correlation of the ratings of linguistic features of the speech samples with evaluations of social status and ethnicity by teacher groups (Black or White).

Factor analysis grouped the teachers into four types, three including both Black and White teachers, one including only White teachers. Teachers found in each type differed in overall experience, experience in inner-city schools, training, sex, religion, marital status, and race. The only characteristic which differentiated the types was race. Black teachers did not consistently rate Black children above White children, whereas White teachers in two of the types
did show racial bias, rating White children consistently above Black children.

The only characteristic in the scales which differentiated the teacher types was their ratings of the children on the pronunciation/standardness scale. One type of teacher overcompensated, rating children whose performance was actually low (low-status children) above children whose performance was actually high (high-status children). A previous study by Williams and Naremore had found that "... the high status children generally exceeded the performance of low status children on both syntactic and functional measures of language" (p. 23). The characteristic of pronunciation deviations and pausal phenomena in a child's speech had the highest correlation with the ratings on the family status and pronunciation/standardness scales.

In a related study, Williams et al. (1971) explored whether visual cues of a child's ethnicity would influence the ratings of the same audiotaped speech samples of the two previous studies. Four video-tapes of fifth- and sixth-grade boys, one Black, one Mexican-American, and two White, each representative of one of three central Texas dialect groups, were selected. For the Black and Mexican-American videotapes, different audiotracks were used, an "ethnic" non-standard English version and a standard English version.

Subjects were forty-four undergraduate education majors attending a speech course for future elementary teachers at the University of Texas. Each viewed three videotapes: one of the Black or Mexican-American child "using" standard English and one "using" nonstandard English, and one of a White child using standard English.
"Ethnic guise" variations of the standard English were also included. Using the two-factor scales of "confidence/eagerness" and "ethnicity/standardness" validated in his previous research, the results for the standard English tapes showed that those paired with the videotapes of the Mexican-American child were rated significantly lower on "confidence/eagerness" than those paired with the Black or White children, with no differences between the latter two. On the "ethnic guise" variations of the standard English tapes, those paired with the Black and the Mexican-American children were also rated more ethnic and nonstandard than when paired with a White child.

The conclusion was that . . . the videotaped image showing the child's ethnicity affects ratings of his language in the direction of racial stereotyping expectations . . . the visual cues of ethnicity either biased the perception of the language cues or simply made them irrelevant to the evaluator" (p. 170). The 1972 study by Cline et al generalized this effect to visual images other than race.

Studies by Lambert, Hodgson, Gardner, and Fillenbaum (1960); Anisfeld, Bogo, and Lambert (1962); and Strongman and Woosley (1967) investigated reactions to various spoken languages based on the rationale that ". . . spoken language is an identifying feature of members of a national or cultural group and any listener's attitude toward members of a particular group should generalize to the language they use" and ". . . hearing the language is likely to arouse mainly generalized or stereotyped (emphasis supplied) characteristics of the group" (Lambert, 1960, p. 44).

Lambert et al. (1960) reported ratings on fourteen personality traits by French-speaking and English-speaking students in Montreal,
Canada. They listened to tapes recorded by four male bilinguals, each reading the same prose passage in English and in French. The subjects did not know they were the same voices. The ratings were correlated with measures of attitude toward French and English Canadians. As expected, the English students rated the tapes in English more favorably than those in French, including the trait of "intelligence." The French students rated the English tapes more favorably as well, but they also rated the French tapes less favorably than did the English students, which was "... interpreted as evidence for a minority group reaction on the part of the French sample" (p. 51). Only the speaker who spoke French with a Parisian accent was not downgraded by either group, indicating they may differentiate between Parisian and Canadian French.

The "minority group reaction" is reminiscent of Labov's observation on the "linguistic self-hatred" of the average New Yorker. It would appear that osmosis of the stereotyped attitude toward the spoken language of one's own cultural group has an influence not only on the listener but on the speaker of that language as well. It has a circular effect—the stereotype of the group members is applied to the spoken language and the stereotype of the spoken language is applied to the group members.

Anisfeld et al. used Lambert's design to determine differences in reaction by Gentile and Jewish subjects to "pure" English and English spoken with a Jewish accent. The subjects were undergraduate second- and third-year psychology students at McGill University in Canada. The speakers were all Jewish, taping the passages in both
styles of spoken English. Their English had been objectively judged as having "pure" English expression.

The findings showed that "... Jewish subjects categorized many more voices as Jewish than did the gentile subjects ... likely due to the greater sensitivity of Jewish subjects to cues of Jewishness in the voices" (p. 230). This is somewhat similar to the findings by Naremore that Black teachers, because of their background and education, were more familiar with Black dialect and thus more able, and willing, to perceive the speech of Black children of high status as high status. Although the Jewish-accented speech was rated lower by both groups of subjects on the traits of height, good looks, and leadership, the Jewish subjects rated the accented speech higher on sense of humor, entertainingness, and kindness.

An "unexpected finding" was that Jewish students had more favorable self-images than the Gentile students on the attitudinal measures. It is surmised that because of the emphasis on intellectual achievement in the Jewish community, Jewish students are accorded more prestige by their family and friends than are Gentile students by theirs with a resultant effect on self-image. This conjecture, however, may in itself be a stereotypical evaluation by the researchers or the Jewish students themselves. In both Lambert and Anisfeld there was no significant correlation between the subjects' ratings of the spoken languages and attitudinal measures toward their own and the other group. "The essential independence of evaluational reactions to spoken languages and attitudes is interpreted as a reflection of the influence of community-wide stereotypes ... " (Lambert et al., p. 51). It may be, however, that the students could have been astute.
enough to be aware of what was being looked for in the attitudinal measures and control their overt expression of attitude. The ratings indicating stereotyped reactions to the spoken language may be a truer measure of their covert attitudes.

A confounding variable in both studies may be the difficulty for the speaker reading passages in the two dialects to maintain the same "personality" in each dialect. Warmth of voice and traces of humor, for example, may change when reading in the dialect of one's own cultural group. Nevertheless, the studies appear to affirm the commonly accepted beliefs about stereotypical reactions to spoken language.

Where Lambert et al. and Anisfeld et al., as well as Williams et al. and Naremore, compared majority and minority groups, generally finding stereotyped reactions in favor of the former, Strongman and Woosley (1967) studied two equal groups. One group was from the north of England, the other from the south, matched by sex, age, and course of study. All were students attending a psychology course.

The students listened to passages tape-recorded by two speakers in both Yorkshire and London accents. There were no significant differences found in the ratings on any of the personality traits, including intelligence. Just as majority and minority groups in the same community have stereotyped views in common favoring the majority group, the already observed "minority group reaction," so "... two groups with approximately the same cultural background, neither ... being a majority or minority group, also tend to hold similar stereotyped views" (p. 167). For example, "Yorkshire" speakers were judged more honest and reliable and "London" speakers
more self-confident by both groups, even though the speakers were actually the same person. Other results favored neither group, which is not similar to the majority/minority studies. This may have been due to neither group being seen as a majority group.

If minority group status or nonstandardness is evident in a person's speech and produces negative reactions, the question arises as to whether the speech should be changed. Students in the field of speech often must correct elements in their own speech--pronunciation and the use of grammar--for example, "... to conform to the practice of the 'educated' members of the community ... a process of status dialect change" (Harms, 1961, p. 164). Harms investigated whether this was a useful goal and whether the "credibility" of a speaker was affected by a listener's assessment of his/her status or background based on speech cues.

Speakers and listeners in the study were from the area in and around Columbus, Ohio. Nine speakers made short recordings of content-free speech--responses to questions asking for such items as the time. The speakers had been objectively classified into three status groups: high status, with advanced degrees and prestige occupations; middle status, with high-school diplomas or with one year of college and middle status occupations; and low status, with an eighth-grade or less education and unskilled occupations (those with sub-standard or incorrect speech).

The listeners were 180 non-college adults classified into the same three status groups according to the same procedure, with age and sex constant. Listening, however, was "free-field" and in a variety of settings from living rooms to fire houses. Four scales
were used, two on speaker status and two on speaker credibility.

Results of the study were significant, showing listeners of all statuses correctly identifying a speaker's status from speech cues alone. They were not able to indicate, however, on what their ratings were based--articulation, vocabulary, pronunciation, grammar, voice quality, etc. In assigning credibility, the same distinctions were made, with high-status speakers being rated the most credible, low-status the least credible. The listeners indicated that both judgments, status and credibility, were made within ten or fifteen seconds of hearing the speech. "It may be that a listener notices pronunciation and other stereotyped features most readily after he responds to some yet-to-be identified microscopic speech cues" (p. 168). Many words can be said in ten seconds, however, with a number of speech characteristics as cues--articulation, pronunciation, grammar, etc.

It is not only ratings of personality traits (including intelligence), ethnicity, and SES, then, that are affected by both speech and language cues, but the credibility of the speaker as well. The more "standard" the speech, the higher the rating of social status and credibility.

Harms points out that there may be some unknown speech factor that cues these assessments before the stereotyped features of pronunciation and/or language are consciously perceived. Expressiveness, that quality of phrasing and/or tone which conveys a feeling or thought beyond the specific meaning of the words, may be such a factor.

A research experiment, the results of which became known as the "Doctor Fox" effect, was undertaken by Naftulin, Ware, and Donelly.
(1973) to determine whether student evaluations of instructors are based on educational content or on variables in the personality of the instructor. They review the results of studies which found, variously, that listeners are influenced by a speaker's "involuntary expressive behavior," audiences by the person introducing a lecturer, learners by an educator's "genuineness," and secondary students by a teacher's "charisma" or "popularity," which creates a "halo effect."

In the Doctor Fox study, the researchers chose a distinguished-looking actor who sounded "authoritative," gave him the title of Dr. Myron L. Fox, an "authority" on the application of mathematics to human behavior, and programmed him "... to teach charismatically and non-substantively on a topic about which he knew nothing" (p. 630). The topic was "Mathematical Game Theory as Applied to Physician Education"; the content was "... irrelevant, conflicting, and meaningless" (p. 631).

The one-hour lecture was presented to a group of eleven psychiatrists, psychologists, and social-work educators at a teacher-training conference and was followed by a half-hour discussion period. Two other groups of subjects, one a similar group of eleven, and a third consisting of thirty-three educators in a graduate course in educational philosophy, saw videotapes of the introduction, lecture, and discussion. All fifty-five subjects responded anonymously to a "satisfaction" questionnaire with significantly more favorable than unfavorable responses. None detected the hoax.

The results were difficult to generalize since the topic was quite abstract and may not be applicable to more concrete topics. However, the result "... suggests that for these learners "style"
was more influential than "content" in providing learner satisfaction" (p. 634). The researchers suggest the possibility of having actors deliver legitimate lectures as an educational innovation.

Overall and Marsh (1982) report on replications of the Doctor Fox study which conclude that "... differences in content consistently explains (sic) much less variance in students' overall evaluations than do differences in expressiveness" and "... in manipulating instructor expressiveness, only ratings on the Enthusiasm dimension were affected" (p. 11). Expressiveness, however, may include speech characteristics since "Doctor Fox" undoubtedly used the standard speech of the educated class.

Enthusiasm of expression, then, may be a factor to be considered in any discussion of speech cues, along with articulation, pronunciation, and use of vocabulary and grammar.

