1975

Perceived effects and relationships of multi-assessor feedback on modifying educator performance behavior

Luvern Robert Eickhoff
Iowa State University

Recommended Citation

Eickhoff, Luvern Robert, "Perceived effects and relationships of multi-assessor feedback on modifying educator performance behavior " (1975). *Retrospective Theses and Dissertations*. 5471.
https://lib.dr.iastate.edu/rtd/5471

EICKHOFF, Luvern Robert, 1930-
PERCEIVED EFFECTS AND RELATIONSHIPS OF MULTI-ASSESSOR FEEDBACK ON MODIFYING EDUCATOR PERFORMANCE BEHAVIOR.

Iowa State University, Ph.D., 1975
Education, higher

Xerox University Microfilms, Ann Arbor, Michigan 48106

Perceived effects and relationships of multi-assessor feedback on modifying educator performance behavior

by

Luvern Robert Eickhoff

A Dissertation Submitted to the
Graduate Faculty in Partial Fulfillment of
The Requirements for the Degree of
DOCTOR OF PHILOSOPHY

Department: Professional Studies
Major: Education (Research and Evaluation)

Approved:

Signature was redacted for privacy.

In Charge of Major Work

Signature was redacted for privacy.

For the Major Department

Signature was redacted for privacy.

For the Graduate College

Iowa State University
Ames, Iowa
1975

TABLE OF CONTENTS

LIST OF TABLES

Page

Page

Page

CHAPTER I. INTRODUCTION

Formal educator (teacher) assessment of teaching performance has been receiving a great deal of attention and use in higher education. Efforts on the subject have given rise to a multitude of questions regarding its purpose, its possible effects upon the improvement of instruction, the use of certain groups of assessors, and the reliability of assessors in measuring educator performance. The interest and efforts heightened in the middle 1960s and have extended into the 1970s. The increased interest was brought on by student reactions, diminishing student enrollment, and the thrust on accountability. Governmental dialogue gave rise to a concern for financial responsibility which gained widespread popularity in education. With this thrust of interest, educator assessment strategies became varied and numerous. Aleamoni (1, p. 1) spoke of the many proposals in his research memorandum stating that:

> There have been many proposals, especially in the past few years, to evaluate instruction. Most of those proposals contain similar elements or areas of concern such as student, peer, and supervisor ratings. If, however, one looks for actual working models of instructional evaluation, it is immediately apparent that schemes involving systematic ratings by peer, supervisor, self, as well as of material, content, etc., are very seldom actualized. More often than not, the student ratings of instructor and instruction appear as the only elements in any of the "working models."

If the purpose of these assessments is to enable the educators to modify their performance and in turn improve instruction, it is imperative that the educators have access to the results of the assessments. Underlying this intended use is the assumption that the educators will use the information to alter or modify their performance. It is an assumption open to question.

Effectiveness versus Performance

While many groups of people have called for accountability in the educational arena, few people have considered the significance of measuring an educators' effectiveness versus his performance. Many will declare that the educator is responsible for the incremental knowledge gained using some form of "output minus input" measure. Menne (30, pp. 5-6) in 1972 pointed out the fallacy of this declaration:

> . . . if you are concerned with teacher effectiveness . . . you are concerned with the difference--Output minus Input and effectiveness in this sense is both generally quite small and difficult to measure. The reason for this is that most of the output is explained by input. Consider that if final grades from the preceding course in a sequence of courses or from a similar course are correlated with present "OUTPUT" or course grades, the correlation will typically be a least 0.70 and very often in the 0.80 to 0.90 range.

It is to say then, that a correlation of 0.70 reveals that approximately 50 percent of the variance in final grades can be explained by the conditions that occurred prior to a given educator having had the opportunity to influence the behavior of a student. Menne reasons that often only 30 percent of variance can be attributable to such factors as educator effort and the student-teacher interaction. He, then, argues that it might be better to utilize an educator performance strategy because of the ease in measuring performance versus effectiveness.

Menne (30, p. 6) distinguished between effectiveness and performance when he argued that:

> . . . the proportion of variance due to teacher influence is a fairly small proportion--perhaps 20 percent, 10 percent or less. This small proportion of variance leads to the practical impossibility of measuring the difference in effectiveness (OUTPUT-INPUT) between teachers

so that it would be fair and accurate to say that one teacher is better or worse than another.

The measurement of an educator's effectiveness is difficult according to Aleamoni and Spencer (5). When discussing the Illinois Course Evaluation Questionnaire they contended that, "The measurement of the effectiveness of instruction is a complex problem. Generally speaking, many schemes measuring effectiveness is of only one kind, student opinion." Aleamoni suggests that measurement of effectiveness may be approached in various ways.

Given that the measurement of an educator's effectiveness is an extremely difficult task, is it reasonable to accept the theory that the measurement of an educator's performance is more practicable? For example, Menne (30, p. 4) states that:

> It should be noted that there are many factors or aspects to a teacher's performance. If performance is rated as a global construct, it is to be expected that some raters will think of factors such as clarity and stimulation value of material presentation; others will think of the teacher's personality, mode of interaction with students or competence in the content. Thus, it is necessary to be concerned about and delimit the rather specific aspects of performance being evaluated in order to measure something when using raters.
>
> Measures of teacher performance are frequently obtained by using administrators or fellow teachers or students as raters. But no matter who does the ratings, there are three conditions that must be present in order to have evidence that a rating scheme does, in fact, measure something.
>
> a) there must be more than one rater;
> b) the raters must closely agree in their ratings;
> c) the ratings must indicate differences between teachers.

