



University of
Massachusetts
Amherst

Identification of elements of bias in teacher ratings.

Item Type	Dissertation (Open Access)
Authors	Cromack, Theodore Robert
DOI	10.7275/g31m-qt03
Download date	2026-06-18 04:31:58
Link to Item	https://hdl.handle.net/20.500.14394/12640



IDENTIFICATION OF ELEMENTS OF BIAS
IN TEACHER RATINGS

A Dissertation Presented
by
THEODORE ROBERT CROMACK

Submitted to the Graduate School of the
University of Massachusetts
in partial fulfillment of the requirements
for the degree of

DOCTOR OF EDUCATION

October 1971

Major Subject Educational Research

(c) Theodore Robert Cromack 1971

All rights reserved


IDENTIFICATION OF ELEMENTS OF BIAS
IN TEACHER RATINGS

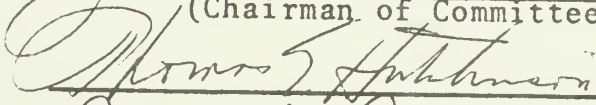
A Dissertation

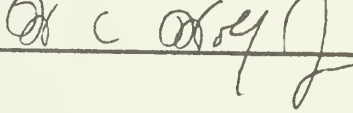
By

THEODORE ROBERT CROMACK

Approved as to style and content by:



(Chairman of Committee)




July 1971

(Month) (Year)

ACKNOWLEDGMENTS

For the assistance which I received in reaching this stage of my educational development, I am indebted to more people than can possibly be named. A few, whose contributions cannot be overlooked, are mentioned below as an indication of my debt of gratitude to them:

Bill Fanslow, Ray Budde, and Roger Peck for allowing me to use their class time.

Students in Teacher Education and in Administration for serving as subjects for this study.

Interns and student teachers for willingly giving their time in the interest of research.

Earl Seidman for his valuable time in the role of "Devil's Advocate" which stimulated me to look beyond empiricism.

Consultants at the University Computer Center for their patience.

My committee: Jim Fortune, Tom Hutchinson, and Bill Wolf, for their criticism, encouragement, and time.

Larry Wightman, Rick deFriesse, and especially Betty Proper for indispensable technical assistance and encouragement. Without such friends, failure would be guaranteed.

The three unnamed teachers for making the study possible.

Finally, my wife, Mary, and our children, Doug, Nina, Hazel, Fred, and Charles, for the untold hardships, frustrations, and disappointments; especially because they are untold.

TABLE OF CONTENTS

APPROVAL	ii
ACKNOWLEDGMENTS	iii
LIST OF TABLES	vii
CHAPTER I. INTRODUCTION	1
Background	1
The Problem	2
Purpose	2
Statement of the Problem	3
Constraints	3
Definitions	3
Variables	3
Independent Variables	4
Active Variables	4
Purpose of Rating	4
Abstraction Level	5
Assigned Variables	5
Willingness to Confront	5
Concept of Average	6
Concept of Dispersion	7
Dependent Variable	7
CHAPTER II. REVIEW OF THE LITERATURE	8
The Hypotheses	12
Reduction of Variance	14
Confirmation of Intertape Variance	14
Experimental Effects	14
CHAPTER III. PROCEDURES	18
Subjects	18
Method	18
Data Collection	19
Videotapes	19
Willingness to Confront	19
Concepts of Average and of Dispersion	19
Purpose of Rating	20
Abstraction Level	20
Data Analysis	20
Rater Variance	20
Experimental Effects	23
Instrumentation	25
Semantic Differential	25
Rating Scale	27

CHAPTER IV.	RESULTS.	29
	Reduction of Variance.	29
	Concept of Average	29
	Ordinary Teacher	30
	Average Teacher.	30
	Reliability.	32
	Data Adjustments	32
	Linear Adjustment.	32
	Proportional Adjustment.	32
	Z Transformed Scores	32
	Item Residual Adjustment	33
	Reliability.	34
	Intertape Variance	37
	Remaining Variables.	37
	Willingness to Confront.	38
	Levels	38
	Hypothesis	40
	Purpose of Rating.	41
	Abstraction Level.	41
	Interactions	43
	Willingness to Confront.	43
	Purpose of Rating.	45
	Teacher and Abstraction Level.	48
	Summary.	50
CHAPTER V.	SUMMARY AND CONCLUSIONS.	52
	Summary.	52
	Reduction of Variance.	52
	Interteacher Variance.	54
	Main Effects	54
	Interaction Effects.	55
	Conclusions.	55
	The Problem.	55
	Are there effects of know- ledge of output use of ratings on raters.	56
	Are there effects of varied levels of abstraction of rating items on rater output	56
	Can knowledge of rater con- cept of average be used to reduce interrater variance	57

Can knowledge of rater differences in concept of average and dispersion be used to reduce inter-rater variance.	58
Does rater willingness to confront interact with rating output use	59
Procedures for Studying the Rating	
Process	60
Realism	60
Sample Size	61
Variables	61
Instrumentation	62
Implications for Further Study.	62
Further Study Profitable.	62
Improvement of Present Study.	63
Concepts of Average and Dispersion.	63
Willingness to Confront	63
Purpose of Rating	63
Abstraction Level	64
Further Study of the Rating	
Process	64
Implications for Teacher Rating	65
REFERENCES.	67
APPENDIX A, SEMANTIC DIFFERENTIAL	69
APPENDIX B, INSTRUCTIONS TO RATERS.	74
APPENDIX C, RATING SCALE.	78

LIST OF TABLES

TABLE 1.	DESIGN FOR RELIABILITY ANALYSIS OF SCORES BASED ON RATINGS MADE BY 48 RATERS OF 3 TEACHERS ON 18 ITEMS	22
TABLE 2.	ANALYSIS OF VARIANCE OF EXPERIMENTAL EFFECTS.	24
TABLE 3.	FACTOR LOADINGS FROM SEMANTIC DIFFER- ENTIAL	26
TABLE 4.	DESCRIPTION OF ITEM RATINGS OF AVERAGE TEACHER.	31
TABLE 5.	CORRELATION COEFFICIENTS AMONG AVERAGE TEACHER MEASURES AND EACH TEACHER TOTAL SCORE.	31
TABLE 6.	RELIABILITY COEFFICIENTS.	35
TABLE 7.	MEANS OF INTERITEM VARIANCES.	36
TABLE 8.	MEANS OF RATINGS OF EACH TEACHER.	38
TABLE 9.	ANALYSIS OF VARIANCE FOR ADJUSTED AND UNADJUSTED SCORES.	39
TABLE 10.	DESCRIPTION OF RATER SEMANTIC DIFFERENTIAL SCORES ON WILLINGNESS TO CONFRONT.	40
TABLE 11.	MEANS OF PURPOSE OF RATING.	41
TABLE 12.	COMPARISON OF ABSTRACTION LEVEL MEANS	42
TABLE 13.	CELL MEANS FOR INTERACTION OF WILLINGNESS TO CONFRONT WITH OTHER VARIABLES	46
TABLE 14.	CELL MEANS FOR INTERACTION OF PURPOSE OF RATING WITH TEACHERS AND ABSTRACTION LEVEL.	48
TABLE 15.	MEANS OF TEACHERS AND ABSTRACTION LEVELS.	49

C H A P T E R I

INTRODUCTION

Background. What is the purpose of school? No one is satisfied that subject-matter knowledge alone is sufficient as a purpose. Other benefits are expected; for example, good citizenship and social development. How does one measure whether the school is meeting its purpose? With the increased emphasis on accountability in education, educators are being forced to face this question. Increased objectivity in decisions regarding teacher retention, promotion, and improvement is demanded. It has been recently reported that the American Federation of teachers is planning to rate teachers. State teachers' organizations are becoming actively involved in developing teacher evaluation procedures. The question to be asked is what is being evaluated?

If achievement is not the criterion of teacher effectiveness, what is? Barr (1950), Ryans (1950), McKeachie, Lin, and Mann (1971), and others have searched for relationships between achievement and other measures, among these other measures were teacher rating by pupils and by outside observers. Their goal was a criteria of "Teacher Effectiveness" which is so desperately needed as explained by Gage (1963). Rosenshine (1968, 1970), and Rosenshine and Furst (1971) have urged researchers to consider the teacher rating form as a viable measure of

teacher effectiveness. Yet ratings have only occasionally shown significant relationship with achievement or any accepted measure of teacher accomplishment.

Teacher rating has been used often but has been studied rarely. We frequently rate but we don't truly know much about rating. We use rating data but we are ignorant of the process employed in collecting that data. Empirical investigation of the rating process is desperately needed if teacher rating is to become a useful measure of teaching effectiveness.

Logic tells us that classroom behavior of the teacher is the key element in education that distinguishes school from other social institutions. Judgment of the teacher's classroom behavior has to be a viable measure of teaching effectiveness. Good (1959, p. 439) defines rating "as a process of estimating, according to some systematized procedure the degree to which an individual person or thing possess any given characteristic. It is suggested that if a method existed to calibrate and assure standardization of a baseline for raters, the rating process could be purified and rating could be a highly effective method for distinguishing quality of teaching.

THE PROBLEM

Purpose. The purpose of this study is to generate and test certain hypotheses concerned with rating; input, process, and output variables, involved in judgment of effectiveness of teaching.

Statement of the Problem. This study seeks to answer the following questions: Are there effects of knowledge of output use of ratings on raters? Are there effects of varied levels of abstraction of rating items on rater output? Can a knowledge of rater concept of average be used to reduce interrater variance? Can a knowledge of rater differences in dispersion (use of extreme rating categories) be used to reduce interrater variance? Does rater willingness to confront interact with rating output use?

Constraints. To clarify the problem statement further and to introduce the elements of the study design, certain limitations and constraints within this study are specified.

A basic assumption of the study is that the best rating has the least interrater variance within each item. In other words, best rating connotes highest interrater reliability.

The design employs supervisory and potentially supervisory personnel rating experienced teachers (on videotapes). The situation simulated is appropriate for supervisors of teachers; there is no intent within the study to represent the investigation as student ratings of teachers.

This study is concerned with variables involved in the rating process which may identify training needs for raters but raters in the study will receive no training.

DEFINITIONS

Variables. The variables under study in this paper are

listed below with their definitions. Further operationalization of these variables may be found under Chapter 3, Instrumentation.

Independent Variables. The generic term "Independent variables" subsumes both active and assigned (experimental and measured) variables (Kerlinger, 1966, pp. 38-43).

Active Variables.

Purpose of rating. Ratings may vary greatly as to the intended output use. Different intended use of ratings may pose different levels of threat to the rater. As a range of levels of threat, raters will be instructed that rating output use will be for one of three purposes: research only, administrative, or diagnostic. The research only level represents a minimum threat to the rater in that it describes the ratings to be for the purpose of analysis of the viability of the rating instrument, e.g., reliability and item analysis. Administrative purposes will be presented by representing anonymous ratings which may subsequently be used as a basis for retention, dismissal or merit raises. Such intended use is expected to pose a degree of threat to the rater to the extent that his output may affect the career and livelihood of the ratee. As a third level of threat to the rater, another situation will be presented whereby the rater will be instructed to complete the form with the intention of discussing the rating output with the ratee. In this situation it is anticipated that the

rater must be prepared to justify why the rater marked each item as he did. The variable labeled "purpose", therefore, is operationally defined by the instructions provided the raters representing three intended uses of the rating output.

Abstraction level. Johnson (1955) suggests that "when an abstract judgment is called for an affective judgment is commonly given". Conversely, he proceeds to say, "when the judgment called for is straightforward and easy, the judge is likely to maintain his set for that judgment and not be diverted into judgment of something else". The abstraction level of the rating scale items will determine whether the rater judges by a feeling or by a more objective assessment. Operationally this variable is defined by categorizing rating items into certain levels of abstraction. The categories (low, medium, and high) were developed by having a number of judges sort a large pool of potential items which had been drawn from existing instruments into five separate abstraction levels. All items were then scaled using Togerson's "Law of Categorical Judgments" (1958) and those six items for each of the scale values representing low, medium, and high levels of abstraction were used in the instrument.

Assigned Variables. Variables which will be measured, not manipulated, include those on which subjects will be blocked and those used as covariables.

Willingness to Confront. The subjects completing the

ratings are to be blocked on a measure of their willingness to confront. Differences in willingness to confront are thought to contribute to bias in ratings, especially when one perceives the rating as leading to a threatening or unpleasant situation. Willingness for confrontation emerged in a factor analysis of semantic differential (Osgood, 1957) data as a measurable trait. Operationally, it will be defined in this study by that factor.