One study which specifically addressed the topic of evaluation of intelligence as perceived from speech cues was Fay and Middleton's 1940 study "Judgment of Intelligence from the Voice as Transmitted over a Public Address System." An audition was held from which twelve speakers, high-school juniors and seniors (ranging in age from fifteen years eight months to nineteen years three months) were chosen. All were self-confident, had oral-reading ability, and had no speech defects. They were divided into three intelligence groups with mean IQ's of 126, 99, and 86.5 as assessed by the Terman Group Test of Mental Ability.

The speakers were neither known nor seen by the listeners, eighty-four psychology students (fifty-six male and twenty-eight female). Each speaker was heard twice, unbeknownst to the listeners.
who were told they would hear twenty-four speakers, each reading a short thirty-second passage, to be rated on a seven-point scale for intelligence.

Results indicated that there were no significant sex differences in the ratings. The reliability of the ratings, however, showed considerable differences, ranging from .13 to .60, with only fair reliability for all the speakers together. Of the three speakers with the highest and lowest mean ratings and with stronger reliability coefficients of .58 to .60, the one with the highest rating had the highest IQ, the one with the next to lowest had the lowest IQ, but one whose IQ was average (100) was given the lowest mean rating of intelligence and had the highest reliability coefficient of .60. "His voice quite clearly seems to be a low-intelligence stereotype. Possibly all the ratings that have relatively high reliabilities indicate voice stereotypes. . . . Our data do seem to prove the existence of voice stereotypes of superior and of inferior intelligence" (pp. 190, 191).

There was no discussion, however, in this study of the elements that make up "voice," some of which may have been different in each of the speakers. Speech patterns as well seem to be part of the overall term of "voice" as used in this study. It is known that all were self-confident, had "oral-reading ability," and no speech defects, but it is not known if there were differences in voice quality, expressiveness, enthusiasm, or even preciseness of articulation or pronunciation which may have cued a confounding variable of social status affecting the perception of intelligence level. The fact, however, that even one speaker with average intelligence was
given the lowest mean rating of intelligence is an indicator of the
unreliability of judging intelligence from "voice," regardless of the
elements in its makeup. Later studies already reviewed investigated
more specific factors in speech and language as cues producing a
variety of stereotypical evaluations by listeners.

Summary

The various hypotheses set forth in the literature suggest
that a number of factors may be implicated in any dyadic relationship
including any examiner-examinee interaction. Whether or not there
is bias in test content itself is a controversial question. In any
case, the examiner must at least be aware of it, or, better yet,
correct for it. Interaction of biosocial and psychosocial attributes
such as sex, race, and personality of both the examiner and examinee,
as well as atmosphere and verbal reinforcement, may all play a role.

The expectancy effect as a significant factor influencing
results, even in scientifically objective experiments, as well as
teacher-student relationships is extensively documented. Many resear­
chers have investigated the biasing effect on an examiner of previous
knowledge of an examinee's socioeconomic status, achievement level,
intelligence, appearance, and/or use of nonstandard English, with
mixed results, especially in IQ testing. The question is raised in
IQ testing as to whether the bias occurs in the administration or the
scoring of the test, or both. Research in this area generally
involved graduate students of psychology as examiners, with pre-IQ
test referral information containing the biasing information.

Speech and language were found to affect listeners'
evaluations of a speaker. Such factors as standardness of pronunciation, articulation, language usage, language and dialect differences, ethnicity, expressiveness, and enthusiasm of a speaker affected a listener's perception and evaluation of a speaker on such characteristics as ability, intelligence, social status, ethnicity, personality traits (such as sense of humor), teaching effectiveness, and credibility. The speech and language cues generally produced stereotypical reactions based on what appear to be community-held stereotypes. Stereotyped attitudes toward a group generalize to the language they speak, and hearing that language generates the stereotyped reaction.

As was pointed out, perception of speech is a psychological phenomenon. Its evaluation develops in the mediational process that occurs covertly in the moment between hearing the speech stimulus and making the judgmental decision. It produces a "halo effect" which is difficult to determine because of its subtlety and for which it is even more difficult to infer a particular cause or make generalizations of the results since all the variables cannot be controlled.

Phonological variations in speech were found to be socially stratified. Deviations from the speech of the educated middle class, commonly accepted as the "standard," cued lower social status and the concomitant judgment of lower ability for not having learned "proper" or "prestigious" speech. A consequent loss of credibility in the speaker occurred as well. The negative evaluations were found to be especially true of any nonstandard speech used by minority groups with the negative assessments given by members of both the majority and minority groups of the community, the latter reaction referred to as "linguistic self-hatred."
The question which no one appears to have investigated specifically, however, is whether an examiner expectancy effect or bias occurs when confronted with an examinee with a speech handicap in a testing situation. As soon as the examinee speaks, does a speech handicap interfere with objectivity, trigger an expectation or bias in the covert mediational process, producing a "halo effect" which affects the administration and/or scoring of the IQ test? The present study investigated whether an examinee's speech handicap does affect the scoring of a WISC-R subtest by psychologists and graduate students of psychology.
CHAPTER III

METHODOLOGY

Type of Research

This research utilized an experimental design. It investigated the effect of an examinee's specific speech handicap on the scoring by trained examiners of audiotaped responses to an intelligence test. The research hypothesis, based on the concepts of the "halo effect" and "self-fulfilling prophecy," explored whether a speech handicap has a biasing effect on the scoring of an intelligence test. A second hypothesis explored whether there is a difference between scoring by practicing psychologists as opposed to graduate psychology students. If this proved to be the case, it would indicate a need for further research into this area as a separate study. The first hypothesis was the one of major concern to this study.

Selection of Subjects

The sample for this study was sixty-four examiners. There were two groups of examiners, thirty-two practicing psychologists and thirty-two graduate students of psychology. Two types of responses, of identical content, were recorded by two examinees, one a child with a speech handicap of multi-articulation errors, and the other a child with no speech handicap. Both were female, aged eight years eleven months. All persons were drawn from the "Michiana" area, an informal yet widely used designation for a geographical area comprising the
southwestern portion of the state of Michigan and the north central part of Indiana. The graduate psychology students were from area universities that offer graduate studies in psychology.

Each of the sixty-four examiners had the required training in the administration and scoring of the WISC-R (or WISC). As indicated previously (p. 8), this training is highly standardized throughout the United States, and detailed instructions are provided in the test manual for both administration and scoring. In order to be selected for participation in the study, the psychologists had to have practiced psychology for at least one year after receiving their graduate degree and administered the WISC-R (or WISC) in their practice a minimum of ten times, at least one of which was during the year immediately prior to participation in this experiment.

**Instrumentation and Test Reliability**

The subtest used was the Comprehension subtest of the WISC-R. For the standardization of the WISC-R, a stratified sampling plan was utilized. The sample included various classes of children in the same representative proportions found in the general population according to the 1970 United States Census. The sample was comprised of 2200 children. This included 100 boys and 100 girls, in each of eleven age groups, ranging from six and one-half through sixteen and one-half years.

Standardization of the WISC-R scores was achieved by converting raw scores to scaled scores ranging from one to nineteen, with a mean of ten and a standard deviation of three. A scaled score of ten is considered the performance of the average child of the given age on the particular scale. It is comparable to a full scale IQ of
one hundred and a percentile rank of fifty. Scaled scores of seven and thirteen are one standard deviation below and above the mean, respectively. "The distributions of scaled scores obtained on the twelve tests by the children in the standardization sample correspond closely to . . . theoretical values" (Wechsler, 1974, p. 26).

The WISC-R tests two areas of intelligence, verbal ability and performance. The Verbal Scale includes the subtests of Information, Similarities, Arithmetic, Vocabulary, and Comprehension; the Performance Scale the subtests of Picture Completion, Picture Arrangement, Block Design, Object Assembly, and Coding, i.e., the perception and manipulation of objects. "This dichotomy is a way of identifying two principal modes by which human abilities express themselves" (ibid., p. 9). With no verbal component, the Performance Scale is considered stronger in its objectivity and resistance to bias.

The split-half technique was utilized for each of the Verbal subtests (odd versus even items) to obtain the reliability coefficients. The coefficients for the Verbal subtests given for age eight and one-half range from .69 to .86, with the Comprehension subtest having a reliability coefficient of .73 (ibid., p. 28). The standard errors of measurement for the Verbal subtests given for this age range from 1.15 to 1.48, with the Comprehension subtest having a standard error of measurement of 1.48. These are in scaled-score units (ibid., p. 30).

The Verbal section is considered more subject to individual judgment in scoring, with the Comprehension subtest particularly prone to bias as it is "... the least reliable and the hardest to score of all WISC subtests" (Babad et al., 1975, p. 268).
For the particular age group of eight and one-half, then, the Comprehension subtest does not have the best reliability of the Verbal subtests; it has the highest standard error of measurement. Furthermore, the Verbal subtests are considered by a number of researchers to be more vulnerable to individual judgment (see p. 8). The scoring of the subtest and the judgments made by the examiner of the responses become more crucial. Therefore, one would have to be wary of judging a child's true ability based on an obtained score on this subtest.

Schroeder and Kleinsasser (1972) indicate that of the various subtests of the WISC-R "intrinsically susceptible to bias," "The first test administered may be most prone to bias" (p. 453). Perceptions of the examinee by the examiner may be corrected in the interactive experience of subsequent subtests. Because of its vulnerability to bias, and because bias may be most evident in the first test administered of any order of subtests, the Comprehension subtest only was used for scoring purposes. The limitation to one subtest eliminates both the experience and order effects of other subtests.

There are seventeen questions on the test (Wechsler, 1974, p. 97). Each can be given 2, 1, or 0 points depending on the quality of the response as judged by the examiner who matches the response against the criteria and sample answers given in the test manual. "In some cases the scoring will be fairly obvious; in others it may be difficult. . . . In these instances he will have to use his own judgment" (ibid., p. 175). The maximum score possible for the subtest is 34 points. The raw score is then converted to a scaled score, with a standard error of measurement given for each of eleven age groups, six and a half to sixteen and a half years.
Procedures

The test was administered in person to an elementary school-aged child, an eight-year eleven-month-old girl, with no speech handicap. The test responses were recorded on a cassette audiotape. A second child of the same age and sex, but with a speech handicap of multi-articulation errors, was recorded, giving the identical responses of the non-speech-handicapped child, on a second cassette audiotape. Appendix D gives the responses. In the case of the speech-handicapped child, the multi-articulation errors are transcribed into the International Phonetic Alphabet or set apart in parentheses. Each cassette was professionally copied in order to give each examiner his or her own copy—thus there were thirty-two copies of each cassette, each one being fixed so it could not be erased. Each recording is approximately five minutes in length.

Professors of the courses in intelligence testing in the department of psychology in each university were contacted by telephone, requesting their cooperation and permission to have their students as examiners in the study. Names of those willing to participate were supplied by the professors. Psychologists working in the school systems were similarly contacted, as well as psychologists in private practice. Anonymity was guaranteed.

Appropriate personal or telephone contact was made by the experimenter with each person who agreed to participate. Instructions were given for participation in the research, and data collection procedures were determined which would best preserve anonymity for each examiner. Departmental secretaries were generally named by members of each group as contact persons for disseminating and
collecting the materials to preserve anonymity. The secretaries were
duly contacted and procedures arranged to maintain anonymity of the
participants.

To avoid any confounding pre-test expectancy effect that might
have occurred if the examiners knew the actual purpose of the study,
they were told that the purpose of the study was to determine the
feasibility and possible uses of scoring audiotaped IQ test responses.
There was no mention of the involvement of a speech handicap or that
there were two types of tapes.

To ensure consistency of technique across examiners, each
examiner was given the age and sex of the child and specific
procedural directions, oral and written, on how to listen to and score
the audiotape, as follows: (1) simulate a "live" situation by
listening to the tape alone in a private area; (2) use the copy
175-187), if needed, to check sample responses, but only while
scoring; (3) listen to only one question and its response at a time;
(4) write the response on the protocol provided; (5) if needed, listen
to the response only once more; (6) score the responses on the
protocol provided; (7) do not listen to any part of the tape again
after finishing the last question or redo any of the scoring; (8) do
not discuss the tape or the scoring with any one else at this time to
avoid any bias.

It was stressed that the procedures needed to be followed
exactly to simulate and make valid comparisons with a "live" testing
situation. Arrangements were made to be sure that each examiner knew
how to use a cassette player and would have one available when needed,
as well as a room or area where he or she could score the tape privately.

Examiners were randomly assigned to one of the two tape conditions. The psychologists were given numbers 101 to 132. A random ordering of these numbers was assigned to an alphabetized list of their names by the Chairman of the Department of Educational and Psychological Services at Andrews University. In a similar manner, a random set of numbers 201 to 232 was assigned to an alphabetized list of the graduate students. The odd numbers were then randomly designated to score the tape with the speech-handicapped responses and the even numbers the tape with the non-speech-handicapped responses. Tapes and materials needed for scoring were numbered in the same way and sealed in similarly numbered envelopes by the experimenter. Each envelope contained a letter (appendix A); the list of procedures (appendix B); the protocol for scoring; the scoring criteria; and the appropriate cassette. These envelopes were placed in larger envelopes by the secretary to the Chairman of the Department of Educational and Psychological Services at Andrews University, who addressed them according to the numbered lists of names given to her by the Chairman. The envelopes were delivered by this experimenter’s assistant to those who had been designated by the participants as contact persons.

When scoring was completed, only the numbered inner envelopes were returned to the contact persons, collected by the assistant, and delivered to the experimenter. Thus at no time did the experimenter know who was given which number, and the contact persons, including the Chairman’s secretary, did not know the contents of the
envelopes or the significance of the two groups of numbers. This preserved anonymity for the examiners, provided a means of checking for the return of the materials, and precluded the possibility of the participants learning that there were two types of tapes before they scored the responses.

The completed inner envelopes were to be returned to the contact persons within one week. Since each examiner was given his or her own tape, the actual scoring was not expected to take more than one-half hour as it was only one subtest, approximately five minutes in length, and the procedure was limited. The fact that all examiners had and scored the tapes within the same time period controlled the effects of such extraneous variables as the occurrence of historical events, maturation processes, or an experience which may affect the internal validity of a design (Campbell & Stanley, 1963).

At the end of the one-week period allotted for scoring purposes, the experimenter called the contact persons to determine if the envelopes had been returned. Those who had not yet returned the materials were contacted. Almost all envelopes were collected within one month of their distribution. A few were delayed due to various circumstances discussed in chapter 4. Letters of appreciation were sent to all (appendix C). The raw scores awarded by the examiners were personally converted to scaled scores.

Results of the study will be available to the five participants who requested them.

**Hypotheses and Statistical Analysis**

The research utilized an experimental 2 x 2 factorial design, with analysis of variance as the statistical tool, and the F-ratio
test for significance of effect, with alpha set at .05.

The two factors (independent variables) were type of response (speech-handicapped and non-speech-handicapped) and type of examiner (practicing psychologist and graduate student of psychology). The dependent variable was the scores awarded by the examiners to the responses to the WISC-R Comprehension subtest.

A two-way analysis of variance design tests three null hypotheses:

Hypothesis 1: There is no significant difference between scores awarded to speech-handicapped and non-speech-handicapped responses.

Hypothesis 2: There is no significant difference between scores awarded by practicing psychologists and graduate students of psychology.

Hypothesis 3: There is no significant interaction between type of response and type of examiner.

Summary

The research was designed to yield the data needed to fulfill the specific purpose of this study--to determine if examiners score identical responses to an intelligence-test subtest differently depending on whether the examinee responding has a speech handicap or not. It provided for further inferential analysis as to whether a speech handicap is a significant factor in an evaluative situation, whether it may be a biasing condition producing a "halo effect" and "self-fulfilling prophecy," and whether there are implications for the "cue validity" of the content of oral responses.
CHAPTER IV

RESULTS

Respondents

To obtain the thirty-two participants for each of the two groups, it was necessary to contact thirty-five practicing psychologists and forty-six graduate students of psychology. Of the psychologists, one did not meet the criteria and two did not wish to participate. Of the graduate psychology students, nine did not meet the criteria and five did not wish to participate.

Those who agreed to participate were very cooperative, many expressing the idea that they had an obligation to the field of psychology to help in any research. A number of the students indicated that the difficulties that they had encountered obtaining participants for their own research prompted their willingness to cooperate.

The packets with the research materials were delivered to all sixty-four participants within a two-day period. Within approximately three weeks after delivery, the materials had been completed and returned by all the psychologists and approximately one-half of the students. Of the latter, all but two returned theirs within the month. One was delayed another month, and another student was substituted for one who had returned to his homeland. Anonymity was maintained throughout the process.
Seven of the respondents, four psychologists and three graduate students, did not check the boxes on the procedures sheet, as requested, stating they had followed the directions as written. This may indicate that they had not done so, or they may have simply neglected to check the boxes. Another psychologist noted his/her error in listening to the tape first before reading the procedures, subsequently proceeding to follow the directions.

Fourteen respondents, five psychologists and nine students, wrote none or only a few of the taped responses on the protocol provided. The Manual cautions that on the Comprehension and two other Verbal subtests "... every effort should be made to record the child's response verbatim and legibly, since subsequent evaluation of responses to these Verbal items is frequently necessary for accurate scoring" (p. 63). There may have been a problem of time for those who did not record the responses on the protocol, or perhaps they saw no need to write the responses since they were on tape and could be listened to a second time. Additionally, psychologists experienced in testing may tend to have confidence in their handling the scoring this way.

Of those having the tape with the speech-handicapped responses, only two, one in each respondent group, noted in writing that the child needed to be referred to a speech pathologist. Three, two psychologists and one student, indicated that they had difficulty understanding some of these responses, but all three gave above average scores. This may or may not have been a compensatory action. One psychologist who gave the next to lowest total raw score of fourteen, indicated that he/she would have probed further those answers
"... not well-articulated." The speech handicap may have affected the scoring negatively in this instance.

Five of the psychologists, three having the tape with the speech-handicapped responses and two the non-speech-handicapped responses, indicated that they would have probed further after one or two of the responses, but apparently out of a need for elaboration of the content of the responses and not because of the articulation. Questions eight, nine, and ten were mentioned by two examiners each.

**Basic Data**

The raw scores awarded by each respondent for the taped responses to the seventeen Comprehension subtest questions are presented in appendix E. Each total raw score is the sum of the points given to the seventeen responses. The raw scores were personally converted to scaled scores. Overall, the total raw scores ranged from thirteen to twenty-two, which converted to a total scaled score range of nine to fifteen. Table 1 presents the range of scores given by each respondent group.

Although one never infers a Full Scale IQ from the scaled score of a single subtest, it is meaningful to consider such a comparison to assess the meaning of this range of scores. In relation to Verbal, Performance, and Full Scale IQ, as well as percentile rank, scaled scores of nine to fifteen may be equated to an IQ range of 95 to 125, or a percentile rank of 37 to 95. Given a mean of ten and a standard deviation of three, they range from one-third of a standard deviation below the mean to one and two-thirds of a standard deviation above the mean, a span of two standard deviations. They equate to an intelligence classification of anywhere from "Average" to "Superior"
TABLE 1
RANGE OF RAW SCORES AND SCALED SCORES WITHIN EACH GROUP

<table>
<thead>
<tr>
<th></th>
<th>Speech-handicapped responses</th>
<th>Non-speech-handicapped responses</th>
</tr>
</thead>
<tbody>
<tr>
<td>Psychologists</td>
<td></td>
<td></td>
</tr>
<tr>
<td>raw scores</td>
<td>13 -- 20</td>
<td>15 -- 21</td>
</tr>
<tr>
<td>scaled scores</td>
<td>9 -- 14</td>
<td>11 -- 15</td>
</tr>
<tr>
<td>Students</td>
<td></td>
<td></td>
</tr>
<tr>
<td>raw scores</td>
<td>15 -- 22</td>
<td>16 -- 20</td>
</tr>
<tr>
<td>scaled scores</td>
<td>11 -- 15</td>
<td>12 -- 14</td>
</tr>
</tbody>
</table>

(Wechsler, 1974, Tables 7 and 8, pp. 25-26). This is a surprisingly wide range of scores.

The psychologists who scored the speech-handicapped responses, gave raw scores which converted to scaled scores below eleven, the lowest of the scaled scores of the other three groups. Table 2 indicates how many in each of the groups produced each total score.

As is shown in table 2, the experienced psychologists awarded a wider range of scores overall than the graduate students as well as a wider range to three of the four sets of scores. Both groups, however, show a wider range of scores for the speech-handicapped than the non-speech-handicapped responses. This may represent the wide variety of effects a speech handicap may have on raters, ranging from a negative bias to one of compensation.

Table 3 gives the percentage of scaled scores which fell within each intelligence classification.
<table>
<thead>
<tr>
<th>RAW SCORES</th>
<th>Range of Scores</th>
</tr>
</thead>
<tbody>
<tr>
<td>13 14 15 16 17 18 19 20 21 22</td>
<td></td>
</tr>
</tbody>
</table>

**Psychologists**
- speech-handicapped responses: 1 1 1 2 5 3 2 1
- non-speech-handicapped responses: 1 1 4 4 5 1

**Students**
- speech-handicapped responses: 1 1 1 5 4 3 1
- non-speech-handicapped responses: 1 4 5 4 2 5

**Scaled Scores**
- Average: Bright: Superior
- 9 10 11: 12 13: 14 15

**Psychologists**
- speech-handicapped responses: 1 1 1 2 8 3
- non-speech-handicapped responses: 1 1 8 5 1

**Students**
- speech-handicapped responses: 1 1 6 7 1
- non-speech-handicapped responses: 1 9 6
<table>
<thead>
<tr>
<th></th>
<th>Average  (SS 9-11)</th>
<th>High Average (Bright) (SS 12-13)</th>
<th>Superior (SS 14-15)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(N) %</td>
<td>(N) %</td>
<td>(N) %</td>
</tr>
<tr>
<td>Psychologists</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>speech-handicapped</td>
<td>(3) 18.75</td>
<td>(10) 62.5</td>
<td>(3) 18.75</td>
</tr>
<tr>
<td>responses</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>non-speech-handicapped responses</td>
<td>(1) 6.25</td>
<td>(9) 56.25</td>
<td>(6) 37.50</td>
</tr>
<tr>
<td>Students</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>speech-handicapped</td>
<td>(1) 6.25</td>
<td>(7) 43.75</td>
<td>(8) 50.0</td>
</tr>
<tr>
<td>responses</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>non-speech-handicapped responses</td>
<td>(0) 0.0</td>
<td>(10) 62.5</td>
<td>(6) 37.5</td>
</tr>
</tbody>
</table>

Eighty-one percent or more of each respondent group gave raw scores which, when converted to scaled scores, placed the responses in the High Average (Bright) to Superior classification (scaled scores of twelve to fifteen). In three of the groups, no more than 6.25 percent gave scores placing the responses in the Average classification (scaled scores of nine to eleven). By contrast, almost 19 percent of the psychologists scoring the speech-handicapped responses placed the responses in the Average category. The scaled scores derived from the raw scores awarded by the psychologists yielded double the number of Superior classifications to the non-speech-handicapped responses than to the speech-handicapped responses. The difference between being classified as Average or High Average (Bright) or Superior...
can be very significant, especially in a school setting.