Concept of Average. A possible reason for ineffective ratings is a failure on the part of raters to specify their concept of average. In spite of definitions of "absolute ratings" as opposed to "comparative ratings", all ratings are compared to some anchoring mechanisms. Normally this anchoring is to a hypothetical middle or average. Johnson Abercrombie (1960) indicates that this hypothetical average differs greatly among individuals. She discusses the need to develop a course to establish the meaning of "average" or "normal" with a group of medical students. Such a concept is alluded to in several studies which discuss the attempt to "capture the rater policy" (Houston & Roscoe, 1969, Christal, 1968). Such a concept is further supported by Guilford's "error of leniency" and "error of central tendency" (Guilford, 1954). Each rater will be requested to complete a rating form on his concept of "average teacher" as a measure of the raters' concept of average and used as a covariable to reduce across rater variance.

Concept of dispersion. Guilford (1954) reports a general tendency for some raters to utilize only middle scales on rating forms while others tend to utilize extremes. Using the rating form completed by each rater for their concept of "average teacher", a measure of rater variance will be calculated to be utilized as a covariable to reduce across rater variance.

Dependent Variable. The dependent variable is the response or output of each rater on the rating form. The term "rating" will utilize a numerical score as a unit of measurement. Each item on the rating form will be scored from one to nine and scores will be summed for each level of abstraction across the items representing that level of abstraction. For the analysis by item, each item score will be considered separately.

C H A P T E R I I

REVIEW OF THE LITERATURE

Although the literature is filled with studies of teacher ratings and of rating scales, few if any of these actually focus upon the rating process. In most studies of teacher rating, ratings are used as criteria for teacher behavior variables. Studies dealing with rating scales generally deal with scale development and reliability and validity procedures. Since this study is primarily focused upon the rating process, the review of literature will also focus upon the process.

The psychological literature is replete with laboratory studies of response variables wherein rating instruments are employed by judges to judge sensory and personality type variables. Such laboratory experiments have dealt largely with variables which influence judges, anchoring effects of rating scales, and other measures relating input to output variables. Little can be found describing the judgment process itself. Such studies are difficult to perform because of the hypothetical constructs necessary to connect the input and output variables.

Johnson (1955) has collected most of the studies relevant to the process of judgment and has developed a list of seven principles dealing with the judgment aspect of the rating

process. He supports his list of general principles of judgment with empirical research and experimentation.

- "1. The judgment may be influenced by stimulus aspects or variables to which attention is not directed by the explicit instructions or by logical implications of the stimulus material.
- "2. The judgment may be influenced by stimulus variables which the judge cannot or does not report.
- "3. Most judges give extra weight to stimulus material, or suggestions, attributed to people whom they regard as experts.
- "4. Most judges give extra weight to stimulus material, or suggestions, attributed to the majority of a group with whom they identify.
- "5. When the judgment called for is avoided, because of its difficulty or for any other reason, judgment of some other stimulus aspect will be made.
- "6. When an abstract judgment is called for, an affective judgment is commonly given.
- "7. When the judgment called for is straightforward and easy, the judge is likely to maintain his set for that judgment and not be diverted into judging something else."

Similarly, Guilford (1954) describes potential errors in rating as follows:

- "Error of leniency...the preference here is to use the term 'leniency error' to apply to a general, constant tendency for a rater to rate too high or too low for whatever reason.
- "Error of central tendency...raters hate to give extreme judgments and thus tend to displace individuals in the direction of the mean of the total group.
- "Halo effect...to force the rating of any trait in the direction of the general impression of the individuals rated.

"Logical error in rating...judges are likely to give similar ratings for traits that seem logically related in the minds of the raters.

"Contrast error...a tendency for a rater to rate others in the opposite direction from himself in a trait.

"Proximity error...adjacent traits on a rating form tend to intercorrelate higher than remote ones, their degree of similarity being presumably equal."

Adjustment of data to remove bias in rating has been rarely attempted. Guilford (1954) describes a procedure for estimating certain of the constant errors and adjusting the data to eliminate these errors. His procedure results in equal rater means by utilizing within data adjustment. Item intercorrelations are changed with his procedure, however; which he explains as removal of "halo errors". Such changes would seem to be removal of "true score" effects which reduce rater discrimination. It would appear more useful to obtain outside measures for adjusting the data.

Measurement of certain rater characteristics should prove useful in adjusting rating data for removal of rating errors. Leniency of rating has been specified by Guilford (1954) as an error prevalent in the rating process. Because this error is associated with the judgment base of the rater's interval processes, it is suggested that the rater's concept of average might identify the leniency tendency (with leniency acknowledged as deviating in either direction from the central position). A measure of the rater's concept of average might

provide a tool for adjusting leniency error to reduce variance across raters, thereby, approaching a true score measurement.

It has also been reported by Guilford (1954) that some raters tend to rate only central categories while other may tend toward a two-valued orientation (Hayakawa, 1940), and rate only extremes. A measure of concept of spread or extreme-ness of ratings should be a useful way of further reducing variance across raters to remove what has been termed by Guilford an "error of central tendency".

Following adjustment of the data, output may be examined by experimentally manipulating or by measuring certain other aspects of the rating process. Raters' knowledge of the use of rating has been suggested by Guilford (1954) as a cause of systematic bias in ratings. It is suggested that the more threatening the use to which the rating will be subjected, the greater the artificial inflation of the rating. When ratings pose the least amount of threat to either the rater or ratee, the more will the rating reflect a true score. Guilford suggests that ratings be secured with the raters ignorant of the use to be made of the ratings. Possible uses may be surmised by the raters in such instances and an unknown element of bias would thereby be interjected.

Errors caused by knowledge of the intended purpose of rating use by the rater are possibly caused by degree of the

potential threat to the rater or the ratee. Such threat is felt to differ among raters in some fashion connected to the rater's willingness for confrontation. It is thought that some of this leniency error bias might be accounted for by blocking on a measure of willingness for confrontation. Such blocking should serve to further reduce rating error.

Abstraction level or level of subjectiveness of rating items appears to contribute to several types of error (Guilford, 1954; Johnson, 1955). Teaching, however, is a complex process, the description of which involves many abstract concepts. If abstraction level were varied, one might obtain an empirical assessment of the degree to which this variable influences errors in rating. Guilford's (1954, p. 296) "rules concerning traits" speak clearly to the abstraction level question. Generally, the rules may be summarized by stating that objective descriptions of traits, limiting a rating to a single trait, reference to an activity or result of an activity, and limiting ratings to past or present accomplishments not future promise, will assure more accurate ratings.

THE HYPOTHESES

To test the hypotheses for this study, three types of analyses are necessary. First are tests of reduction of variance. An ANOVA will be performed five times--on raw data, on rating scores adjusted linearly for concept of average, on rating scores adjusted for average only using a proportional

method, on rating scores adjusted for average and dispersion using a Z-score transformation, and on rating scores adjusted through item covariance for average and dispersion. Each time the ANOVA is calculated, a reliability coefficient will be computed to test the reduction of variance across raters within item.

Second, hypotheses will be tested to confirm that adjustment of the data has not removed variance due to differences in teachers. Simultaneously the data will be tested concerning increase in variance across items resulting from adjustment of the data. Interteacher variance must be retained or ability of the instrument to discriminate may be seriously reduced. Also, across item variance should not be increased as a sacrifice for gaining reliability within items.

The third category of testing involves experimental effects hypotheses concerning the remaining variables. These hypotheses are to be tested on that set of data selected through the above tests; either raw rating scores or one of the adjustments depending on the data yielding the highest reliability.

The null hypotheses with which this study is concerned are grouped according to the three types of analyses to be performed. The parameters being tested are defined as follows:

α = the effects of levels of willingness to confront,

β = the effects of levels of purpose of rating,

γ = the effects of teachers (tapes),

δ = the effects of abstraction level of items.

Reduction of Variance.

- I. There will be no difference between reliability of raw scores (r_{XX}) and reliability of scores adjusted by covarying raters' concept of average ($r_{XX(a)}$).

$$H_0: r_{XX} = r_{XX(a)}$$

- II. There will be no difference between reliability of raw scores (r_{XX}) and reliability of scores adjusted by covarying raters' concept of average and concept of dispersion ($r_{XX(ad)}$).

$$H_0: r_{XX} = r_{XX(ad)}$$

Confirmation of intertape variance.

- III. There will be no difference between ratings of teachers (tapes).

$$H_0: \gamma_1 = \gamma_2 = \gamma_3$$

Experimental Effects.

- IV. There will be no difference between scores of raters classified as low willingness to confront and scores of raters classified as high willingness to confront.

$$H_0: \alpha_1 = \alpha_2$$

- V. There will be no difference in rating scores among the three levels of purpose of rating.

$$H_0: \beta_1 = \beta_2 = \beta_3$$

- VI. There will be no difference in rating scores among the three abstraction levels.

$$H_0: \delta_1 = \delta_2 = \delta_3$$

VII. There will be no interaction of levels of willingness to confront and levels of purpose of rating.

$$H_0: \alpha\beta_{11} = \alpha\beta_{12} = \alpha\beta_{13} = \alpha\beta_{21} = \alpha\beta_{22} = \alpha\beta_{23}$$

VIII. There will be no interaction of levels of willingness to confront and levels of teachers.

$$H_0: \alpha\gamma_{11} = \alpha\gamma_{12} = \alpha\gamma_{13} = \alpha\gamma_{21} = \alpha\gamma_{22} = \alpha\gamma_{23}$$

IX. There will be no interaction of levels of willingness to confront and levels of abstraction.

$$H_0: \alpha\delta_{11} = \alpha\delta_{12} = \alpha\delta_{13} = \alpha\delta_{21} = \alpha\delta_{22} = \alpha\delta_{23}$$

X. There will be no interaction of levels of willingness to confront, levels of purpose of rating, and levels of teachers.

$$\begin{aligned} H_0: \alpha\beta\gamma_{111} &= \alpha\beta\gamma_{112} = \alpha\beta\gamma_{113} = \alpha\beta\gamma_{121} = \alpha\beta\gamma_{122} = \alpha\beta\gamma_{123} \\ &= \alpha\beta\gamma_{131} = \alpha\beta\gamma_{132} = \alpha\beta\gamma_{133} = \alpha\beta\gamma_{211} = \alpha\beta\gamma_{212} = \alpha\beta\gamma_{213} \\ &= \alpha\beta\gamma_{221} = \alpha\beta\gamma_{222} = \alpha\beta\gamma_{223} = \alpha\beta\gamma_{231} = \alpha\beta\gamma_{232} = \alpha\beta\gamma_{233} \end{aligned}$$

XI. There will be no interaction of levels of willingness to confront, levels of purpose of rating, and levels of abstraction.

$$\begin{aligned} H_0: \alpha\beta\delta_{111} &= \alpha\beta\delta_{112} = \alpha\beta\delta_{113} = \alpha\beta\delta_{121} = \alpha\beta\delta_{122} = \alpha\beta\delta_{123} \\ &= \alpha\beta\delta_{131} = \alpha\beta\delta_{132} = \alpha\beta\delta_{133} = \alpha\beta\delta_{211} = \alpha\beta\delta_{212} = \alpha\beta\delta_{213} \\ &= \alpha\beta\delta_{221} = \alpha\beta\delta_{222} = \alpha\beta\delta_{223} = \alpha\beta\delta_{231} = \alpha\beta\delta_{232} = \alpha\beta\delta_{233} \end{aligned}$$

XII. There will be no interaction of levels of willingness to confront, levels of teachers, and levels of abstraction.

$$\begin{aligned}
 H_0: \alpha\gamma\delta_{111} &= \alpha\gamma\delta_{112} = \alpha\gamma\delta_{113} = \alpha\gamma\delta_{121} = \alpha\gamma\delta_{122} = \alpha\gamma\delta_{123} \\
 &= \alpha\gamma\delta_{131} = \alpha\gamma\delta_{132} = \alpha\gamma\delta_{133} = \alpha\gamma\delta_{211} = \alpha\gamma\delta_{212} = \alpha\gamma\delta_{213} \\
 &= \alpha\gamma\delta_{221} = \alpha\gamma\delta_{222} = \alpha\gamma\delta_{223} = \alpha\gamma\delta_{231} = \alpha\gamma\delta_{232} = \alpha\gamma\delta_{233}
 \end{aligned}$$

XIII. There will be no interaction of levels of willingness to confront, levels of purpose of rating, levels of teachers, and levels of abstraction.