The means of the raw scores and their equivalent scaled scores obtained from each group are presented in tables 4 and 5. Each cell represents sixteen respondents, for a total N of 64, 32 psychologists and 32 graduate students of psychology.

### TABLE 4
MEANS OF RAW SCORES

<table>
<thead>
<tr>
<th></th>
<th>Speech-handicapped responses</th>
<th>Non-speech-handicapped responses</th>
<th>Overall Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Psychologists</strong></td>
<td>16.9375</td>
<td>17.9375</td>
<td>17.4375</td>
</tr>
<tr>
<td><strong>Students</strong></td>
<td>18.5</td>
<td>18.125</td>
<td>18.3125</td>
</tr>
<tr>
<td><strong>Overall Mean</strong></td>
<td>17.71875</td>
<td>18.03125</td>
<td></td>
</tr>
</tbody>
</table>

### TABLE 5
MEANS OF SCALED SCORES

<table>
<thead>
<tr>
<th></th>
<th>Speech-handicapped responses</th>
<th>Non-speech-handicapped responses</th>
<th>Overall Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Psychologists</strong></td>
<td>12.50</td>
<td>13.25</td>
<td>12.875</td>
</tr>
<tr>
<td><strong>Students</strong></td>
<td>13.375</td>
<td>13.3125</td>
<td>13.34375</td>
</tr>
<tr>
<td><strong>Overall Mean</strong></td>
<td>12.9375</td>
<td>13.28125</td>
<td></td>
</tr>
</tbody>
</table>

There appears to be a difference between the psychologists and the students in their scoring of the speech-handicapped responses, with the students giving them a higher mean score than the
psychologists. There appears to be some difference as well between the scoring by the psychologists of the two types of responses. The difference between the psychologists as a group and the students as a group was greater than the difference between the speech-handicapped responses as a group and the non-speech-handicapped responses as a group.

Testing the Hypotheses

To determine whether the observed differences in the means were statistically significant, a number of hypotheses were tested by analysis of variance.

Hypothesis 1: There is no significant difference between the scores awarded to speech-handicapped and non-speech-handicapped responses.

Hypothesis 2: There is no significant difference between the scores awarded by practicing psychologists and graduate students of psychology.

Hypothesis 3: There is no significant interaction between type of response and type of examiner.

The hypotheses were stated in the order of primary concern to the study. In the analysis, however, the test of Hypothesis 3 was discussed first as this was essential to the testing of the two main effects hypotheses.

Before any comparisons could be made, it was necessary to determine if there was any interaction between rater (psychologists and graduate students of psychology) and child (speech-handicapped and non-speech-handicapped responses). Individual cell means and overall raw score and scaled score means were presented in tables 4
and 5. Tables 6 and 7 present the analysis of variance for raw and scaled scores, separately.

**TABLE 6**

**ANALYSIS OF VARIANCE OF OVERALL RAW SCORES BY RATER AND CHILD**

<table>
<thead>
<tr>
<th>Source of Variation</th>
<th>Sum of Squares</th>
<th>Degrees of Freedom</th>
<th>Mean Square</th>
<th>Calculated F-Ratio</th>
<th>Probability of F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rater</td>
<td>12.2500</td>
<td>1</td>
<td>12.2500</td>
<td>5.117</td>
<td>.02732*</td>
</tr>
<tr>
<td>Child</td>
<td>1.5625</td>
<td>1</td>
<td>1.5625</td>
<td>0.653</td>
<td>.42233</td>
</tr>
<tr>
<td>Interaction</td>
<td>7.5625</td>
<td>1</td>
<td>7.5625</td>
<td>3.159</td>
<td>.08057</td>
</tr>
<tr>
<td>Residual</td>
<td>143.6250</td>
<td>60</td>
<td>2.393750</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>165.0000</td>
<td>63</td>
<td>2.6190472</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Indicates significant statistic.

**TABLE 7**

**ANALYSIS OF VARIANCE OF OVERALL SCALED SCORES BY RATER AND CHILD**

<table>
<thead>
<tr>
<th>Source of Variation</th>
<th>Sum of Squares</th>
<th>Degrees of Freedom</th>
<th>Mean Square</th>
<th>Calculated F-Ratio</th>
<th>Probability of F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rater</td>
<td>3.5156250</td>
<td>1</td>
<td>3.515625</td>
<td>3.392</td>
<td>.07046</td>
</tr>
<tr>
<td>Child</td>
<td>1.8906250</td>
<td>1</td>
<td>1.890625</td>
<td>1.824</td>
<td>.18190</td>
</tr>
<tr>
<td>Interaction</td>
<td>2.6406250</td>
<td>1</td>
<td>2.640625</td>
<td>2.548</td>
<td>.11571</td>
</tr>
<tr>
<td>Residual</td>
<td>62.187485</td>
<td>60</td>
<td>1.0364580</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>70.234360</td>
<td>63</td>
<td>1.1148310</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
The critical F-ratio with 1 and 60 degrees of freedom at the probability level of .05 is 4. Both raw-score and scaled-score tables showed no significant interaction between rater and child, and overall comparisons could be made.

As shown in tables 6 and 7, the F-ratio for child main effect was not significant and thus Hypothesis 1 could not be rejected. There was no significant difference between overall scores awarded to speech-handicapped and non-speech-handicapped responses.

The rater main effect was significant in the analysis of raw scores presented in table 6, and it approached significance in the scaled-score analysis in table 7. Hypothesis 2 was rejected with respect to raw scores, the actual scores awarded by the rater. With respect to scaled scores, the hypothesis could not be rejected, though the F-ratio was approaching significance.

Analysis of Simple Effects

The fact that there was no significant interaction indicated that main effects could be studied. Simple effects were not necessary. However, an analysis of simple effects was done to determine if there were a significant difference between each pair of means. For this purpose, the pooled mean square of all cells was used to achieve greater validity. For simple effects it is legitimate to use as error variance the pooled mean square error from the ANOVA table, provided the assumption of homogeneity of variance is upheld.

For this purpose, the mean and variance of each cell was determined. To reduce the likelihood of a Type I error in testing the assumption of homogeneity of variance, a probability level of .01
was used. Table 8 presents the mean and variance of each cell for raw and scaled scores.

TABLE 8
MEAN AND VARIANCE OF EACH CELL

<table>
<thead>
<tr>
<th></th>
<th>Speech-handicapped Responses</th>
<th>Non-speech-handicapped Responses</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>Standard Deviation</td>
</tr>
<tr>
<td>(Raw Scores)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Psychologists</td>
<td>16.9375</td>
<td>1.8428</td>
</tr>
<tr>
<td>Students</td>
<td>18.4999</td>
<td>1.6733</td>
</tr>
</tbody>
</table>

F_{max} 2.57899*

| (Scaled Scores)        |      |                   |          |      |                   |          |
| Psychologists           | 12.5000 | 1.4142           | 1.99996  | 13.25 | .9309            | .866575 |
| Students                | 13.3750 | .9574            | .916615  | 13.3125 | .6021         | .362524 |

F_{max} 5.51677

With four variances, the critical $F_{max}$ with $p = .01$ is 5.5. Homogeneity of variance of the raw scores was upheld with an $F_{max}$ of 2.6. Hence the error mean square from the analysis of variance could be used for the simple effects tests on the raw scores.

Homogeneity of variance of the scaled scores was not upheld, having an $F_{max}$ of 5.51677. Therefore, the analysis of simple effects utilized the estimate of variance of only the relevant two cells at a time.
Test of Simple Effects

Table 9 shows the results of the test of simple effects for the raw scores using the error mean square from the analysis of variance.

**TABLE 9**

**SIMPLE EFFECTS—RAW SCORE MEANS**

<table>
<thead>
<tr>
<th></th>
<th>Psychologists v. Students</th>
<th>t</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Rater)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Speech-handicapped responses</td>
<td>16.9375</td>
<td>18.4999</td>
<td>-2.856*</td>
</tr>
<tr>
<td>Non-speech-handicapped responses</td>
<td>17.9375</td>
<td>18.125</td>
<td>-0.343</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Speech-handicapped v. Non-speech-handicapped responses</th>
<th>t</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Child)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Psychologists</td>
<td>16.9375</td>
<td>17.9375</td>
<td>-1.828</td>
</tr>
<tr>
<td>Students</td>
<td>18.4999</td>
<td>18.125</td>
<td>0.685</td>
</tr>
</tbody>
</table>

Using a two-tailed t-test, the critical t with 60 degrees of freedom at the probability level of .05 is + 2. The difference between the raw scores awarded by the psychologists and the students to the speech-handicapped responses was statistically significant with a t of -2.856.

Table 10 shows the test of simple effects for the scaled scores using the estimate of variance of only the relevant two cells at a time.
<table>
<thead>
<tr>
<th></th>
<th>Psychologists v. Students</th>
<th>t</th>
<th>df</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Rater)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Speech-handicapped responses</td>
<td>12.50</td>
<td>13.375</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>-2.0494*</td>
<td>27.8</td>
<td>0.049</td>
<td></td>
</tr>
<tr>
<td>Non-speech-handicapped</td>
<td>13.25</td>
<td>13.3125</td>
<td></td>
<td></td>
</tr>
<tr>
<td>responses</td>
<td>-0.2255</td>
<td>27.11</td>
<td>0.823</td>
<td></td>
</tr>
</tbody>
</table>

|                              | Speech-handicapped         | Non-speech-handicapped | t     | df   | p   |
|------------------------------| v.                          | responses              |       |      |     |
| (Child)                      |                           |                     |       |      |     |
| Psychologists                | 12.50                     | 13.25             | -1.7719| 28  | 0.087|
| Students                     | 13.375                    | 13.3125           | 0.2210 | 30  | 0.827|

Using a two-tailed t-test, the critical t with 30 degrees of freedom is ± 2.042. The difference between the scaled scores converted from the raw scores given by the psychologists and the students to the speech-handicapped responses was statistically significant with a t of -2.0494.

Hypothesis 2 could be rejected for both raw and scaled scores as there was a statistically significant difference between the scores awarded by practicing psychologists and graduate students of psychology to the speech-handicapped responses.
CHAPTER V

SUMMARY, CONCLUSIONS, IMPLICATIONS
AND RECOMMENDATIONS

The study was based on the concept of expectancy effects, that is, the effects expectations have on the behavior of participants in any human interaction. It revolved, in particular, around the "halo effect" and the "self-fulfilling prophecy"--the influence that the initial positive or negative impression of a person has on an evaluator whose subsequent ratings of that person are in the direction of the impression (the halo), thus fulfilling the evaluator's expectation--the prophecy fulfilled by the prophet. The investigation was limited to the interaction that occurs in an intelligence-testing situation.

Chapter 5 presents a summary of the study including its purpose, an overview of the related literature, the methodology, and the summary and discussion of the findings. It also presents the conclusions of the study, implications of the findings, and recommendations for further research.

Summary

Purpose of the Study

The specific purpose of this study was to determine if examiners score identical responses to an intelligence-test subtest.
differently depending on whether the responding examinee has a speech handicap or not. It was hypothesized that there would be a significant difference in the scoring of the two types of responses, speech-handicapped and non-speech-handicapped, with lower mean scores for the speech-handicapped responses. A secondary hypothesis was that there would be a significant difference between the scoring by psychologists and graduate students of psychology.

The effect of various attributes of both the examiner and the examinee in an intelligence-test interaction has been investigated over the years, although it has not become an issue as major as that of bias in the test content itself. A variable that does not appear to have been considered in this body of research is that of an examinee's speech handicap and the effect it might have on an examiner's scoring of test responses.

Overview of Related Literature

The studies reported in the literature review were divided as much as possible into four main sections: Test Bias, Experimenter or Examiner Effect, Expectancy or Halo Effect and Self-Fulfilling Prophecy, and Expectancy or Halo Effect and Spoken Language.

Test Bias

The literature on test bias reported on the various court cases and controversies that developed over the issue of whether or not the content of intelligence tests is culturally biased. The cases arose from the inappropriate or questionable placement of minority children in special education classes. Judges in the cases handed down conflicting rulings on whether the tests penalized children
from different linguistic and cultural backgrounds. Use of intelligence tests has been banned in public schools in several states. It has also been questioned and, in some instances, abandoned in France as well. The same issue has also entered the areas of reading and speech where clinicians have been alerted to be aware of the potential bias in materials and tests used with children who use nonstandard dialects of English, and to adjust materials and assessments accordingly.

**Experimenter or Examiner Effect**

The role of the examiner is a factor which is ignored or merely alluded to in the literature on test bias. The research on the experimenter/examiner effect investigated the influence of various biosocial and psychosocial attributes of the examiner: sex, socio-economic status (SES), warmth of personality, verbal and nonverbal manner, and experience. Studies were done in a variety of fields such as experimental research with animals, projective testing, intelligence testing, and survey research. Some reviews of the research found that the experimenter/examiner was a significant variable affecting every aspect of a situation from the choice of research design to the interaction with subjects and the interpretation of data.

In the area of intelligence testing, only warmth of examiner and examiner's positive verbal statements had a significant effect, producing positive test results. Other factors were generally found to be non-significant: experience level of the examiner was a negligible factor in obtaining reliable results; studies on race of
examiner were considered poorly designed; the idea that Black children perform better with White examiners was labeled a "myth"; sex of examiner had an effect in some studies where subjects obtained higher IQ's with female examiners, but had no significant effect in other studies. Results of the research on the experimenter/examiner effect in intelligence testing were inconsistent and at times contradictory, making generalizable statements difficult and tenuous.

Expectancy or "Halo Effect" and "Self-Fulfilling Prophecy"

Whether in schools, industry, intelligence testing, or other fields, studies generally reported that an evaluator's expectations, produced by either positive or negative information, affected subjects or biased judgmental decisions in such a way that end results fulfilled the expectations. Studies on teacher expectations and industrial experiments produced classic examples of these effects (Pygmalion in the Classroom and the "Hawthorne Effect," respectively).

In the area of intelligence testing, results were inconclusive. In some studies, IQ scores were found to be affected significantly by examiner expectations induced by pre-test biasing information on an examinee's achievement level, SES, appearance, and/or intelligence. Other studies found no significant difference between mean scores of high and low expectancy groups. Whether the effect occurred in the administration or scoring of the test was of concern in those studies that utilized written or taped responses as a control for administration. Others did not differentiate between the two. None investigated the effect of a speech handicap as a possible biasing factor on an examiner.
Expectancy or "Halo Effect" and Spoken Language

Studies in this area investigated the effect of either of two aspects of language on a listener's attitude. They reported on the effect of various speech characteristics such as pronunciation or articulation, or the comparative effect of two different languages such as English and French.

Phonological and dialectal variations were found to be socially stratified, producing social stereotyping. "Standard" and, therefore, "correct" forms were socially, not linguistically, based on that form used by the educated middle class. Those who use "non-standard" speech were viewed variously as having a "deficit," implying a cognitive as well as a social handicap; discriminated against in schools and the job market; associated with low SES, often "working-class" status; and afforded less "credibility." Standard speech was considered more prestigious and equated with linguistic effectiveness, and higher SES, even by those using the nonstandard speech, a type of "linguistic self-hatred."

Studies on reactions to speakers with ethnic or accented spoken languages of various national, cultural, or ethnic groups found that hearing the language elicits stereotypical judgments in listeners based on the characteristic stereotypes of the group using the language. Such speakers were rated lower on various scales including SES, desirability, and personality (including intelligence) even by members of the minority group which used the accented language. In one study, evaluations were found to occur within ten to fifteen seconds of hearing the speech or language characteristics.
Similar to studies on visual cues, speech and language characteristics were generally found to bias or render irrelevant the actual verbal content of a message, producing a "halo effect" in the listener and fostering ratings in the direction of stereotypical expectations.

There appears to be no investigation into the effect of a speech handicap in the judgmental process. Articulation errors of omission, distortion, and/or substitution of sounds produce non-standard speech. Like the nonstandard speech investigated in the literature review, it may cue low SES or even minority group status since it is not the speech of the majority of members of a community; may produce judgments of lower ability for not having learned "proper" speech, with a loss of credibility as well; and may elicit a negative expectancy or "halo effect" in the listener.

Methodology

This was an experimental study utilizing a 2 x 2 factorial research design. One factor was level of experience of rater, either practicing psychologists or graduate students of psychology, all of whom had the required training in the administration and interpretation of the Wechsler Intelligence Scale for Children-Revised (WISC-R). The second factor was type of examinee, speech-handicapped or non-speech-handicapped.

Since there were no studies evident in the examination of related literature which utilized a speech handicap as a variable in the interactive environment of an intelligence-test situation, this research explored whether an examinee's specific speech handicap has
a biasing effect on the scoring by trained examiners of audiotaped responses to an intelligence test. It also explored secondarily whether there is a difference between scoring by practicing psychologists as opposed to graduate psychology students.

Selection of Subjects

The sample for this study comprised two groups of examiners, thirty-two practicing psychologists and thirty-two graduate students of psychology, for a total sample of sixty-four examiners. Two examinees, one a child with a speech handicap of multi-articulation errors and the other a child with no speech handicap, recorded responses of identical content. Both were female, aged eight years eleven months. All persons were drawn from the southwestern region of Michigan and the north central region of Indiana. All examiners met the requirements for participation in the study.

Sixteen of the practicing psychologists and sixteen of the graduate students scored the audiotaped responses by the child with the speech handicap. The other sixteen of each group of examiners scored the audiotaped responses by the child with no speech handicap.

Instrumentation

The Comprehension subtest of the WISC-R was utilized to obtain the test responses. The WISC-R was well-standardized on a large sample of various classes of children, aged six and one-half to sixteen and one-half years, representative of the population of the United States. The Comprehension subtest only was utilized as it is one of the subtests considered more vulnerable to bias and because bias may be most evident in the first test administered in any order of subtests. The seventeen questions of the Comprehension subtest
can produce a maximum of thirty-four total raw-score points, which convert to a maximum of nineteen total scaled-score points having a mean of ten and a standard deviation of three. The subtest has a reliability coefficient of .73 and a standard error of measurement of ± 1.48 for the scaled scores for the age group of eight and one-half, one of eleven age groups for which statistical properties are given and the one nearest to the age of the examinee in this study, i.e., age eight years and eleven months (Wechsler, 1974, pp. 27-29).

Procedures

The test was administered in person to an elementary school-aged child, an eight-year eleven-month-old girl, with no speech handicap and recorded on a cassette audiotape. A second child of the same age and sex but with a speech handicap of multi-articulation errors was recorded giving the identical responses on a second cassette. The cassettes, each approximately five minutes in length, were professionally copied in sufficient number to give each examiner his/her own cassette for scoring purposes.

Subjects (examiners) were obtained through personal or telephone contact. Instructions for participation were given and arrangements made for data collection procedures that preserved the anonymity of each participant and avoided any mention of the speech handicap factor in the experiment.

Examiners were randomly assigned to one of the two tape conditions. The appropriate tapes and materials needed for scoring were delivered and collected when completed by persons other than the experimenter to preserve the assured anonymity of the respondents. Within one month all materials were returned except by two students,
for whom materials were returned within the next month. The raw scores awarded by the raters were personally converted to scaled scores.

Statistical Analysis

To analyze the data collected, the research utilized an experimental 2 x 2 factorial design with analysis of variance as the statistical tool, and the F-ratio test for significance of effect, with alpha set at .05. Simple effects were also analyzed, as was range of scores.

Summary and Discussion of Findings

This section presents a summary and discussion of the results of the study with regard to range of scores, means of scores, and three null hypotheses.

Range of Scores

The overall raw scores awarded by the total group of examiners for all the taped responses ranged from 13 to 22, which convert to a scaled-score range of 9 to 15. Given a mean of ten and a standard deviation of three, they range from one-third of a standard deviation below the mean to one and two-thirds of a standard deviation above the mean, a difference of two standard deviations. Although IQ is never inferred from a single subtest, scaled scores of 9 to 15 may be compared to IQ's of 95 to 125, intelligence categories ranging anywhere from "Average" to "Superior." The difference between being classified Average or Superior can be very significant, especially in a school setting. If this range were to be extrapolated to IQ's of 100 and 70, an Average child would be classified as Borderline. If extrapolated to IQ's of 130 and 100, a Very Superior child would
be classified as Average. A difference of 30 IQ points is cause for concern.

This finding is similar to the research results reported in the literature where different examiners obtained significantly different IQ's on the same children (Cattell, 1937; Schwarz & Flanigan, 1971) and where low interscorer agreement was observed in the scoring of test responses (Schwartz, 1966; Sattler et al., 1969).

In addition, the range of scores produced by the psychologists overall (SS 9-15) as well as for each type of response (SS 9-14 for the speech-handicapped responses; SS 11-15 for the non-speech-handicapped responses) was greater than that produced by the graduate students overall (SS 11-15) and for each type of response (SS 11-15 for the speech-handicapped responses; SS 12-14 for the non-speech-handicapped responses).

This finding differs from that of Auffrey and Robertson (1972) where the scoring range of the "novice" group was greater than that of the "expert" group. Although they found significant scoring variability within each experience level, they concluded that there was greater scoring reliability with more experience. The opposite result of the present research is similar to that of Plumb and Charles (1955) where the experience factor was negligible in the analysis of interscorer agreement, with the experienced psychologists failing to agree on a wider percentage of the responses than the graduate students. Sattler and Theye's 1967 review of the research also found "... little evidence that trained experimenters differ from less trained experimenters with respect to obtaining reliable results" (p. 355).
The wider range of scores awarded by the psychologists, then, suggests either greater scoring consistency on the part of the students or that the students may be more wary of using extreme scores. Both groups, however, awarded a wider range of scores to the speech-handicapped responses than to the non-speech-handicapped responses, which may represent the wide variety of effects a speech handicap may have on examiners, ranging from a negative bias to one of compensation.