$$\begin{aligned}
 H_0: \alpha\beta\gamma\delta_{1111} &= \alpha\beta\gamma\delta_{1112} = \alpha\beta\gamma\delta_{1113} = \alpha\beta\gamma\delta_{1121} = \alpha\beta\gamma\delta_{1122} \\
 &= \alpha\beta\gamma\delta_{1123} = \alpha\beta\gamma\delta_{1131} = \alpha\beta\gamma\delta_{1132} = \alpha\beta\gamma\delta_{1133} = \alpha\beta\gamma\delta_{1211} \\
 &= \alpha\beta\gamma\delta_{1212} = \alpha\beta\gamma\delta_{1213} = \alpha\beta\gamma\delta_{1221} = \alpha\beta\gamma\delta_{1222} = \alpha\beta\gamma\delta_{1223} \\
 &= \alpha\beta\gamma\delta_{1231} = \alpha\beta\gamma\delta_{1232} = \alpha\beta\gamma\delta_{1233} = \alpha\beta\gamma\delta_{1311} = \alpha\beta\gamma\delta_{1312} \\
 &= \alpha\beta\gamma\delta_{1313} = \alpha\beta\gamma\delta_{1321} = \alpha\beta\gamma\delta_{1322} = \alpha\beta\gamma\delta_{1323} = \alpha\beta\gamma\delta_{1331} \\
 &= \alpha\beta\gamma\delta_{1332} = \alpha\beta\gamma\delta_{1333} = \alpha\beta\gamma\delta_{2111} = \alpha\beta\gamma\delta_{2112} = \alpha\beta\gamma\delta_{2113} \\
 &= \alpha\beta\gamma\delta_{2121} = \alpha\beta\gamma\delta_{2122} = \alpha\beta\gamma\delta_{2123} = \alpha\beta\gamma\delta_{2131} = \alpha\beta\gamma\delta_{2132} \\
 &= \alpha\beta\gamma\delta_{2133} = \alpha\beta\gamma\delta_{2211} = \alpha\beta\gamma\delta_{2212} = \alpha\beta\gamma\delta_{2213} = \alpha\beta\gamma\delta_{2221} \\
 &= \alpha\beta\gamma\delta_{2222} = \alpha\beta\gamma\delta_{2223} = \alpha\beta\gamma\delta_{2231} = \alpha\beta\gamma\delta_{2232} = \alpha\beta\gamma\delta_{2233} \\
 &= \alpha\beta\gamma\delta_{2311} = \alpha\beta\gamma\delta_{2312} = \alpha\beta\gamma\delta_{2313} = \alpha\beta\gamma\delta_{2321} = \alpha\beta\gamma\delta_{2322} \\
 &= \alpha\beta\gamma\delta_{2323} = \alpha\beta\gamma\delta_{2331} = \alpha\beta\gamma\delta_{2332} = \alpha\beta\gamma\delta_{2333}
 \end{aligned}$$

XIV. There will be no interaction of levels of purpose of rating and levels of teachers.

$$\begin{aligned}
 H_0: \beta\gamma_{11} &= \beta\gamma_{12} = \beta\gamma_{13} = \beta\gamma_{21} = \beta\gamma_{22} = \beta\gamma_{23} = \\
 &\beta\gamma_{31} = \beta\gamma_{32} = \beta\gamma_{33}
 \end{aligned}$$

XV. There will be no interaction of levels of purpose of rating and levels of abstraction.

$$\begin{aligned}
 H_0: \beta\delta_{11} &= \beta\delta_{12} = \beta\delta_{13} = \beta\delta_{21} = \beta\delta_{22} = \beta\delta_{23} = \\
 &\beta\delta_{31} = \beta\delta_{32} = \beta\delta_{33}
 \end{aligned}$$

XVI. There will be no interaction of levels of purpose of rating, levels of teachers, and levels of abstraction.

$$\begin{aligned}
 H_0: \quad & \beta\gamma\delta_{111} = \beta\gamma\delta_{112} = \beta\gamma\delta_{113} = \beta\gamma\delta_{121} = \beta\gamma\delta_{122} = \beta\gamma\delta_{123} \\
 & = \beta\gamma\delta_{131} = \beta\gamma\delta_{132} = \beta\gamma\delta_{133} = \beta\gamma\delta_{211} = \beta\gamma\delta_{212} = \beta\gamma\delta_{213} \\
 & = \beta\gamma\delta_{221} = \beta\gamma\delta_{222} = \beta\gamma\delta_{223} = \beta\gamma\delta_{231} = \beta\gamma\delta_{232} = \beta\gamma\delta_{233} \\
 & = \beta\gamma\delta_{311} = \beta\gamma\delta_{312} = \beta\gamma\delta_{313} = \beta\gamma\delta_{321} = \beta\gamma\delta_{322} = \beta\gamma\delta_{323} \\
 & = \beta\gamma\delta_{331} = \beta\gamma\delta_{332} = \beta\gamma\delta_{333}
 \end{aligned}$$

XVII. There will be no interaction of levels of teachers and levels of abstraction.

$$\begin{aligned}
 H_0: \quad & \gamma\delta_{11} = \gamma\delta_{12} = \gamma\delta_{13} = \gamma\delta_{21} = \gamma\delta_{22} = \gamma\delta_{23} = \\
 & \gamma\delta_{31} = \gamma\delta_{32} = \gamma\delta_{33}
 \end{aligned}$$

If any main effects or interaction effects are significant, i.e., if the null hypothesis is rejected, $p < .05$, a multiple range test will be made between treatment means. Following Winer (1962), a Newman-Kuels modified "q" statistic will be used.

CHAPTER III

PROCEDURES

This chapter will discuss the subjects used in the study, the methods for collecting and analyzing the data, and the instrumentation involved in the study. Detailed instructions to raters and a copy of the instruments used are included in Appendices A, B, and C.

Subjects.

Generalizability required that this study be performed with raters who are involved in rating teachers or who will become involved in rating teachers. Utilizing graduate students from the University of Massachusetts, School of Education, Center for Teacher Education and Center for Administration, satisfied these requirements. The forty-eight subjects recruited for this study were enrolled in courses within those centers and were either currently teacher supervisors or potentially teacher supervisors. Though predominantly male, there were nine female subjects, probably representative of the proportion of female supervisors in the total population.

Method.

Data Collection.

Videotapes. Three male science teachers in departmentalized elementary schools were videotaped while teaching actual 4th, 5th, or 6th grade classes. Though each class was taped for varying lengths of time, an eight minute clip of each of the three classes was selected for stimuli for the experiment. All teachers selected were unknown to the raters and represented different levels (quality) of teaching based on a subjective judgment. This judgment was later confirmed by the rating data collected. Videotapes were randomly ordered and shown in the same order to all subjects.

Willingness to confront. Each rater was administered a Semantic Differential booklet (see appendix A) which contained the concept "Willingness for confrontation" with ten bipolar scales. From development of the semantic differential, it had been determined that a factor relating to willingness to confront existed in eight of these ten scales. The median for the total of eight scales under this concept was used to divide the group of raters into "high" and "low" willingness levels.

Concepts of Average and of Dispersion. Prior to rating the tapes, two measures were taken of the rater's concept of average. The semantic differential, described above, also contained ten bipolar scales under the concept of "Ordinary Teacher". Factor analysis performed during development of the scale identified a factor using nine of the ten scales.

This measure was taken as one of the scores for "Average Teacher".

Following classification of subjects into "High" or "Low" willingness to confront, prior to rating the tapes, each subject was asked to complete a rating scale on their concept of average teacher (instructions are contained in appendix B). The mean ratings of average teacher were used as measures of each raters' concept of average teacher and the standard deviations as their concept of dispersion.

Purpose of Rating. Subsequent to assignment of each rater to a level of willingness to confront, raters were randomly assigned to a purpose of rating, representing three levels thought to simulate levels of threat to the rater. These purposes, described in appendix B, were "research only", "Administrative", and "Diagnostic" as defined by the instructions provided each rater. Each rater was provided these instructions and asked to read them carefully before the videotapes were exposed. Attached to each instruction were three rating scale forms for rating of the three teachers.

Abstraction Level. Each rating scale contained eighteen rating items, six for each of three levels of abstraction (low, medium, and high). Each subject rated all three teachers (videotapes) and their concept of average teacher on all 18 items.

Data Analysis.

Rater Variance. Rating scores were adjusted in two

manners, linear adjustment and covariant adjustment, for measures obtained on "Average Teacher". Correlational results from the semantic differential measure of "Ordinary Teacher" indicated that this measure was of no value for adjusting data, therefore, all adjustments were made with the measure on "Average Teacher" ratings. Details of the adjustments are provided in the next chapter dealing with results of the analysis. The purpose of adjusting was to remove effects of the raters' concept of average, and a combination of their concepts of average and of dispersion from the ratings of each individual teacher.

To examine the effects of removal of these concepts across raters, a four-way analysis of variance for reliability was performed as described in Medley and Mitzel (1963). The analysis of variance was performed five times: on raw rating scores, scores linearly adjusted for concept of average, scores covaried for concept of average with proportional variance by items, scores covaried for concepts of average and of dispersion using correlation of rater means and standard deviations to obtain a Z-score transformation, and scores covaried for concepts of average and of dispersion using item score residuals. The reliability analysis of variance design in general form is shown in Table 1, below, with the components used to estimate true score and observed score for calculation of a reliability coefficient (Medley and Mitzel, 1963).

TABLE 1
 DESIGN FOR RELIABILITY ANALYSIS OF SCORES BASED ON
 RATINGS MADE BY 48 RATERS OF 3 TEACHERS ON 18 ITEMS

SOURCE OF VARIATION	DF	OBTAINED		EXPECTED MEAN SQUARES
		MEAN	SQUARE	
Teachers	(t-1) 2	S_t^2	$864\sigma_t^2 + 18\sigma_{ti}^2 + \sigma^2$	
Raters	(N-1) 47	S_r^2	$54\sigma_r^2 + 18\sigma_{ti}^2 + 3\sigma_{ri}^2 + \sigma^2$	
Items	(i-1) 17	S_i^2	$144\sigma_i^2 + 3\sigma_{ri}^2 + \sigma^2$	
Teachers X Raters	(t-1)(N-1) 94	S_{tr}^2	$18\sigma_{tr}^2 + \sigma^2$	
Teachers X Items	(t-1)(i-1) 34	S_{ti}^2	$48\sigma_{ti}^2 + \sigma^2$	
Raters X Items	(N-1)(i-1) 799	S_{ri}^2	$3\sigma_{ri}^2 + \sigma^2$	
Residual	(t-1)(N-1)(i-1) 1598	S^2	σ^2	
TOTAL	(Ntri-1) 2591			

COMPONENTS

- (1) $\sigma_t^2 = \frac{1}{ri}(S_t^2 - S_{tr}^2 - S^2)$ = Estimate of True Score Variance
 (2) $\sigma_{tr}^2 = \frac{1}{i}(S_{tr}^2 - S^2)$ = Component of Observed Score Variance
 (3) $\sigma_{ti}^2 = \frac{1}{r}(S_{ti}^2 - S^2)$ = Component of Observed Score Variance
 (4) $\sigma^2 = S^2$ = Component of Observed Score Variance

Where:

$$\rho_{XX} = \frac{\sigma_T^2}{\sigma_X^2} \text{ or } r_{ri} = \frac{(ri)^2 \sigma_t^2}{ri(ri\sigma_t^2 + i\sigma_{tr}^2 + r\sigma_{ti}^2 + \sigma^2)}$$

The highest reliability coefficient would indicate data with the least interrater variance and thereby identify which data was to be utilized in examining experimental effects.

Experimental Effects. Both raw data and data which had been adjusted through Z-score covariance were examined for experimental effects. Main effects and interaction effects were obtained from an Analysis of Variance based on the following linear structural model (Dayton, 1970):

$$Y_{ijkmg} = \mu + \alpha_k + \beta_j + \pi_i(jk) + \gamma_m + \delta_g + \alpha\beta_{kj} + \alpha\gamma_{km} + \alpha\delta_{kg} + \beta\gamma_{jm} + \beta\delta_{jg} + \gamma\delta_{mg} + \alpha\beta\gamma_{kjm} + \alpha\beta\delta_{kjg} + \alpha\gamma\delta_{kmg} + \beta\gamma\delta_{jmg} + \alpha\beta\gamma\delta_{kjmg} + \gamma\pi_{mi}(jk) + \delta\pi_{gi}(jk) + \gamma\delta\pi_{mgi}(jk) + \epsilon_{ijkmg}$$

Where:

Y_{ijkmg} = The observed rating based on i raters nested within j purposes and k levels of willingness to confront with repeated measures on m teachers and g levels of abstraction,

μ = the mean rating,

α = the effects of levels of willingness to confront,

β = the effects of levels of purpose of rating,

γ = the effects of teachers (tapes),

δ = the effects of abstraction level of items,

π = the effects of the individual raters, and

ϵ = the unaccounted for error.