The scaled scores derived from the raw scores awarded by the psychologists yielded double the number of Superior classifications (six) to the non-speech-handicapped responses than to the speech-handicapped responses (three), with the reverse being true of the students (six vs. eight). The psychologists' raw scores also yielded more Average classifications (three) to the speech-handicapped responses than to the non-speech-handicapped responses (one). Whereas the students' raw scores yielded only one Average classification to the speech-handicapped responses and none to the non-speech-handicapped responses. Since the content of both types of responses was the same, this again suggests the differences in value judgments elicited by the type of response.

Means of Scores

Overall the graduate students awarded higher mean scores than did the psychologists to each type of response. As suggested by other researchers (Meitus, 1973; Rosenthal, 1968), students may attempt to second-guess the purpose of the experiment and produce results opposite to those they believe are expected. It suggests as well a type of "Hawthorne effect" where participation in an experiment produces positive results.
The students awarded a higher mean score than did the psychologists to the speech-handicapped responses. This mean was higher than that of any other set of scores, with, as already noted, a higher number of scores placed in the Superior classification (eight vs. six, six and three) than by any other scoring group. This finding is suggestive of the overcompensatory "halo effect" noted by Burgemeister (1962) in the testing of subjects with cerebral palsy: "Motivated by a feeling of sympathy . . . the examiner easily believes his hope, i.e., that the child knows more than he can express, and hence overestimates the child's ability" (p. 117).

The difference between the scoring by the psychologists of the two types of responses was greater than that of the students. Overall, the difference between the psychologists as a group and the students as a group was greater than the difference between the speech-handicapped responses as a group and the non-speech-handicapped responses as a group. Which of the observed differences in the means was statistically significant was determined by the testing of three null hypotheses.

Testing the Hypotheses

The data were analyzed using a two-way analysis of variance with alpha set at .05. Each null hypothesis was rejected or retained according to the F-ratio test of significance. Before the main effects could be analyzed, it was necessary to test hypothesis 3 first to determine if there was any significant interaction between type of response and type of examiner.

Hypothesis 3: There is no significant interaction between type of response and type of examiner.
The ANOVA table indicated that there was no significant interaction present. Therefore this hypothesis was retained, allowing the testing of the two main effects hypotheses, the overall comparisons of means.

Hypothesis 1: There is no significant difference between the scores awarded to speech-handicapped and non-speech-handicapped responses.

Analysis of the data revealed no significant main effect for type of response either in raw or scaled scores. The hypothesis was retained as there was no significant difference between overall scores awarded to speech-handicapped and non-speech-handicapped responses.

Since there were no other apparent studies in the literature review with a speech handicap as a variable in a testing situation, this finding cannot be compared with others with exact equivalency. If one considers a speech handicap in the same general category as a nonstandard speech pattern, this overall finding does not agree with those studies where nonstandard speech, as evidenced by the varied pronunciation of certain sounds, or nonstandard oral language patterns, such as Black dialect or accented language, were considered less desirable and/or elicited negative stereotypical judgments (Anisfeld et al., 1962; Brookover et al., 1982; Harms, 1961; Granger et al., 1977; Labov, 1964, 1966; Lambert et al., 1960; Naremore, 1971; Williams, 1970). These studies generally involved rating the oral speech and language patterns directly on scales evaluating SES, effectiveness, desirability, credibility, and personality (which included intelligence). The present study utilized a more elliptical approach to obtain the effect of a type of speech pattern.
In addition, the subjects in these studies were usually teachers, students, or lay citizens, not psychologists and graduate students of psychology as in the present study. In the area of intelligence testing, where the subjects were equivalent, pre-test biasing information was involved in the investigation of expectancy or "halo effect." Conclusions were mixed. If one considers a speech handicap as similar to this type of condition, albeit inferred not specifically stated, then the overall finding is similar to those where there was no significant difference between the mean scores of the high and low-expectancy groups (Auffrey & Robertson, 1972; Dangel, 1971; Gillingham, 1970; Goetz, 1977; Mason, 1976; Sneed, 1975). However, the finding is dissimilar to those reporting significantly lower mean scores for the low-expectancy groups, those labeled or inferred as "dull," "underachieving," and/or "disadvantaged" even though they were actually of the same levels as the high-expectancy groups, and the responses were the same or similar (Babad et al., 1975; Sattler et al., 1970; Schroeder & Kleinsasser, 1972).

Another element to be considered, however, in this discussion is the possible masking effect that may have occurred in the comparison of the overall means of the two types of responses. Although the analysis of variance indicated that there was no significant interaction between type of examiner and type of response, the students awarded the highest mean score to the speech-handicapped responses and the psychologists awarded the lowest mean score to these same responses. This apparent interaction between type of examiner and type of response may have masked to some extent the comparison of overall means.
Hypothesis 2: There is no significant difference between the scores awarded by practicing psychologists and graduate students of psychology.

Analysis of the data revealed a significant main effect for type of examiner in the analysis of the raw scores, and it approached significance in the analysis of the scaled scores. Hypothesis 2 was rejected with respect to raw scores, the actual scores awarded by the examiners. The difference between the means of the overall raw scores awarded by each type of examiner was statistically significant, with the students awarding higher mean raw scores than the psychologists.

The hypothesis could not be rejected with respect to scaled scores although the F-ratio approached significance. Since scaled scores are scores to which raw scores are converted to attain equivalent scores among the subtests, the standardization of scores requires compressing a wide range of raw scores (1 to 34 in this subtest) into a narrower range of scaled scores (1 to 19). A certain amount of information may be lost in the smoothing process, which suggests greater significance in the raw scores since they are the actual scores awarded by the examiners, and there is no need in this study for any comparison with other subtests.

Although it was not necessary to study simple effects since main effects could be studied given the fact that there was no significant interaction, an analysis of simple effects was done to determine if there were a significant difference between each pair of means. Using a probability level of .01 to reduce the likelihood of a Type I error (concluding a significant difference exists where it does not), the assumption of homogeneity of variance was tested and
upheld for the raw scores but not the scaled scores. Therefore, for the simple effects tests of the raw scores, the pooled mean square error from the analysis of variance could be used; for the scaled scores, the estimate of variance of only the relevant two cells at a time was used.

Utilizing two-tailed t-tests with a probability level of .05, only the difference between the raw scores awarded by the psychologists and the students to the speech-handicapped-responses and the difference between the scaled scores to which they were converted were found to be statistically significant.

Hypothesis 2 was thus rejected for both raw and scaled scores as there was a statistically significant difference between the scores awarded by practicing psychologists and graduate students of psychology to the speech-handicapped responses. The psychologists gave the lowest mean score to these responses, the students the highest.

The fact that there was no statistically significant difference between the scores awarded by the psychologists and the students to the non-speech-handicapped responses suggests that level of experience and/or reliability of scoring is not a factor. This is similar to the finding in Sattler and Theye's 1967 review of studies in the area of intelligence testing that experience as a measure for obtaining reliable results was non-significant. It is similar as well to other studies where experience of examiner as a factor was found to be negligible (Plumb & Charles, 1955; Sattler et al., 1969; Schwartz, 1966).

The statistically significant difference found between the scores awarded to the speech-handicapped responses suggests the
possibility that the speech handicap may have had a different effect overall on each type of examiner, a negative "halo effect" on the group of psychologists scoring these responses, and a compensatory effect on the students scoring these same responses. The negative "halo effect" is similar to those studies where nonstandard speech and language patterns elicited negative judgments (Anisfeld et al., 1962; Brookover et al., 1982; Harms, 1961; Granger et al., 1977; Labov, 1964, 1966; Lambert et al., 1960; Naremore, 1971; Williams, 1970), as well as those studies where pre-test or inferred biasing information produced a negative expectancy or "halo effect" in the scoring of similar test responses (Babad et al., 1972; Sattler et al., 1970; Schroeder & Kleinsasser, 1972).

The compensatory effect on the students may be a type of over-compensatory "halo effect" noted by Burgemeister (1962) in the overestimation of the test responses of children with cerebral palsy out of sympathy and hope the child knows more than he can express, or it may be an attempt to second-guess the purpose of the experiment and produce opposite results (Meitus, 1973; Rosenthal, 1968).

Whether one focuses on the lower mean raw and scaled score awarded by the psychologists to the speech-handicapped responses or the higher mean score awarded by the graduate students to these same responses, the statistically significant difference between the two indicates that a speech handicap is a variable to be considered in the evaluative interaction of an intelligence-test situation.

Conclusions

The results of the present study lead to conclusions which are generalizable considering that the content and procedures for
administration and scoring of the WISC-R are highly standardized and that the training of psychologists throughout the United States in its use can therefore be assumed to be similar. However, given the fact that this is apparently the only study which has investigated the effect of speech-handicapped responses on the scoring of intelligence-test responses, replications of the study or studies with similar variables are needed to provide further data on the degree of generalizability of these conclusions.

On the basis of the results of the analysis of the data, the following conclusions may be drawn:

1. The wide range of scores, a range of two standard deviations, awarded by the examiners overall corroborates the body of research which has found widely different IQ's obtained by different examiners as well as low interscorer agreement within each response.

2. The wider range of scores awarded by the practicing psychologists as a group as contrasted with those of the graduate students as a group confirms the research finding that the examiners' experience level is not a significant factor in obtaining more reliable or consistent results in the scoring of test responses.

3. The wider range of scores awarded by each group of examiners to the speech-handicapped responses exemplifies the wide variety of effects a speech handicap has on examiners, ranging from one that is negative to one of compensation, i.e., a negative to a positive "halo effect."

4. Although the difference between the speech-handicapped responses as a group and the non-speech-handicapped responses as a group was not significant, the difference between the psychologists...
as a group and the graduate students as a group was. However, the significant difference was in their scoring of the speech-handicapped responses and not in the non-speech-handicapped responses. Given these facts, the conclusion is drawn that an examinee's speech handicap does influence examiners in their scoring of speech-handicapped responses to an intelligence test.

5. The psychologists awarding the lowest mean score to the speech-handicapped responses and the graduate students the highest, with the difference being a significant one, leads to the conclusion that the effect produced by the speech handicap may be in either of two directions, a negative or a positive/compensatory "halo effect."

Implications

As already reported, a number of studies and reviews of the research literature concluded that the rater/examiner was an important variable in any dyad, including intelligence testing, and that responses with non-standard speech characteristics and/or language patterns, as well as pretest biasing information may trigger an expectancy/"halo effect" in the rater/examiner and produce a self-fulfilling prophecy in the direction of the "halo."

Nonstandard speech patterns were viewed variously in the literature as inferior, indicative of lower intelligence, a cognitive deficit for not having learned standard speech, less credible, not prestigious, and/or of lower SES. Expectations were produced by societal prejudices, social experiences, beliefs about IQ and cultural deficits, and stereotypical reactions to the group using those patterns (usually of minority group status) which have generalized
to the speech patterns. Teachers, graduate students, and lay citizens were generally the subjects.

In recent court cases in California, Illinois, Maryland, Michigan, and New York, psychologists were implicated in the misplacement of minority children in special education classes or sued for misclassification of students as mentally impaired who were only speech or learning impaired. This indicates that psychologists are not immune to the factors which cue expectations. This is reinforced by the conclusions reported in those studies which found various biobehavioral and psychosocial attributes (such as sex or warmth of examiner) as well as pre-test biasing information on various aspects of the examinee to have an effect on the administration and/or scoring of intelligence tests.

If a nonstandard speech pattern is viewed as inferior, indicative of lower intelligence, and/or lower SES, then hearing such a pattern would be equivalent to pre-test biasing information labelling an examinee as "dull," "underachieving," or "disadvantaged," producing a low-expectancy condition as soon as the nonstandard speech pattern is heard. One study found judgments were made within ten to fifteen seconds after hearing a speech pattern; another considered the perception of speech a psychological problem with stereotyping common.

If the perception of speech is psychological, therefore covert, and evaluation of speech almost instantaneous, implying a stereotypical reaction, then determining which speech characteristic produces which type of rating is difficult. Compounding this difficulty is the fact that factors of "voice," "enthusiasm of
expression," "low SES," and "minority group status" implicated in the research studies of necessity also include speech characteristics, such as pronunciation and articulation. The reverse, then, must be true of the speech handicap of multi-articulation errors utilized in the present study to induce an expectancy/"halo effect." Of necessity, speech characteristics are conveyed through a voice, with expression, with SES, and possibly even minority group status perceived by the listeners (examiners). In order to reach any conclusions, however, it must be assumed that the speech handicap consisting of multi-articulation errors of omission, substitution, and distortion of sounds was the overriding variable in this study. This is affirmed by the fact that seven of the examiners scoring the speech-handicapped responses, five psychologists and two graduate students, mentioned the articulation factor on their protocols.

In addition, there were no visual cues in this study, no verbal or nonverbal interactions between the examinee and those scoring the responses, the content of the responses was the same in both conditions, and there was no significant difference in the scoring of the non-speech-handicapped responses by the two groups of examiners. Given the fact that these variables were controlled, the speech handicap of multi-articulation errors becomes the one major variable which produced the significant difference found in this study between the scoring by the psychologists and that of the graduate students of psychology of the speech-handicapped responses. Since examiners may be more wary in experimental situations, it may be that this factor "... would play an even larger role in standard test situations" (Sattler et al., 1970, p. 175). The finding supports the
hypothesis that a speech handicap triggers an expectancy or "halo effect" in the examiner producing a "self-fulfilling prophecy" in the scoring of intelligence-test responses.

The "halo effect" was found to be two-directional, either positive or negative. This finding, indicative of the wide range of scores covering two standard deviations produced by the examiners in this study, implicates the examiner as an additional important variable to be considered in the interpretation of any IQ score, especially that of a child who is "different." The examiner needs to question his/her role in the interaction and the validity of his/her interpretations and scoring of test responses; and he/she needs to be aware that the scoring of an intelligence test may not be as objective as he/she would like to believe. Psychologists are subject to the same influences as any other member of society. It is not purposive bias that is implied but the more subtle expectancy or "halo effect" that influences judgmental decisions.

Practicing psychologists, for example, tend to see mainly children with learning or other problems. Some of these children may also have a speech handicap. This view of speech handicap combined with learning problem may generalize to any child with a speech handicap, reinforcing the view held by some that nonstandard speech implies a cognitive deficit for not having learned standard speech. This would account for the lower mean score awarded by the psychologists to the speech-handicapped responses, i.e., scores in the direction of the negative "halo effect." In addition, five psychologists compared to only one student scoring the speech-handicapped responses noted that they had trouble with the articulation and/or would have probed further.
those and/or other responses, "Scoring, probing, and inquiring may be affected by the experimenter's impression of whether the subject may be able to answer the questions" (Sattler & Theye, 1967, p. 351). In the scoring of the non-speech-handicapped responses, identical in content to the speech-handicapped responses, only one psychologist and no students indicated a need for further probing of the responses.

The higher score awarded by the graduate students to the speech-handicapped responses, i.e., in the direction of the positive/compensatory "halo effect," may be attributed to a number of possibilities. As already noted, the students may have suspected the purpose of the experiment and "bent over backward" to counter any expectations; just being part of the experiment may have produced a more positive attitude and results, a type of "Hawthorne effect"; they may have sympathized with a child with a speech handicap, assumed the child knew more than was expressed and overestimated the child's ability; or, being students, they may have followed Wechsler's Manual more closely to check responses and so gave more accurate scores. (Scoring by this researcher of the test responses places them in the 'Superior' category, raw score 20, scaled score 14, which is how the child who gave the original responses actually achieves in school.) Practicing psychologists, having more experience in testing, may not refer to the Manual as frequently as students who have not yet done as much testing.

If one accepts the assumption that the graduate students' higher score for the speech-handicapped responses is the more accurate one, then it would presumably follow that it is the psychologists' lower score for these same responses, significantly different
statistically from that of the students, that can be implicated as
the effect that is produced by speech-handicapped responses in the
scoring of intelligence-test responses. One may then propose that
a speech handicap of multi-articulation errors affects the cue valid-
ity of the verbal responses, triggers a negative "halo effect," which
negates the content of the responses and reduces accuracy of judgments
which are made in the direction of the "halo." An expectation becomes
a "self-fulfilling prophecy."

Recommendations

A. For further research:

Since this was the only apparent study investigating the
effect of a speech handicap on the scoring of intelligence-test
responses, there is a need for further research utilizing these var-
iables. Because of the difficulties inherent in inducing a "halo
effect," replications of this study could determine the degree of
generalizability of the findings. The recommendations for further
research are:

1. Replicate the study using a sample of examiners or examin-
ners from other geographical areas.

2. Replicate the study investigating the effect of other
types of speech handicaps, such as stuttering, or some obvious differ-
ences in voice quality (hoarseness, nasality), to determine which non-
standard speech characteristics cue which effects.

3. Obtain a large enough sample of examiners in order to add
sex of examiner as an additional independent variable to the analysis
of variance, as well as allow analysis of the scoring of each
response. These factors, sex of examiner and response ambiguity, were utilized as variables in different studies. They would add additional insights, as "... it is conceivable that some situational and/or experimenter or subject effects only occur on test items or on test responses which are more ambiguous or less highly structured" (Sattler & Theye, 1967, p. 257).

B. For the practice of psychology:

If one assumes that the lower mean score produced by the psychologists was the less accurate one, and that the negative effect may be attributable to the expectations produced by their experience in the field, then there may be implications for the professional preparation and practice of psychologists in intelligence-test scoring. The recommendations proposed are:

1. Have practicing psychologists periodically attend refresher courses or workshops on the scoring of intelligence tests to reexamine their judgments and interpretations and update their knowledge of the research in the area of examiner bias and expectancy effects.

2. Include in the workshops, or in the coursework for training psychologists, the topic of a speech handicap to develop understanding of the various types, what they imply, and their possible effects on listeners' judgments.

3. Have any child with a speech handicap who is being considered for any type of special placement tested by a second examiner, either within the same time frame with a similar test, or with the same test after the recommended time lapse. Given the wide range of
scores and the significant difference in the scoring of the speech-handicapped responses by the two types of examiners and by different examiners of the same type in the present study, it is important to be "... cognizant of the impact of different examiners on the child as well as their evaluation of the subject's responses" (Schwarz & Flanigan, 1971, p. 263).

4. Have anyone who utilizes intelligence scores to evaluate someone--teachers, administrators, parents, employers, psychologists themselves--made aware not only of test reliability but of the examiner as a variable. "Judgment of an individual's score on special abilities and IQ tests, etc., must not only be viewed in light of which test was used, but must also take into consideration the previously ignored variable of the specific administrator. ... All persons using test scores must recognize the strong influence of E and make decisions accordingly" (Kintz et al., 1965, p. 230).
APPENDIX A

Letter to Participants
Dear Colleague:

Your willingness to participate in my research study as we discussed on the telephone is very much appreciated. Enclosed are all the materials you will need except for a cassette player to listen to and score the five-minute recording.

The scoring needs to be completed and returned in the enclosed self-addressed envelope by April 29, 1983.

If you should have any questions on the procedures, please do not hesitate to call me at (219) 272-7404. Results of the study will be made available to any participant who requests them.

Thank you again for your time and cooperation.

Sincerely,

Julia Peyser
Doctoral Candidate
Andrews University

Enclosures - 4
List of procedures
Protocol for scoring
Scoring criteria
Cassette tape recording of the WISC-R Comprehension subtest
APPENDIX B

List of Procedures
PROCEDURES

IMPORTANT: These procedures need to be followed exactly to simulate and make valid comparisons with a "live" testing situation. The subtest is being administered to a girl eight years eleven months of age.

1. Simulate a "live" situation by listening to your tape alone in a private area.

2. Use the scoring criteria enclosed, if needed, to check with sample responses, but only while scoring.

3. Listen to only one question and its response at a time.

4. Write the response on the protocol enclosed.

5. If needed, listen to the response only once more.

6. Score the response on the protocol enclosed.

7. Do not listen to any part of your tape again after finishing the last question or redo any of the scoring.

8. Please do not discuss your tape or your scoring with anyone else at this time to avoid any bias.

9. Return all materials in the enclosed self-addressed envelope to the same place where distributed by April 29, 1983, on which date they will be collected by my assistant.

10. Please check the boxes after the following statements and return with the other materials:

   The procedures for scoring were followed as directed . . . .  

   I have returned all materials requested:

   Cassette   Protocol   Scoring criteria   Procedures
APPENDIX C

Letter of Appreciation
May 25, 1983

Dear Colleague:

Thank you for participating in my dissertation research study. The time you took out of what I know is a busy schedule to listen to and score the tape is greatly appreciated.

I was impressed with the fact that virtually everyone contacted was so cooperative. Thank you again and best wishes to you in your own research.

Cordially,

Julia Peyser
APPENDIX D

Responses to Comprehension Subtest Questions
Non-speech-handicapped Responses with No Errors

1. Wash it off and put a band-aid on it.
2. Take it up to the cas'... cash register and tell her that we found this on the floor.
3. a. Call the cops and... and call the fire department.
   b. Yell fire.
4. a. If somebody robs a place or something.
   b. To keep the country safe.
5. Say you're sorry.
6. Ask her to stop fighting.
7. It's warmer and stronger.
8. a. The cop can... write the license plate down.
   b. You know who it's... who it belongs to.
9. a. They can teach them a lesson.
   b. They won't do anything wrong.
10. To send them.
11. Nobody'll get sick from it.
12. a. It might be a robber.
   b. It gives to the poor.
13. I don't know.
14. I don't know.
15. Because you might hurt somebody's feelings.
16. 'Cause it's soft and warm.
17. I don't know.
Errors in Speech-Handicapped Responses

1. Wash it off and put a band-aid on it.
2. Take it up to the cash and tell her (that) we found this on the floor.
3. a. Call the cop(s) and call the fire department.
    b. Yell, yell fire.
4. a. If somebody rob(s) a place or something.
    b. (To) keep the country safe.
5. tex you're start.
6. et her to tap fighting.
7. It(s) warmer and stony.
8. a. The cop can write their plate...lars plate down.
    b. You know who it barrns to.
9. a. They can teach am a lost.
    b. They won't do anything run.
10. To stend them.
11. Nobody('ll) won't get sick from it.
12. a. (It) might be a robber.
    b. It give(s) to the poor.
13. I don't know.
14. I don't know.
15. (Because) you might hurt somebody(s) feeling(s).
16. "Cause it(s) soft an(d) warm.
17. I don't know.