The general model for the ANOVA is shown in Table 2, below:

TABLE 2
ANALYSIS OF VARIANCE OF EXPERIMENTAL EFFECTS

SOURCE OF VARIATION		DF	MEAN SQUARE EFFECTS
Willingness to Confront(A)	(k-1)	1	Main effect for A
Purpose of Rating (B)	(j-1)	2	Main effect for B
Teachers (C)	(m-1)	2	Main effect for C
Abstraction Level (D)	(g-1)	2	Main effect for D
A X B	(k-1)(j-1)	2	Interaction A X B
A X C	(k-1)(m-1)	2	Interaction A X C
B X C	(j-1)(m-1)	4	Interaction B X C
A X D	(k-1)(g-1)	2	Interaction A X D
B X D	(j-1)(g-1)	4	Interaction B X D
C X D	(m-1)(g-1)	4	Interaction C X D
Subjects:A X B	kj(i-1)	42	Error term for A, B, & A X B
A X B X C	(k-1)(j-1)(m-1)	4	Interaction A X B X C
A X B X D	(k-1)(j-1)(g-1)	4	Interaction A X B X D
A X C X D	(k-1)(m-1)(g-1)	4	Interaction A X C X D
B X C X D	(j-1)(m-1)(g-1)	8	Interaction B X C X D
Subjects X C:A X B	kj(i-1)(m-1)	84	Error term for C, A X C, B X C, and A X B X C
Subjects X D:A X B	kj(i-1)(g-1)	84	Error term for D, A X D, B X D, and A X B X D
A X B X C X D	(k-1)(j-1)(m-1)(g-1)	8	Interaction A X B X C X D
Subjects X C X D:A X B	kj(i-1)(m-1)(g-1)	168	Error term for C X D, A X C X D, B X C X D, and A X B X C X D
TOTAL	(ijkmg-1)	431	

Instrumentation.

Semantic Differential. The semantic differential, developed by Osgood, Suci, and Tannenbaum (1957), has been extensively used as a method of tapping connotative meanings. In the scale, which is made up of a series of bipolar adjectives, certain of the adjectives were found to have high evaluative loadings when applied to many objects or concepts. Utilizing factor analytic techniques to sort out adjectives and concepts which describe a factor in an evaluative manner, the semantic differential has been accepted as a valid measurement technique (Tannenbaum, 1956; Brinton, 1969).

From Osgood, et. al. (1957) a list of nine adjective pairs were selected which reported high loadings on the evaluative factor. Because the concepts relating to confrontation seemed to be concerned with approach-avoidance, one additional scale was added, approach-avoid, even though it violated the traditional use of the semantic differential in that it was a verb pair not an adjective pair. These ten scales were then matched to six concepts selected by the experimenter in consultation with his advisor; three concepts represented various aspects of average and three represented various connotations of willingness to confront. This semantic differential was administered to 168 undergraduate students in education at the University of Massachusetts, School of Education, a somewhat similar population from which the subjects for the study were

TABLE 3

FACTOR LOADINGS FROM SEMANTIC DIFFERENTIAL

BI-POLAR SCALES	C O N C E P T S											
	ORDINARY TEACHER		AVERAGE PERSON		AVERAGE COST		ARGUMENT		WILLINGNESS FOR CON- FRONTATION		HELPING A PERSON IN TROUBLE	
	I	II	I	II	I	II	I	II	I	II	I	II
Good-bad	.80	.04	.31	-.02	.08	-.16	.03	-.05	.03	-.86	.05	-.06
Strong-weak	.77	.14	.33	.02	.08	-.08	.13	-.06	.01	-.79	.11	-.11
Active-passive	.83	.06	.41	-.04	-.13	-.14	.11	-.02	.01	-.81	.04	-.25
Hard-soft	-.07*	.00	-.08	-.01	-.11	-.05	.05	-.03	.09	-.38*	-.07	-.04
Clean-dirty	.55	-.01	.22	-.02	.06	.04	-.02	.03	.04	-.38*	.09	-.02
Valuable-worthless	.83	.14	.35	-.01	-.10	-.12	-.06	.01	-.06	-.81	-.12	-.15
Pleasant-unpleasant	.81	-.05	.25	-.09	.01	-.03	-.02	.04	-.05	-.49	.08	.05
Honest-dishonest	.73	-.01	.10	-.05	-.12	-.03	.11	-.02	-.12	-.66	-.12	-.10
Positive-negative	.81	-.07	.34	-.03	-.01	-.11	-.00	.08	-.17	-.74	-.07	-.12
Approach-avoid	.72	.08	.14	-.03	.06	-.26	.06	.01	-.00	-.80	-.10	-.11

NOTE: Factor I, "Ordinary Teacher" composed of nine scales under the concept ordinary teacher. *scale not used.

Factor II, "Willingness to Confront" composed of eight scales under the concept willingness for confrontation. *scales not used.

drawn. The concepts and scales with loadings obtained from orthogonally rotated factor analysis are shown in Table 3, below. The two strongest of the ten factors which emerged when the diagonals were set to one and a limit of ten factors called for, were described as "Concept of Ordinary Teacher" and "Concept of Willingness to Confront". All other factors were ignored as being of no interest in this study.

The concept of average teacher was cross-validated through use of the rating instrument described below.

Rating Scale. A rating scale was developed which provided three levels of abstraction of items. Because abstraction level of items is a matter of judgment, the scale was developed in the following manner: 90 items were gleaned from commonly used rating scales and the items were sorted by ten persons into five categories or levels of abstraction. The ten persons represented both teachers and non-teachers, supervisors and non-supervisors. Using a scaling method involving Torgerson's (1958) law of categorical judgment, the least squares solution to arrive at scale values was employed. Items with scale values ranging nearest the median and items with scale values on each the highest and lowest end of the scale were then selected for further scaling.

One hundred and forty-six junior and senior practice teachers were given forty-seven items selected from the above procedure as an instrument on which they were requested to

categorize each of the forty-seven items into one of five categories. These categories, as before, ranged from most abstract to most concrete. A number of these subjects failed to complete the entire instrument resulting in a total of 127 completed forms which were then used to develop the scale values for the items. Category boundaries of the 47 scale values were found using Torgerson's (1958) law of categorical judgment, Condition D, Class II. (Condition D assumes that dispersion of stimuli is constant, dispersion of category boundaries is constant and that correlations between stimuli and boundaries are constant. Class II involves replication over individuals, each stimuli being presented once to each subject.) The six items which had the highest scale values, the six nearest the median scale value, and the six with the lowest scale values were selected to represent low, medium, and high abstraction levels.

A copy of the instrument with scale values for abstraction levels is included in Appendix C. As stated above, the instrument utilizes items from commonly used teacher rating scales, however, format changes have been made to provide uniformity of item format.

CHAPTER IV

RESULTS

This chapter will deal with the data collected to answer the questions posed in the problem statement, namely: Are there effects of knowledge of output use of ratings on raters? Are there effects of varied levels of abstraction of rating items on rater output? Can knowledge of rater concept of average be used to reduce interrater variance? Can knowledge of rater differences in dispersion (use of extreme rating categories) be used to reduce interrater variance? Does rater willingness to confront interact with rating output use?

In dealing with these questions, results will be presented in terms of the hypotheses as stated in the null form. The order of presentation corresponds to the order in which the data were analyzed as described in the previous chapter.

REDUCTION OF VARIANCE

Concept of Average.

Two measurements were attempted on concept of average for each rater. The first measurement was obtained from a factor identified on a semantic differential score for the concept of "Ordinary Teacher". The second measurement was obtained from each rater's rating of their concept of "Average Teacher" on the identical rating form used for the rating of each teacher.

Ordinary Teacher. The mean semantic differential score across all raters was 58.46 when a seven point scale was summed across nine scales. The rating of 6.50 for each scale indicates a positive effect toward the ordinary teacher (4 being the neutral point). A standard deviation of 5.71 for the nine scales summed represents less than two-thirds units per scale; very little deviations of ratings.

However, intercorrelation of the "Ordinary Teacher" concept with "Average Teacher" ratings and ratings of each of the other teachers approximated zero, indicating that the semantic differential score represented a different measure than the "Average Teacher" rating. Intercorrelations are shown in Table 5. No further use was made of these data except as support for the non acceptance of the null hypotheses discussed below.

Average Teacher. Means and standard deviations of the ratings of Average Teacher by all raters is shown in Table 4. It should be noted that the medium abstraction level scores were higher than either the low or high abstraction level.

Each raters' concept of the average teacher was correlated with his total rating of each of the three teachers. Results of these correlations are shown in Table 5, along with the correlation coefficients generated between "Ordinary Teacher" concept and each of the ratings. The only significant correlation was between Teacher 1 and Teacher 2.

TABLE 4
DESCRIPTION OF ITEM RATINGS OF AVERAGE TEACHER
ABSTRACTION LEVEL

ITEM	LOW		MEDIUM		HIGH	
1	5.77	1.72	5.77	1.81	6.00	1.81
2	4.58	1.80	5.85	1.07	4.94	2.14
3	4.92	2.21	6.02	1.64	5.29	1.47
4	6.15	1.49	6.58	1.92	5.48	1.97
5	6.35	1.23	5.63	1.61	5.98	1.67
6	5.79	1.49	6.15	1.27	5.54	1.49
TOTAL	33.56	7.30	36.00	6.82	33.23	7.83

TABLE 5
CORRELATION COEFFICIENTS AMONG
AVERAGE TEACHER MEASURES AND EACH TEACHER TOTAL SCORES

	AVERAGE TEACHER	TEACHER 1	TEACHER 2	TEACHER 3
Ordinary Teacher	0.0282	0.0795	0.0628	0.0487
Average Teacher		-0.0978	-0.1919	-0.2159
Teacher 1			0.4586***	0.2233
Teacher 2				0.0802

***p < .001

N=48

Reliability.

Data Adjustments. The rating data was adjusted through four methods which will be referred to as Linear adjustment, Proportional adjustment, Z-score transformation, and Item Residual adjustment.

Linear adjustment. Raw item scores for each raters' rating of "Average Teacher" were subtracted from raw item scores for each raters' rating of each teacher. To avoid negative numbers, a constant of nine was then added to the difference. This yielded a set of scores for each item for each teacher by each rater called linear adjusted scores.

Proportional adjustment. Using the correlations of means of "Average Teacher" with each teacher across all raters, a covariance adjustment was made of each teachers' mean rating for each rater. This covariance adjustment was made to remove any variance due to differences in concept of average among raters. The predicted (adjusted mean score for each rater on each teacher) was then apportioned to each of the items using that proportion in the original raw rating attributed to each item. The resultant item scores for each teacher by each rater is referred to as proportional adjusted score.

Z Transformed scores. To adjust the data for differences in both mean and dispersion characteristics of raters, a Z-score was computed. The means and standard deviations of "Average Teacher" and each individual teacher were correlated.

The resultant two correlation coefficients ($r_{\bar{X}\bar{A}}$ and $r_{S_{\bar{X}}S_{\bar{A}}}$) were used to predict new item scores for each teacher as rated by each rater using the following formula:

$$\hat{X}_{tij} = S_{ti} Z_{tij} + \bar{X}_{ti}$$

Where:

\hat{X}_{tij} = Predicted score for each teacher (t) for each rater (i) for each item (j).

S_{ti} = the covariant adjustment for standard deviation from $r_{S_{\bar{X}}S_{\bar{A}}}$, for each teacher and each rater.

\bar{X}_{ti} = the covariant adjustment for means from $r_{\bar{X}\bar{A}}$, for each teacher and each rater.

$$Z_{tij} = \frac{X_{tij} - \bar{X}_{ti}}{S_{ti}}$$

Where:

X_{tij} = Each raw rating score (tij)

\bar{X}_{ti} = Mean scores (ti)

The resultant adjusted score is referred to as Z transformed score.