Substitutions and insertions of sounds are in the International Phonetic Alphabet; omissions of sounds or words are in parentheses.
APPENDIX E

Raw and Scaled Scores of Each Respondent
| Question #1 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | Total Raw Score | Total R.S. Score |
|------------|---|---|---|---|---|---|---|---|---|----|----|----|----|----|----|----|----------------|-----------------|
| Psychologists' Speech-Handed responses | 101 | 2 | 2 | 2 | 1 | 2 | 0 | 1 | 2 | 1 | 2 | 0 | 2 | 1 | 0 | 0 | 1 | 0 | 17 | 13 |
| Non-speech Handled responses | 102 | 2 | 2 | 2 | 2 | 0 | 2 | 1 | 2 | 1 | 2 | 1 | 0 | 0 | 0 | 1 | 0 | 17 | 13 |
| | | 4 | 2 | 2 | 1 | 2 | 2 | 0 | 1 | 2 | 2 | 2 | 2 | 1 | 1 | 1 | 0 | 1 | 0 | 17 | 13 |
| | | 6 | 2 | 2 | 2 | 1 | 0 | 1 | 2 | 3 | 1 | 2 | 2 | 2 | 1 | 2 | 0 | 0 | 0 | 17 | 13 |
| | | 8 | 2 | 2 | 2 | 1 | 0 | 1 | 2 | 3 | 1 | 2 | 2 | 2 | 1 | 2 | 0 | 0 | 0 | 17 | 13 |
| | | 10 | 2 | 2 | 2 | 1 | 0 | 1 | 2 | 3 | 1 | 2 | 2 | 2 | 1 | 2 | 0 | 0 | 0 | 17 | 13 |
| | | 12 | 2 | 2 | 2 | 1 | 0 | 1 | 2 | 3 | 1 | 2 | 2 | 2 | 1 | 2 | 0 | 0 | 0 | 17 | 13 |
| | | 14 | 2 | 2 | 2 | 1 | 0 | 1 | 2 | 3 | 1 | 2 | 2 | 2 | 1 | 2 | 0 | 0 | 0 | 17 | 13 |
| | | 16 | 2 | 2 | 2 | 1 | 0 | 1 | 2 | 3 | 1 | 2 | 2 | 2 | 1 | 2 | 0 | 0 | 0 | 17 | 13 |
| | | 18 | 2 | 2 | 2 | 1 | 0 | 1 | 2 | 3 | 1 | 2 | 2 | 2 | 1 | 2 | 0 | 0 | 0 | 17 | 13 |
| | | 20 | 2 | 2 | 2 | 1 | 0 | 1 | 2 | 3 | 1 | 2 | 2 | 2 | 1 | 2 | 0 | 0 | 0 | 17 | 13 |
| | | 22 | 2 | 2 | 2 | 1 | 0 | 1 | 2 | 3 | 1 | 2 | 2 | 2 | 1 | 2 | 0 | 0 | 0 | 17 | 13 |
| | | 24 | 2 | 2 | 2 | 1 | 0 | 1 | 2 | 3 | 1 | 2 | 2 | 2 | 1 | 2 | 0 | 0 | 0 | 17 | 13 |
| | | 26 | 2 | 2 | 2 | 1 | 0 | 1 | 2 | 3 | 1 | 2 | 2 | 2 | 1 | 2 | 0 | 0 | 0 | 17 | 13 |
| | | 28 | 2 | 2 | 2 | 1 | 0 | 1 | 2 | 3 | 1 | 2 | 2 | 2 | 1 | 2 | 0 | 0 | 0 | 17 | 13 |
| | | 30 | 2 | 2 | 2 | 1 | 0 | 1 | 2 | 3 | 1 | 2 | 2 | 2 | 1 | 2 | 0 | 0 | 0 | 17 | 13 |
| | | 32 | 2 | 2 | 2 | 1 | 0 | 1 | 2 | 3 | 1 | 2 | 2 | 2 | 1 | 2 | 0 | 0 | 0 | 17 | 13 |

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
<table>
<thead>
<tr>
<th>Question #1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
<th>16</th>
<th>17</th>
</tr>
</thead>
<tbody>
<tr>
<td>Students' speech-handicapped responses</td>
<td>201</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>19</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>19</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>17</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>18</td>
</tr>
<tr>
<td></td>
<td>9</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>15</td>
</tr>
<tr>
<td></td>
<td>11</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>20</td>
</tr>
<tr>
<td></td>
<td>13</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>18</td>
</tr>
<tr>
<td></td>
<td>15</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>20</td>
</tr>
<tr>
<td></td>
<td>17</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>19</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>21</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>23</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>25</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>27</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>30</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

| Student's non-speech-handicapped responses | 296 | 18.5 | 24 | 13 | 375 |
| 202 | 2 | 2 | 1 | 0 | 2 | 1 | 2 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 16 |
| 204 | 2 | 2 | 2 | 2 | 0 | 1 | 2 | 2 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 17 |
| 206 | 2 | 2 | 2 | 2 | 0 | 1 | 2 | 2 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 19 |
| 208 | 2 | 2 | 2 | 2 | 0 | 1 | 2 | 2 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 18 |
| 210 | 2 | 2 | 2 | 2 | 0 | 1 | 2 | 2 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 18 |
| 212 | 2 | 2 | 2 | 2 | 0 | 1 | 2 | 2 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 18 |
| 214 | 2 | 2 | 2 | 2 | 0 | 1 | 2 | 2 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 18 |
| 216 | 2 | 2 | 2 | 2 | 0 | 1 | 2 | 2 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 18 |
| 218 | 2 | 2 | 2 | 2 | 0 | 1 | 2 | 2 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 18 |
| 220 | 2 | 2 | 2 | 2 | 0 | 1 | 2 | 2 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 18 |
| 222 | 2 | 2 | 2 | 2 | 0 | 1 | 2 | 2 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 18 |

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
SELECTED BIBLIOGRAPHY

American Psychological Association, Monitor, November 1979, p. 2; November 1980, pp. 6-8.


128


Michigan State Board of Education. Special Education Rules, 2 July 1980.


Sneed, Glenn A. "An Investigation of Examiner Bias, Teacher Referral Reports, and Socioeconomic Status with the WISC-R." (Ph.D., University of Toledo, 1975.) Dissertation Abstracts (1976): 4367-A.


Vita

JULIA B. PEYSER
1820 Ribourde Drive, South Bend, Indiana, 46628

EDUCATION:

<table>
<thead>
<tr>
<th>Year</th>
<th>Degree</th>
<th>Institution</th>
<th>Field</th>
</tr>
</thead>
<tbody>
<tr>
<td>1978 - present</td>
<td>Doctoral student - Educational Psychology</td>
<td>Andrews University, Berrien Springs, Michigan.</td>
<td></td>
</tr>
<tr>
<td>1968 - 1973</td>
<td>Post-Master's program - Administration and Supervision; Radio and instructional TV summer program</td>
<td>Hofstra University, Hempstead, New York.</td>
<td></td>
</tr>
<tr>
<td>1948</td>
<td>M.A. - Speech and Drama</td>
<td>Columbia University Teachers College, New York, New York.</td>
<td></td>
</tr>
<tr>
<td>1943</td>
<td>Valedictorian - Harrison High School, Harrison, New York.</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

PROFESSIONAL EXPERIENCE:

<table>
<thead>
<tr>
<th>Year</th>
<th>Position</th>
<th>Institution</th>
<th>Responsibilities</th>
</tr>
</thead>
<tbody>
<tr>
<td>1978 - present</td>
<td>Speech Correctionist</td>
<td>Niles Community Schools, Niles, Michigan.</td>
<td>Member of multi-disciplinary team testing, diagnosing, and developing educational plans for children referred as having language/learning problems. Teacher of children with speech, language, and/or learning problems.</td>
</tr>
<tr>
<td>1975 - 1978</td>
<td>Adjunct Lecturer</td>
<td>Indiana University at South Bend, South Bend, Indiana.</td>
<td>Taught Business and Professional Speaking, Public Speaking, Oral Interpretation of Literature. Chairman, Adjunct Faculty Representatives to Academic Senate. Member, Academic Personnel Committee.</td>
</tr>
<tr>
<td>1973 - 1975</td>
<td>Principal</td>
<td>Kennedy School, Penn-Harris-Madison School Corporation, Oceola, Indiana.</td>
<td>Had full responsibility for 530 students grades four to six, thirty-one full and part-time certified and non-certified staff, and twenty parent volunteers. Reorganized the school to provide alternative educational programs including multi-grade teams and self-contained classes. Implemented a management-by-objectives teacher evaluation system. Presented a workshop on &quot;Teacher Evaluation using Management by Objectives&quot; for conference of district administrators.</td>
</tr>
<tr>
<td>1972 - 1973</td>
<td>Assistant Principal</td>
<td>Northport Public Schools, Northport, New York.</td>
<td>Shared the responsibilities in two schools for 1385 students grades Kindergarten to six and 119 full and part-time personnel. Guided a revision of the mathematics program. Chaired the multi-disciplinary team meetings—a system for the diagnosis and remediation of children with learning disabilities.</td>
</tr>
</tbody>
</table>
1966 - 1972
Speech and Language Therapist - Northport Public Schools, Northport, New York. Taught speech and auditory perception skills to students with speech, language, and/or learning problems. Co-chaired the development of a multi-disciplinary team approach to working with children with learning problems that was implemented system-wide. Presented workshops on "Auditory Perception" and "How Can a Parent Stimulate Learning?".

1965 - 1966
Lecturer - Speech Arts and Sciences Department, Hofstra University, Hempstead, New York. Taught Fundamentals of Speech, Voice and Diction, and was speech evaluator of elementary-school teacher candidates.

1954 - 1965

CERTIFICATION:
Michigan: Speech Correction (K-12); English, Speech, and Spanish (9-12); all subjects (7-8).
Indiana: Elementary Administrator and Supervisor
New York: School Administrator and Supervisor; Supervisor of Secondary Education; Teacher of English in the Secondary School and of Speech, K to 12.

COMMUNITY SERVICE:

South Bend, In.
1983
Member, American Association of University Women.
1981
Presenter, Communications Seminar, South Bend Chamber of Commerce Leadership Program.
1977 - 1978
President, Midwest Chamber Orchestra Women's Division and member of the board.
1977
Presenter, Nuffield Math Program Workshop, NEA Teachers' Conference, South Bend.
1976 - 1977
Program Vice President, Indiana University at South Bend Faculty Women's Club.
1976
Judge, Rotary International High School Speech Contest.
1975, 1976
Guest panelist, "Women in Careers", Indiana University at South Bend education classes.
1974, 1975
Guest lecturer, "Women in Administration", Indiana University at South Bend interdisciplinary classes.

New York:
1962 - 1985
1959 - 1962
Director, After-School Recreation Program (Cultural Arts), Walnut Street School P. T. A., Uniondale, New York.
1957 - 1958
Leader, Cub Scout group, Uniondale, New York.