Item Residual adjustment. Each rater's item score was adjusted through covarying for the relationship between all raters' average teacher item scores and each teachers' item score. This was accomplished by calculating the coefficient of correlation between each item on average teacher (across raters) and each item on each teacher. The resultant 64 coefficients were used with standard deviations of average teacher and each

teacher as a beta weight for calculating residual item scores for each rater. These scores are referred to as the item residual adjustment.

Reliability. Two hypotheses concerning reliability were stated:

HYPOTHESIS I: There will be no difference between reliability of raw scores (r_{xx}) and reliability of scores adjusted by covarying raters' concept of average ($r_{xx(a)}$).

$$H_0: r_{xx} = r_{xx(a)}$$

HYPOTHESIS II: There will be no difference between reliability of raw scores (r_{xx}) and scores adjusted by covarying raters' concepts of average and dispersion ($r_{xx(ad)}$).

$$H_0: r_{xx} = r_{xx(ad)}$$

To test for reduction of variance across raters within items, an analysis of variance was performed on each of the five sets of data, raw and four adjusted scores, following Medley and Mitzel's (1963) procedure for estimating true score and observed score variance. It was assumed that a decrease in interrater variance would be reflected in an increase in reliability. Table 6, reflects that reliability coefficients within abstraction levels were virtually the same for raw scores, linear adjustments, and proportional adjustments. A slight increase was found in the Z transformed data and a

slight decrease in item residual adjustment data. When calculations were performed on each abstraction level independently, results were mixed.

TABLE 6
RELIABILITY COEFFICIENTS

RELIABILITY	SCORES ADJUSTED BY				
	RAW SCORES	LINEAR	PROPOR- TIONAL	Z TRANS- FORMATION	ITEM RESIDUAL
Within Abstraction Level	0.9750	0.9750	0.9752	0.9763	0.9679
Abstraction Level 1	0.9411	0.9411	0.9417	0.9410	0.9404
Abstraction Level 2	0.9646	0.9646	0.9649	0.9666	0.9548
Abstraction Level 3	0.9556	0.9556	0.9562	0.9559	0.9514

To assure that the adjustments had not reduced differences across teachers, which would have voided the discriminating ability of the data, and to verify that adjustments had not increased variance across items, an additional test was performed. The procedure was to calculate an ANOVA using variances as data in the following manner:

Variations for each item across raters were calculated for each of the three teachers on each of the four sets of data. (Item residual adjustment data was dropped due to the reduction in reliability.) A two-way analysis of variance yielded $F(2,204) = 3.27$ for teachers and $F(3,204) = 88.54$ for the four

types of data. Differences among the four teachers remained significant demonstrating that discrimination existed after adjustment. The F-value for data, significant $p < .001$, indicates that variance had been changed by adjusting the data. Examination of the mean variance for each type of data (Table 7) reveals that only the linear adjustment increased variance while other adjustments did not significantly effect the raw mean variance. As may be noted, similar changes occurred over all teachers on interitem variance. On this basis, the linearly transformed data was not used in any further analysis.

TABLE 7

MEAN OF INTERITEM VARIANCES

	RAW DATA	LINEAR	ADJUSTMENT		MEANS
			PROPOR- TIONAL	Z TRANS- FORMATION	
Teacher 1	3.27	5.92	3.25	3.25	3.92
Teacher 2	2.75	5.60	2.66	2.69	3.42
Teacher 3	2.90	6.24	2.80	2.80	3.68
Means	2.97	5.92	2.90	2.92	3.68

Neither Hypothesis I nor II may be rejected in the null form. Because initial reliabilities of the raw data were high, there was little room for improvement through adjustment. Failing to reject the null hypotheses, further analysis was performed on the raw scores. However, since there seemed to be an indication that the Z transformed scores yielded a slightly

improved reliability, did not distort across item variance, and did not reduce interteacher variance, these data were also selected to be analyzed.

INTERTAPE VARIANCE

Differences in teaching abilities are essential if one is to generalize beyond the experimental situation. During the selection of the teachers to be videotaped, care was taken to assure that there would be differences in teaching. The following hypothesis was established to be tested both before and after adjustment of the data:

HYPOTHESIS III: There will be no difference among ratings of different teachers.

$$H_0: \gamma_1 = \gamma_2 = \gamma_3$$

Hypothesis III was rejected in the null form on the basis of differences found in raw data as well as in adjusted data. Table 8, indicates that significant differences exist among all of the teachers with Teacher 3 reflecting the highest mean rating, Teacher 1, the middle mean rating and Teacher 2, the lowest mean rating.

REMAINING VARIABLES

To test for the significance of the remaining variables, a four-way analysis of variance was calculated. Subsequent paragraphs refer to this ANOVA which is shown in Table 9.

TABLE 8
MEANS OF RATINGS OF EACH TEACHER

	RAW SCORES FOR TEACHER			Z TRANSFORMED SCORES FOR TEACHER		
	1	2	3	1	2	3
Means	5.43	4.21	6.85	5.43	4.21	6.86
Standard Deviation	1.37	1.07	1.46	1.98	1.85	1.92
Newman-Kuels "q ="	7.67**	16.60**		7.77**	16.88**	
		8.93**			9.04**	

**p<.01

Significant F-values were found for Teachers, Abstraction level, and interaction between Teachers and Abstraction level. All other F-values were non-significant.

Willingness to Confront.

Levels. The semantic differential score for willingness to confront was calculated for each rater and raters were classified into a high or low willingness level. Classification was accomplished by using the median score, 41.5, as the dividing point, i.e., those scoring 41 or below were classified as "Low" willingness to confront and those scoring 42 or above as "High" willingness to confront. The distribution of Ss on this variable is shown in Table 10. The means of the two groups were significantly different ($t_{(46)} = 9.00$) at $p < .001$. The variances were not significantly different when tested with an

TABLE 9
ANALYSIS OF VARIANCE FOR ADJUSTED AND UNADJUSTED SCORES

SOURCE OF VARIATION	Z TRANSFORMED SCORES			RAW SCORES	
	DF	MEAN SQUARE	F	MEAN SQUARE	F
Willingness to Confront(A)	1	196.3982	N.S.	121.3912	N.S.
Purpose of Rating (B)	2	197.2320	N.S.	237.7940	N.S.
Teachers (C)	2	9084.8408	64.08***	9057.5648	62.83***
Abstraction Level (D)	2	33.7393	3.30*	35.7037	3.37*
A X B	2	167.2464	N.S.	116.3218	N.S.
A X C	2	28.8463	N.S.	39.5926	N.S.
B X C	4	144.1041	N.S.	138.3704	N.S.
A X D	2	23.3189	N.S.	13.5093	N.S.
B X D	4	5.4524	N.S.	4.9051	N.S.
C X D	4	24.7992	3.61*	21.3738	3.09*
Subjects:A X B	42	269.0253		287.4163	
A X B X C	4	152.8659	N.S.	155.6690	N.S.
A X B X D	4	4.4576	N.S.	4.9398	N.S.
A X C X D	4	4.4347	N.S.	5.2419	N.S.
B X C X D	8	3.2938	N.S.	1.9502	N.S.
Subjects X C:A X B	84	141.7710		144.1505	
Subjects X D:A X B	84	10.2170		10.6028	
A X B X C X D	8	8.8156	N.S.	11.5995	N.S.
Subjects X C X D:A X B	168	6.8703		6.9248	
TOTAL	431				

***p<.001

*p<.05

F ratio, even though there appeared to be a larger variance in the lower group than in the higher group.

TABLE 10
DESCRIPTION OF RATER SEMANTIC DIFFERENTIAL SCORES
ON WILLINGNESS TO CONFRONT

	TOTAL GROUP	LOW GROUP	HIGH GROUP
Means	40.77	35.63	45.92
Standard Deviations	6.62	4.98	3.09
Variances	43.84	24.77	9.55
Number	48	24	24
Significance		$t(46) = 9.00***$	

*** $p < .001$

Hypothesis. The following hypothesis in the null form was to be tested concerning the classification of the raters into two levels of willingness to confront:

HYPOTHESIS IV: There will be no difference between scores of raters classified as low willingness to confront and scores of raters classified as high willingness to confront.

$$H_0: \alpha_1 = \alpha_2$$

As may be seen from Table 9, the F-value was not significant for main effects of willingness to confront. Mean ratings are 32.34 for low willingness and 33.69 for high willingness for Z transformed data and 32.45 and 33.51, respectively for

for raw data. The null hypothesis cannot be rejected.

Purpose of Rating.

Each rater was assigned to a purpose of rating randomly within his willingness category. Purposes were classified into three levels on the basis of instructions provided (see appendix B). The three levels, representing levels of threat were "Research only" (level 1), "Administrative" (level 2), and "Diagnostic" (level 3). The null hypothesis to be tested for this variable follows:

HYPOTHESIS V: There will be no difference in rating scores among the three levels of purpose of rating.

$$H_0: \beta_1 = \beta_2 = \beta_3$$

This null hypothesis cannot be rejected on the basis of lack of significant main effects as shown on Table 9. Means for purpose of rating are shown in Table 11, below:

TABLE 11
MEANS OF PURPOSE OF RATING

	PURPOSE 1	PURPOSE 2	PURPOSE 3
Z Transformed Scores	33.85	31.68	33.52
Raw Scores	34.03	31.55	33.38

Abstraction Levels.

Items were randomly ordered on the rating scale. Appendix C contains a copy of the rating scale utilized and identifies

the level of abstraction of each item as determined by scaling procedures described in chapter 3. All references herein to items reflects the item order after sorting by abstraction level, i.e., items 1-6 are low abstraction level, items 7-12, medium abstraction level, and items 13-18, high abstraction level. The hypothesis to be tested is stated as follows in the null form:

HYPOTHESIS VI: There will be no difference in rating scores among the three levels of abstraction.

$$H_0: \delta_1 = \delta_2 = \delta_3$$

Table 9, above, reflects that there were main effects on this variable significant at $p < .05$ level. Using the Newman-Kuels procedure, a multiple comparison of ordered means indicated that level 1 (lowest abstraction level) was significantly below level 3 (highest abstraction level) which was the next highest rating. The highest ratings were in abstraction level 2, medium level of abstraction level, but these were not significantly above level 3. (See Table 12, below).

TABLE 12
COMPARISON OF ABSTRACTION LEVEL MEANS

	LEVEL 1	LEVEL 2	LEVEL 3
Z Transformed Scores	32.46	33.35	33.24
Raw Scores	32.41	33.30	33.24

Null hypothesis VI is therefore rejected in favor of the alternate hypothesis that lower levels of abstraction yield lower ratings.

Interactions.

Willingness to Confront. Seven interaction hypotheses were postulated concerning willingness to confront. These hypotheses are stated in null form below:

HYPOTHESIS VII: There will be no interaction of levels of willingness to confront and levels of purpose of rating.

$$H_0: \alpha\beta_{11} = \alpha\beta_{12} = \alpha\beta_{13} = \alpha\beta_{21} = \alpha\beta_{22} = \alpha\beta_{23}$$

HYPOTHESIS VIII: There will be no interaction of levels of willingness to confront and levels of teachers.

$$H_0: \alpha\gamma_{11} = \alpha\gamma_{12} = \alpha\gamma_{13} = \alpha\gamma_{21} = \alpha\gamma_{22} = \alpha\gamma_{23}$$

HYPOTHESIS IX: There will be no interaction of levels of willingness to confront and levels of abstraction.

$$H_0: \alpha\delta_{11} = \alpha\delta_{12} = \alpha\delta_{13} = \alpha\delta_{21} = \alpha\delta_{22} = \alpha\delta_{23}$$

HYPOTHESIS X: There will be no interaction of levels of willingness to confront, levels of purpose of rating and levels of teachers.

$$H_0: \alpha\beta\gamma_{111} = \alpha\beta\gamma_{112} = \alpha\beta\gamma_{113} = \alpha\beta\gamma_{121} = \alpha\beta\gamma_{122} = \alpha\beta\gamma_{123} = \alpha\beta\gamma_{131} = \alpha\beta\gamma_{132} = \alpha\beta\gamma_{133} = \alpha\beta\gamma_{211} =$$

$$\alpha\beta\gamma_{212} = \alpha\beta\gamma_{213} = \alpha\beta\gamma_{221} = \alpha\beta\gamma_{222} = \alpha\beta\gamma_{223} =$$

$$\alpha\beta\gamma_{231} = \alpha\beta\gamma_{232} = \alpha\beta\gamma_{233}$$

HYPOTHESIS XI: There will be no interaction of levels of willingness to confront, levels of purpose of rating and levels of abstraction.

$$H_0: \alpha\beta\delta_{111} = \alpha\beta\delta_{112} = \alpha\beta\delta_{113} = \alpha\beta\delta_{121} = \alpha\beta\delta_{122} =$$

$$\alpha\beta\delta_{123} = \alpha\beta\delta_{131} = \alpha\beta\delta_{132} = \alpha\beta\delta_{133} = \alpha\beta\delta_{211} =$$

$$\alpha\beta\delta_{212} = \alpha\beta\delta_{213} = \alpha\beta\delta_{221} = \alpha\beta\delta_{222} = \alpha\beta\delta_{223} =$$

$$\alpha\beta\delta_{231} = \alpha\beta\delta_{232} = \alpha\beta\delta_{233}$$

HYPOTHESIS XII: There will be no interaction of levels of willingness to confront, levels of teachers and levels of abstraction.

$$H_0: \alpha\gamma\delta_{111} = \alpha\gamma\delta_{112} = \alpha\gamma\delta_{113} = \alpha\gamma\delta_{121} = \alpha\gamma\delta_{122} =$$

$$\alpha\gamma\delta_{123} = \alpha\gamma\delta_{131} = \alpha\gamma\delta_{132} = \alpha\gamma\delta_{133} = \alpha\gamma\delta_{211} =$$

$$\alpha\gamma\delta_{212} = \alpha\gamma\delta_{213} = \alpha\gamma\delta_{221} = \alpha\gamma\delta_{222} = \alpha\gamma\delta_{223} =$$

$$\alpha\gamma\delta_{231} = \alpha\gamma\delta_{232} = \alpha\gamma\delta_{233}$$

HYPOTHESIS XIII: There will be no interaction of levels of willingness to confront, levels of purpose of rating, levels of abstraction, and levels of teachers.

$$H_0: \alpha\beta\gamma\delta_{1111} = \alpha\beta\gamma\delta_{1112} = \alpha\beta\gamma\delta_{1113} = \alpha\beta\gamma\delta_{1121} =$$

$$\alpha\beta\gamma\delta_{1122} = \alpha\beta\gamma\delta_{1123} = \alpha\beta\gamma\delta_{1131} = \alpha\beta\gamma\delta_{1132} =$$

$$\alpha\beta\gamma\delta_{1133} = \alpha\beta\gamma\delta_{1211} = \alpha\beta\gamma\delta_{1212} = \alpha\beta\gamma\delta_{1213} =$$

$\alpha\beta\gamma\delta$	$=\alpha\beta\gamma\delta$	$=\alpha\beta\gamma\delta$	$=\alpha\beta\gamma\delta$	$=$
1221	1222	1223	1231	
$\alpha\beta\gamma\delta$	$=\alpha\beta\gamma\delta$	$=\alpha\beta\gamma\delta$	$=\alpha\beta\gamma\delta$	$=$
1232	1233	1311	1312	
$\alpha\beta\gamma\delta$	$=\alpha\beta\gamma\delta$	$=\alpha\beta\gamma\delta$	$=\alpha\beta\gamma\delta$	$=$
1313	1321	1322	1323	
$\alpha\beta\gamma\delta$	$=\alpha\beta\gamma\delta$	$=\alpha\beta\gamma\delta$	$=\alpha\beta\gamma\delta$	$=$
1331	1332	1333	2111	
$\alpha\beta\gamma\delta$	$=\alpha\beta\gamma\delta$	$=\alpha\beta\gamma\delta$	$=\alpha\beta\gamma\delta$	$=$
2112	2113	2121	2122	
$\alpha\beta\gamma\delta$	$=\alpha\beta\gamma\delta$	$=\alpha\beta\gamma\delta$	$=\alpha\beta\gamma\delta$	$=$
2123	2131	2132	2133	
$\alpha\beta\gamma\delta$	$=\alpha\beta\gamma\delta$	$=\alpha\beta\gamma\delta$	$=\alpha\beta\gamma\delta$	$=$
2211	2212	2213	2221	
$\alpha\beta\gamma\delta$	$=\alpha\beta\gamma\delta$	$=\alpha\beta\gamma\delta$	$=\alpha\beta\gamma\delta$	$=$
2222	2223	2231	2232	
$\alpha\beta\gamma\delta$	$=\alpha\beta\gamma\delta$	$=\alpha\beta\gamma\delta$	$=\alpha\beta\gamma\delta$	$=$
2233	2311	2312	2313	
$\alpha\beta\gamma\delta$	$=\alpha\beta\gamma\delta$	$=\alpha\beta\gamma\delta$	$=\alpha\beta\gamma\delta$	$=$
2321	2322	2323	2331	
$\alpha\beta\gamma\delta$	$=\alpha\beta\gamma\delta$			
2332	2333			

Table 9 reveals that there were no significant interactions concerning willingness to confront with any of the other variables. Table 13, below, shows the means of all interactions involved in the above hypotheses. All interaction hypotheses concerning willingness to confront fail to be rejected in the null form. Willingness to confront has no significant effects in this study.

Purpose of rating. In addition to interaction with willingness to confront, purpose of rating involves three additional interaction hypotheses. These three hypotheses are listed below in the null form:

HYPOTHESIS XIV: There will be no interaction of levels of purpose of rating and levels of teachers.

TABLE 13
 CELL MEANS FOR INTERACTION OF WILLINGNESS TO CONFRONT
 WITH OTHER VARIABLES
 (Z TRANSFORMED SCORES)

ABSTR- ACTION LEVEL	TEACH- ER LEVEL	WILLINGNESS TO CONFRONT						WILLINGNESS TO CONFRONT	
		LEVEL 1			LEVEL 2			LEVEL LEVEL	
		PURPOSE 1	LEVEL 2	LEVEL 3	PURPOSE 1	LEVEL 2	LEVEL 3	1	1
1	1	33.20	29.20	29.18	34.36	32.45	36.41	30.53	34.41
	2	23.52	19.10	25.98	23.54	24.09	26.71	22.87	24.78
	3	40.36	44.17	37.22	44.35	37.90	42.54	40.58	41.60
2	1	33.75	31.75	31.25	35.08	30.41	34.68	32.25	33.39
	2	24.82	24.32	28.04	25.35	26.01	28.82	25.73	26.73
	3	41.36	44.07	37.31	43.15	37.55	42.54	40.91	41.08
3	1	34.10	30.88	29.55	34.56	30.56	35.38	31.51	33.50
	2	25.00	22.57	29.22	26.04	23.70	28.41	25.60	26.05
	3	42.16	42.85	38.31	44.65	38.65	41.76	41.11	41.68
ALL	1	33.68	30.61	29.99	34.67	31.14	35.49	31.43	33.76
	2	24.44	22.00	27.74	24.98	24.60	27.98	24.73	25.85
	3	41.29	43.69	37.61	44.05	38.03	42.28	40.87	41.45
1		32.36	30.82	30.80	34.09	31.48	35.22	31.33	33.59
2	ALL	33.31	33.38	32.20	34.53	31.32	35.35	32.96	33.73
3		33.75	32.10	32.36	35.08	30.97	35.18	32.74	33.74
ALL	ALL	33.14	32.10	31.78	34.57	31.26	35.25		

$$H_0: \beta\gamma_{11} = \beta\gamma_{12} = \beta\gamma_{13} = \beta\gamma_{21} = \beta\gamma_{22} = \beta\gamma_{23} = \\ \beta\gamma_{31} = \beta\gamma_{32} = \beta\gamma_{33}$$

HYPOTHESIS XV: There will be no interaction of levels of purpose of rating and levels of abstraction.

$$H_0: \beta\delta_{11} = \beta\delta_{12} = \beta\delta_{13} = \beta\delta_{21} = \beta\delta_{22} = \beta\delta_{23} = \\ \beta\delta_{31} = \beta\delta_{32} = \beta\delta_{33}$$

HYPOTHESIS XVI: There will be no interaction of levels of purpose of rating, levels of teachers, and levels of abstraction.

$$H_0: \beta\gamma\delta_{111} = \beta\gamma\delta_{112} = \beta\gamma\delta_{113} = \beta\gamma\delta_{121} = \beta\gamma\delta_{122} = \\ \beta\gamma\delta_{123} = \beta\gamma\delta_{131} = \beta\gamma\delta_{132} = \beta\gamma\delta_{133} = \beta\gamma\delta_{211} = \\ \beta\gamma\delta_{212} = \beta\gamma\delta_{213} = \beta\gamma\delta_{221} = \beta\gamma\delta_{222} = \beta\gamma\delta_{223} = \\ \beta\gamma\delta_{231} = \beta\gamma\delta_{232} = \beta\gamma\delta_{233} = \beta\gamma\delta_{311} = \beta\gamma\delta_{312} = \\ \beta\gamma\delta_{313} = \beta\gamma\delta_{321} = \beta\gamma\delta_{322} = \beta\gamma\delta_{323} = \beta\gamma\delta_{331} = \\ \beta\gamma\delta_{332} = \beta\gamma\delta_{333}$$

Because of the lack of significance as shown on Table 9, above, none of the null hypotheses concerning purpose of rating can be rejected. The means for purpose of rating interacting with teachers and with abstraction levels are shown in Table 14, below. In spite of the seemingly substantial differences in cell means by teacher and purpose, statistically they are not significant. The means range from teacher 2, purpose 2, across nearly 20 points to teacher 3, purpose 1. Even though the difference is non-significant, the tendency indicates the more threatening the purpose the more stringent the rating.

This is the same tendency reflected in main effects for purpose of rating and is further supported in the interactions with abstraction level, Table 14.

TABLE 14
CELL MEANS FOR INTERACTION OF PURPOSE OF RATING
WITH TEACHERS AND ABSTRACTION LEVEL

ABSTR- ACTION LEVEL	TEACH- ER LEVEL	PURPOSE			PURPOSE		
		1	2	3	1	2	3
1	1	33.78	30.82	32.80			
	2	23.53	21.60	26.35	33.22	31.15	33.01
	3	42.36	41.03	39.88			
2	1	34.41	31.08	32.97			
	2	25.09	25.17	28.43	33.92	32.35	33.77
	3	42.25	40.81	39.93			
3	1	34.33	30.72	32.47			
	2	25.52	23.14	28.81	34.42	31.53	33.77
	3	43.40	40.75	40.83			
ALL	1	34.17	30.87	32.74			
	2	24.71	23.30	27.86			
	3	42.67	40.86	39.95			

Teacher and Abstraction Level. Hypothesis XVI, stated below in the null form represents the remaining interaction hypothesis.

HYPOTHESIS XVI: There will be no interaction of levels of teachers and levels of abstraction.

$$H_0: \gamma_{11}^{\delta} = \gamma_{12}^{\delta} = \gamma_{13}^{\delta} = \gamma_{21}^{\delta} = \gamma_{22}^{\delta} = \gamma_{23}^{\delta} = \\ \gamma_{31}^{\delta} = \gamma_{32}^{\delta} = \gamma_{33}^{\delta}$$

Table 9, reflects an F-value significant $p < .05$ for this interaction in both Z transformed scores and unadjusted scores. Means for both sets of data are shown for this interaction in Table 15, below. Multiple comparison, using the Newman-Kuels modified q-statistic, of the interactions between abstraction level and teachers, reveals that only within teacher 2, the lowest rated teacher, does abstraction level retain the significance found in the main effects. All other significant differences range across teachers. Across abstraction levels, within teachers 1 and 3, there are no significant differences.

TABLE 15
MEANS OF TEACHERS AND ABSTRACTION LEVELS
ABSTRACTION LEVEL

TEACHER	1		2		3		MEAN	
	Z	RAW	Z	RAW	Z	RAW	Z	RAW
1	32.47	32.48	32.82	32.79	32.50	32.50	32.60	32.59
2	23.82	23.83	26.23	26.10	25.82	25.83	25.29	25.26
3	41.09	40.92	41.00	41.00	41.39	41.40	41.16	41.10
MEAN	32.46	32.41	33.35	33.30	33.24	33.24	--	--

SUMMARY

Small, non-significant differences in reliabilities were found between raw, unadjusted scores and scores adjusted by Z-score transformation of rating data. Though non-significant, the changes were in the direction of reduction of variance across raters. Further, the adjustment affected neither inter-item variance nor differences across teachers, indicating that discriminatory ability of the data was retained.

Significant differences were found across abstraction level scores. The cell means for abstraction levels were tested using the Newman-Kuels multiple comparison test. This test requires that the means be ordered from lowest to highest. As a result, abstraction level 1 was compared with the next higher mean, abstraction level 3. These means were significantly different ($p < .05$). When comparing level 3 with the highest mean, level 2, there was no significant difference. Comparison of level 1 with level 2, lowest with highest, failed to yield significance because the test reduces the degrees of freedom. It appears that the more abstract the item, the more inflationary bias there is present.

Ratings across teachers was highly significant ($p < .001$) for both raw data and Z transformed data. Interaction between abstraction level and teacher level was significant only for teacher 2, the lowest rated teacher. These findings indicate that abstraction level is most significant when there is reason

(or when one is able) to distinguish differences in teaching abilities.

Willingness to confront and purpose of rating failed to yield significant F values. Possible reasons for this are discussed in the next chapter.

CHAPTER V

SUMMARY AND CONCLUSIONS

This chapter will provide a summary of the procedures and results in order of the hypotheses stated previously. It will also present conclusions in terms of the problem statement as answers to the questions posed within the problem. In conclusion, there will be implications for further study presented, identifying procedures which might provide a more rigorous treatment of the topic.

SUMMARY

Reduction of Variance.

Videotape film clips, 8 minutes in length, of three teachers were viewed by 48 school supervisors and potential supervisors for the purpose of rating the teachers. The rating instrument consisted of 18 items which provided three subscores for different levels of abstraction of rating scale items. The items had previously been scaled by educators in terms of their level of abstraction. Prior to the viewing of the film clips, each rater had completed a semantic differential booklet for a measure of his willingness to confront and for a measure of his concept of ordinary teacher. Each rater also had completed a copy of the rating scale describing his concept of an "Average Teacher". In all, five instruments were completed by each

rater providing six measures: Concept of willingness to confront, concept of ordinary teacher, concept of average teacher, and ratings of each of three teachers videotaped while conducting actual classes.

The six measures were intercorrelated yielding ten intercorrelations of which only the coefficient of correlation between teacher 1 and teacher 2 ratings were significant. Concepts of ordinary teacher and of average teacher did not correlate significantly with each other nor with ratings of any of the teachers.

Reliability coefficients were calculated, using an ANOVA model (Medley and Mitzel, 1964), for the raw ratings of the three teachers and for four adjusted ratings: linear adjustment, proportional adjustment, Z-score transformation adjustment, and item residual adjustment. Coefficients of reliability ranged from 0.9679 with the item residual adjustment to 0.9763 for the Z transformed adjustment. None of the differences were significant indicating that the adjustment to the data did not improve reliability which was the operational definition of reduction of interrater variance. The Z transformed adjustment did result in a slight increase in reliability over the raw score adjustment, however. Null hypotheses I and II concerning reliability of raw scores and adjusted scores could not be rejected based on the results cited above.

All adjustments were then examined to insure that adjusting

the data had not increased across item variance or reduced across teacher differences which would have voided the benefits to be gained with reduction of within item variance. Significant differences in variance occurred only with the linear adjustment. On the basis that the Z transformed scores tended to reduce interrater variance and did not damage the across item variance, further analysis was performed on these data as well as on raw score data.

Interteacher Variance.

Hypothesis III was concerned with maintaining differences across teachers to assure generalizability of results. Differences in mean ratings for the three teachers were significant, causing rejection of the null hypothesis that differences would not exist across teachers. Significant differences were found both before and after adjustment, indicating that three different levels of teaching were actually involved, and adjustment of the data had not removed these differences.

Main Effects.

Hypotheses IV, V, and VI were concerned with main effects of willingness to confront, purpose of rating, and abstraction level of items. The null hypothesis for abstraction level was the only one (Hypothesis VI) which could be rejected based on a significance level of $p < .05$. The remaining hypotheses could not be rejected. The mean for abstraction level 1, lowest

level of abstraction, was significantly below abstraction level 3, highest level of abstraction level, which was lower than the mean of abstraction level 2, though not significantly so.

Interaction Effects.

Remaining hypotheses, VII through XVII, were concerned with testing interaction effects. Only hypothesis XVII, the interaction between teacher levels and abstraction levels, could be rejected in the null form. All other interaction effects were nonsignificant. Multiple comparison tests identified that only means for teacher 2, the lowest rated teacher, interacted with abstraction level differences. Abstraction level effects were not significant within teachers 1 and 3.

CONCLUSIONS

The Problem.

Are there effects of knowledge of output use of ratings on raters? Are there effects of varied levels of abstraction of rating items on rater output? Can knowledge of rater concept of average be used to reduce interrater variance? Can knowledge of rater differences in dispersion (use of extreme rating categories) be used to reduce interrater variance? Does rater willingness to confront interact with rating output use? Answers to these questions are offered as found in this study in the following paragraphs.

Are there effects of knowledge of output use of ratings on raters?

Within this study, knowledge of output use was termed purpose of rating and was operationally defined by specific instructions provided each rater. (See Appendix A). In order to control for other variables within the administration of the experiment, each rater was provided, randomly, a sheet which specified the purpose for which he was to rate the teacher. Questioning of several of the raters following the rating procedure indicated that several were unaware of their purpose of rating.

Because the procedure was conducted with graduate students, the reality of the purposes may not have been realized. It is suggested that many of those who had more threatening purposes may not have participated in their assigned role (purpose) in sufficient earnest to reflect differences that might appear in situations with greater evidence of reality.

Therefore, whether there are effects of knowledge of output use of ratings on raters cannot be adequately addressed within this study. It may only be stated that no effects were found.

Are there effects of varied levels of abstraction of rating items on rater output?

Levels of abstraction of rating items had been established through an extensive scaling process. All analysis was per-

formed using subscores for each level of abstraction. It is apparent that within this study there were effects on the rater output of abstraction level. Not only were there main effects reflecting lower ratings for the least abstract items, but the interactive effects with teacher 2, the lowest rated teacher, would confirm that constant errors, as defined by Guilford, and inflation of ratings as described by Johnson, can be reduced if rating items are more concrete. The highest mean ratings, which are likely to be the ratings with the greatest error, are those at the mid point of a scale of abstraction. These mid-abstraction items are undoubtedly the most commonly used item in teacher rating scales.

Can knowledge of rater concept of average be used to reduce interrater variance?

Rating scores were adjusted for raters' concept of average as measured by having each rater rate his "picture" of the average teacher. Only very minor differences occurred in reliability between unadjusted rating scores and rating scores adjusted for average. None of the differences in reliability were significant.

In addition to the raters' concept of average measured by rating an average teacher on the rating form, a measure was taken using a semantic differential with the concept "Ordinary Teacher". These two measures were not significantly related

to each other or to any of the teacher ratings.

More accurate measures or more powerful adjustment techniques may be required to tap the concept of average element which may be a bias in the ratings. Perhaps there also is a need to adjust for an interval distance from the mean, rather than for the mean. A third alternative is that the concept of average might not be the critical variable but the value of the rater being above or below the average of all raters may be the key variable.

Can knowledge of rater differences in concept of average and of dispersion (use of extreme rating categories) be used to reduce interrater variance?

Three adjustments were made in an attempt to reduce interrater variance through knowledge of rater differences in concept of average and of dispersion. One of these methods yielded a very slight increase in reliability (nonsignificant) while others either decreased the reliability (nonsignificantly) or held it constant. Using a Z-score transformation with the average teacher rating mean and standard deviation covaried out of each teacher's rating, increased reliability 0.0013 over the raw unadjusted data. While nonsignificant, this increase tended to reflect a desired improvement in reliability. A linear adjustment, subtraction of each raters' "Average Teacher" item score from his rating score for each teacher, yielded the identical reliability as raw scores but caused a

substantial increase in interitem variance. From the trend reflected in the Z-score transformation, it was hopeful that item residual covariance might provide even greater reliability. However, reliability was slightly reduced with this adjustment.

Knowledge of rater differences in concept of average and of dispersion shows a glimmer of promise as a method for reducing interrater variance. Whether other adjustments or other measures would provide greater increased reliability is still open to question. One possible method might be to ask each rater to complete two other rating forms, the poorest teacher they have ever seen and the best teacher they have ever seen, to better anchor their average and dispersion points. The suggestion made earlier concerning categorizing raters as above or below the group mean might be extended to categorizing them as central or extreme raters for a more fruitful control.

Does rater willingness to confront interact with rating output use?

Willingness to confront was measured by a factor obtained from data on the semantic differential which each rater completed prior to his rating of the teachers. The lack of significant interaction with purpose of rating indicates that rater willingness to confront has no significant relationship to rating output use.

In addition to the explanation offered under differences in rating output use, above, it is suggested that the method

of measuring willingness to confront may not have been sufficiently refined to identify any interaction which might be present. Willingness to confront was used as a dichotomous variable even though it was measured as continuous. By arbitrarily dividing the raters into high and low willingness on the basis of median score, there may have been insufficient distance between the two groups to adequately distinguish between those in the high willingness group and those in the low. A more discriminating measure of this variable or use of the variable as a covariant might prove more fruitful.

PROCEDURES FOR STUDYING THE RATING PROCESS

While findings in this study are meager concerning the rating process, experience has been rich in terms of procedures for studying such a process. Adequate attention to elements within the three categories below would increase our knowledge of the process of rating.

Realism. Like many of the variables we would wish to study in education, transfer from laboratory to classroom is difficult. Study within the classroom often involves many uncontrollable variables, however, only if the rating situation has practical and realistic implications to the raters will responses be generalizable and genuine. Performing such a study in the classroom would increase the rater involvement and increase the likelihood that he takes his role in earnest.

Limited variables. Confounding effects may be more easily

"research only" and "diagnostic" in a live situation would be supported by the experience within this study. Alternatively, different categories might be designed representing varied levels of threat which would be appropriate to a particular rating situation such as rating of student teachers on either videotapes or in the classroom. Essentially, the situation must be real and there must indeed be a felt threat on the part of the rater in one or more instances and a lack of threat in one or more other instances.

Sample size. While there can be no rule of thumb, it should be obvious that one has a greater probability of gaining significance if a large sample is used than if a small sample is used. In replicating this study, it would be most logical to increase sample size to at least 15 per cell. For maximum generalizability and opportunity for significance, 30 per cell would be desirable. Locating 180 willing raters as a sample of a population of administrators would be horrendous but with slight revision of variables, student ratings might be used and larger samples could readily be obtained.

Variables. It is not always easy to reduce the number of variables within a study but better controls and better experimenter-subject rapport would result from a study with fewer variables. Also, the measurement of a continuous variable such as willingness to confront and attempting to dichotomize it, is unwise. If this variable cannot be measured in a dichotomous

manner, it undoubtedly should be either disregarded or employed as a covariable. The gain in measurement as a covariable, however, might be lost in the potential significance of this variable interacting with purpose of rating.

Instrumentation. Reliability of the rating scale was sufficiently high that there was little room for demonstrating an increase through adjustment procedures. Reduction in reliability of the rating might automatically result from using a larger sample of teacher behavior or it might be necessary to test rating items to select items on which raters tend to disperse their ratings to a greater extent. If raters increased their within item variance, more potential would be available for reducing that variance experimentally. Typically reliability of ratings tend to range in the vicinity of .60 to .70. Such ratings would be expected to have a much greater probability of being increased than ratings ranging from .96 upward.

IMPLICATIONS FOR FURTHER STUDY

Further Study Profitable. Results of this study provide considerable evidence that further study could be highly fruitful. Not only are the findings concerning abstraction level important for rating scale development, but general tendencies concerning rater's concept of average and of dispersion lead to the conclusion that the rating process can be refined. Several suggestions are offered in the following paragraphs for further refinement of methods for identification and removal

of elements of bias in ratings.

Improvement of Present Study. From the lessons learned in this study, principle elements which can be refined for better controls are offered below:

Concept of Average and Dispersion. Three rating forms should be completed by each rater prior to rating teachers. These forms should represent the poorest teacher, the best teacher and the average teacher as pictured by each rater. Such measures would yield more precise adjustments to reduce interrater differences than those used in this study. Further, these measures should be examined as possible categorizing variables to identify "High", "Low", "Central", and "Extreme" raters. Such categories could be used to "block" the raters for control rather than adjust for increased reliability.

Willingness to Confront. Validity of this variable as measured is subject to question in this study. Also, measurement on a continuous scale creates definite problems when attempting to dichotomize. It may be possible to measure this variable using a situational test, rather than an attitudinal scale, with more accuracy. If it is a normally distributed variable, it may be in order to use it as a covariable, rather than a blocking variable.

Purpose of rating. Purposes of rating must be believable and contain the threat of a real situation to assure that possible differences are accounted for. The rating should be

performed in an actual classroom setting where the rater can see the teacher during and after the rating. Collapsing levels to only two purposes may be in order since the administrative purpose and diagnostic purpose interchanged order across teachers and across abstraction levels.

Abstraction Level. The range of abstraction level for the items used was relatively small. Out of a possible range of five, scaled items ranged from 0.83 to 2.30, or a range of approximately one and one-half. Most rating scale items are of this small range and tend toward the middle to upper level of abstraction. Construction of some highly concrete rating scale items to be used with the present more abstract ones might yield even greater significance. Further refinement might include construction of parallel items of different levels of abstraction.

Further Study of the Rating Process. The urgency for more precision in teacher rating demands that rater processes be purified. The fact that two raters, observing the same teacher behavior vary in their qualitative judgment of that behavior, bespeaks a difference in the internal processes involving judgment. There is a need to examine additional variables in an attempt to identify their contribution to rating bias. Some of these additional variables which might contribute to bias include:

Rater basis or criteria for rating a given subject on a

given trait. Assessment of each rater's criteria for rating of each trait might serve to identify some of the characteristics which differentiate among raters. Attention to the rating situation might also be enhanced and could influence reliability.

Confidence level of ratings. Confidence weighting of test responses has been demonstrated to increase reliability. Would confidence weighting procedures reduce interrater variance? Would raters' willingness to confront by accounted for by confidence weighting.

Influence of not observed traits. When traits are not observed, raters tend to either ignore that item or to rate it on the basis of an inference drawn from those traits that were observed. Given no instructions concerning unobserved traits, can raters be predicted to respond with an inferential judgment or ignore an item? How does this difference in response pattern effect the rating data?

Rater knowledge of ratee and subject-matter. There is evidence that raters overestimate traits of those ratees whom they know. Can this overestimation be quantified and removed from ratings? If the observer has knowledge of the subject-matter being taught, does his ratings differ from an unsophisticated observer?

Implications for Teacher Rating. Studies of the process of teaching suffer from a serious lack of criteria for "Effective Teaching". Coupled with pupil achievement, teacher rating

could provide an acceptable measure of "Effective Teaching". Unfortunately, the rating process must be refined before one can utilize rating data for such research. Identification of certain elements of bias in the rating process and method for removing this bias are possible. Concrete rating items contain less bias than more abstract ones. Knowledge of rater characteristics may enhance rating data through adjustment of the data to remove differences in average and dispersion rating strategies.

Administrators should be able to use these techniques for removing bias to improve rating of their teachers. Hard personnel decisions concerning school faculties are being forced. These decisions cannot be based on the traditional "standardized test" results but must be supported with concrete judgments.

Schools of education continue to send out supervisors for their student teacher programs armed with rating scales. How accurate, reliable, and valid these scales is still subject to question. Removal of elements of bias can make the judgments of effective student teaching more objective and useful, not only for diagnosing weaknesses and strengths of the students but for improving student teacher programs and screening potential teachers.

REFERENCES

- Barr, Avril S. Wisconsin Studies of the Measurement and Prediction of Teacher Effectiveness. Madison, Wisconsin: Dembar Publications, Inc., 1961.
- Brinton, James E. Deriving an Attitude Scale from Semantic Differential Data. In Semantic Differential Technique, James G. Snider and C. E. Osgood (editors). Chicago: Aldine Publishing Co., 1969.
- Christal, Raymond E. Selecting a Harem--And Other Applications of the "Policy Capturing Model." Journal of Experimental Education. 36:35-41, 1968.
- Dayton, C. Mitchell. Design of Educational Experiments. New York: McGraw-Hill, 1970. p. 283.
- Gage, N. L. Paradigms for Research on Teaching. Handbook of Research on Teaching, N. L. Gage (editor). Chicago: Rand McNally, 1963.
- Good, Carter V. (editor) Dictionary of Education. 2nd ed. New York: McGraw-Hill, 1959.
- Guilford, J. P. Psychometric Methods. New York: McGraw-Hill, 1954.
- Hayakawa, S. I. Language in Thought and Action. New York: Harcourt, Brace and Co., 1949.
- Houston, Samuel R. and John T. Roscoe. The Use of the Judgment Analysis Technique in Predicting Success in Graduate Education. California Journal of Educational Research. 20:162-169, 1969.
- Johnson, Donald M. The Psychology of Thought and Judgment. New York: Harper and Brothers, 1955.
- Johnson Abercrombie, M. L. The Anatomy of Judgment. New York: Basic Books, Inc., 1960.
- Kerlinger, Fred N. Foundations of Behavioral Research. New York: Holt, Rinehard and Winston, 1966.
- McKeachie, W. J., Yi Guang Lin, and William Mann. Student Ratings of Teacher Effectiveness: Validity Studies. American Educational Research Journal. 8:435-445, 1971.

- Medley, Donald M. and Harold E. Mitzel. Measuring Classroom Behavior by Systematic Observation. Handbook of Research on Teaching, N. L. Gage (editor). Chicago: Rand McNally, 1963. pp. 247-328.
- Osgood, Charles E., George J. Suci, and Percy H. Tannenbaum. The Measurement of Meaning. Urbana, Illinois: University of Illinois Press, 1957.
- Remmers, H. H. Rating Methods in Research on Teaching. Handbook of Research on Teaching, N. L. Gage (editor). Chicago: Rand McNally, 1963. pp. 329-378.
- Rosenshine, Barak. Evaluation of Classroom Instruction. Review of Educational Research. 40:279-300, 1970.
- Rosenshine, Barak. To Explain: A Review of Research. Educational Leadership. 26, No. 3:303-309, 1968.
- Rosenshine, Barak and Norma Furst. Current and Future Research on Teacher Performance Criteria. Research on Teacher Education: A Symposium, B. O. Smith (editor). Englewood Cliffs, N. J.: Prentice-Hall, 1971.
- Stahlnaker, J. M. and Remmers, H. H. Can Students Discriminate Traits Associated with Success in Teaching. Journal of Applied Psychology. 12:602-610, 1928.
- Tannenbaum, Percy H. Initial Attitude Toward Source and Concept as Factors in Attitude Change Through Communication. Public Opinion Quarterly. 20:413-425, 1965.
- Torgerson, Warren S. Theory and Methods of Scaling. New York: John Wiley & Sons, 1958. pp. 168-179.
- Winer, B. J. Statistical Principles in Experimental Design. New York: McGraw Hill, 1962. p. 80.

(Instructions Continued)

If you consider the concept to be neutral on the scale, both sides of the scale equally associated with the concept, or if the scale is completely irrelevant, unrelated to the concept, then you should place your check-mark in the middle space:

safe : : : X : : : dangerous

IMPORTANT:

- (1) Be sure that you place your check-mark in the middle of spaces, not on the boundaries.
: Th X : : : Not X this : : :
- (2) Be sure to check every scale for every concept--do not omit any.
- (3) Never put more than one check-mark on a single scale.

Sometimes you may feel as though you've had the same item before on the test. This will not be the case, so do not look back and forth through the items. Do not try to remember how you checked similar items earlier. Make each item a separate and independent judgment. Work at fairly high speed through the booklet. Do not worry or puzzle over individual items. It is your first impressions, the immediate "feelings" about the items, that are important. On the other hand, please do not be careless, because we want your true impressions.

ORDINARY TEACHER

(Concept)

good : : : : : bad

weak : : : : : strong

passive : : : : : active

hard : : : : : soft

dirty : : : : : clean

valuable : : : : : worthless

unpleasant : : : : : pleasant

honest : : : : : dishonest

negative : : : : : positive

approach : : : : : avoid

WILLINGNESS FOR CONFRONTATION
(Concept)

good : : : : : bad

weak : : : : : strong

passive : : : : : active

hard : : : : : soft

dirty : : : : : clean

valuable : : : : : worthless

unpleasant : : : : : pleasant

honest : : : : : dishonest

negative : : : : : positive

approach : : : : : avoid

INSTRUCTIONS FOR RATING AVERAGE TEACHER

Picture in your mind an "average" teacher. Picture this teacher in a classroom teaching students. Subject matter and sex of the teacher are irrelevant. Consider that this "average" teacher possesses certain characteristics, and like most of us, he has strong points and weak points. Now, considering this hypothetical average teacher, and visualizing him in a classroom, rate him on the attached rating form. Mark each quality listed on the form for that teacher whom you have pictured as "average".

— 100% —

INSTRUCTIONS

I have been asked by the superintendent and a principal of a school system to develop a rating form which can be used to rate their teachers. Your ratings will be anonymous to all but me. Results of the ratings will be used for research purposes in the development of the rating forms. Please complete the forms by rating the teachers on the video tapes which you will see. Every item must be marked, please.

Instructions provided subjects assigned to Purpose 1, Research

INSTRUCTIONS

I have been asked by the superintendent and a principal of a school system to obtain some objective ratings of some teachers. These ratings are to be used to help the teachers to improve their teaching. Immediately following the rating sessions, the teachers will be brought in for you to discuss why you rated them as you did and what they might do to improve their teaching. Please complete the forms by rating the teachers on the video tapes which you will see. Every item must be marked, please.

Instructions provided subjects assigned to Purpose 3, Diagnostic

INSTRUCTIONS

I have been asked by the superintendent and a principal of a school system to obtain some objective ratings of some teachers. These ratings will be used to help the school administration to make decisions on issuing contracts and establishing salary. Results of the ratings will be provided the administrators but your ratings will remain anonymous to all but me. Please complete the forms by rating the teachers on the video tapes which you will see. Every item must be marked, please.

Instructions provided subjects assigned to Purpose 2, Administrative

TEACHER RATING SCALE

Teacher: _____ Your Name: _____

INSTRUCTIONS: The following lines represent characteristics commonly noted when describing effective or not effective teachers. Please place a check mark (✓) on that part of each line which would indicate how you would rate this teacher on each of the characteristics. Each line must be checked.

- 1. Did the teacher evaluate his courses by keeping in contact with the learners?
 Scale Value 1.07--High Level
 Maintained contact _____ Occasional contact _____ No contact _____
- 2. Pupils actively participate in classroom discussions and activities.
 Scale Value 2.30--Low Level
 Active participation _____ Occasional participation _____ No participation _____
- 3. Teacher employs a variety of approaches in presenting new materials.
 Scale Value 2.16--Low Level
 Uses variety _____ Lacks variety _____
- 4. Does the learner have a feeling of accomplishment concerning this lesson?
 Scale Value 0.83--High Level
 Yes _____ No _____
- 5. Was the content of this lesson meaningful?
 Scale Value 0.93--High Level
 Completely meaningful _____ Not meaningful _____
- 6. Selection of content. Content is appropriate for aims, level, and method.
 Scale Value 1.55--Medium Level
 Appropriate _____ Inappropriate _____
- 7. The pupils' curriculum is enriched through the use of a variety of materials to supplement the basic program.
 Scale Value 2.09--Low Level
 Yes _____ No _____
- 8. Interest in subjects:
 Scale Value 1.55--Medium Level
 Always appears full of subject _____ Mildly interested _____ Subject seems irksome _____

- 9. Selection of Materials, Materials and resources are related to content and complement method.
 Scale Value 2.03--Low Level
 Well selected _____ Poorly selected _____
- 10. Knowledge of subject:
 Scale Value 2.06--Low Level
 Seems highly knowledgeable _____ Preparation adequate _____ Seems poorly prepared _____
- 11. Self-reliance and confidence:
 Scale Value 1.63--Medium Level
 Always sure of himself _____ Fairly self-confident _____ Hesitant, timid, uncertain with poise _____
- 12. Did the teacher listen to and respect ideas different from his own?
 Scale Value 2.10--Low Level
 Always _____ Sometimes _____ Never _____
- 13. Pupils and teacher share the enjoyment of humorous situations.
 Scale Value 1.64--Medium Level
 Yes _____ No _____
- 14. Were the objectives of the course developed in an understandable manner?
 Scale Value 1.55--Medium Level
 Completely understandable _____ Not understandable _____
- 15. Attitude of the teacher:
 Scale Value 1.16--High Level
 Openminded _____ Biased _____
- 16. Personality of the teacher:
 Scale Value 1.04--High Level
 Interesting _____ Poor _____
- 17. Presentation of material:
 Scale Value 1.64--Medium Level
 Makes subject crystal-clear _____ Mechanical, uninspiring _____ Very hard to follow _____
- 18. What is your overall evaluation of the teacher?
 Scale Value 0.87--High Level
 Excellent _____ Poor _____