From these three conditions, then, an effective, useful, and successful educator performance assessment strategy would be one whereby a variety of inputs are utilized in the process. The development of such a strategy necessitates the involvement of students, peers, and possibly administrators.

In such a scheme, the question may be asked, does it make any difference who the assessors are, so long as the three conditions are met? The use of such a scheme for higher education is a difficult task since there is a lack of such a model.

Assessment Schemes

While there has been a great deal of lip service to the needs and purposes of educator performance assessment, adequate and acceptable assessment procedures are tenuous at best. This attention has focused primarily on the use of the student's assessment of the educator. Frequently the development of measuring instruments have not involved the educator. Some authors declared that the success or failure of the development and use of measuring instruments depends to a critical degree on the involvement of the educator.

Aleamoni (1, p. 3) conceived of one scheme in the assessment of the educator's performance when he proposed that:

> One possible approach that could be used to begin establishing a total instructional evaluation scheme is to have departmental executive committees and/or chairmen begin asking candidates to suggest names of qualified individuals to evaluate their instruction . . . , etc. A review committee of four could then be selected consisting of three faculty members and one student with at least one member being taken from the candidate's suggested list. This committee would then be charged to conduct a thorough evaluation . . . which would be used along with student ratings in arriving at a recommendation for rank, pay, and tenure of the faculty member.

The impetuous surge of student assessments was met with faculty resistance. The principle source of resistance arose from many assessment schemes developed by groups of individuals usually not qualified to construct such instruments. In other words, faculty in higher education were concerned with whether or not student assessment schemes asked appropriate questions, measured anything in terms of one's performance, or benefited anyone within the realm of the educational goals.

The concern over student assessment schemes is a valid one. Usually, most assessment schemes begin with an item pool, selected from a multitude of questions. From the pool, selected items make up the assessment form. Because of this, assessment form scheme developers should consider items which can be used to discriminate between educators. For an example, if two educators who teach the same course are compared, then they should appear differently on the student assessments. Simply because of individual differences, educators would not rate equally well. To determine whether or not an item discriminates between educators, a statistical technique sensitive to between group and within group responses is necessary. Menne and Tolsma (31) declared that there is insufficient discussion concerning techniques for evaluating the ability of an item to discriminate. This remains a question open to investigation.

If student assessment schemes, or any assessment scheme by any other group of assessors, are to be considered reliable measures of educator behavior, then items should have the ability to distinguish among educators. The need, therefore, of the statistical technique to determine whether or not items discriminate is obvious.

Hidlebaugh (21) points out that, as a solution to the problem of assessment by a single individual, multiple evaluator systems have been suggested. He suggests such systems would provide a solution to the "one-sided" aspect of assessment. Proponents of multi-assessor strategies point out that, in order for an assessment scheme to be equitable and objective, the various "publics" with which the teacher associates should be involved. These "publics" encompass administrators, peer educators, and students. Such a strategy would appear then to be a shrewd practice to have different "publics" assess in the process of educator performance. It is one of the purposes of this study to investigate the ability of these "publics" to discriminate between educators, as measured by the Iowa State University Student Rating Instrument.

While such strategies are not now in general use, what is needed is some evidence that multi-assessor schemes demonstrate possible new approaches to the assessment of one's performance. Such a scheme demands an approach which is different from previous attempts. While in the past, assessments were developed by other than faculty, and ultimately affected them, the different approach demands that assessment and feedback begin and end with the educator. If the purpose of this assessment is to improve instruction an approach aimed at the educator is apparent. Hidlebaugh (21, p. 27) wrote that: "Even though there is a vast quantity of student data collected on courses and teaching, rarely have these results been used to measure modification of the educator's performance." Aleamoni and Hexner (4), investigating the effect of different sets of instruction on student course and instructor evaluation, contended that many elements of the instructional setting need to be assessed by several different audiences. They further contend that most assessment schemes rest solely on the use of student assessment. It should be noted that students are able to provide reliable and valid evaluations of instructional quality. This conclusion has come to be recognized by Costin, Greenough, and Menges (11), and Aleamoni, (3). Frequently in the past, the results of these student assessments have been tabulated for student use in the selection of courses and for use by administrators as a form of assessing one's "effectiveness" in the classroom.

Equilibrium Theory

If the declared purpose of educator assessment is for the improvement in instruction, educators must have access to the tabulated results. Usually student assessment has not been made available to educators for their use in possible modification of their performance. Exception to this has been the work of Aleamoni and his associates. Modification of educator performance behavior perhaps arises from what has been discussed by some authors as "equilibrium theory" noted by Gage, Runkel, and Chatterjee (16) and Daw and Gage (12). Based upon equilibrium theory, it might be assumed educators value assessment so that they modify their performance when assessment, by assessor groups, is more or less favorable than the educator's self-concept. Accordingly, when assessor feedback creates a condition of "imbalance" (Hieder quoted in 30, p. 1), "asymmetry" (Newcomb quoted in 30, p. 1), or "dissonance" (Festinger quoted in 30, p. 1), educators will change in the direction desired by assessors, in order to establish a condition of equilibrium. Measurement of performance modification, based upon feedback and equilibrium theory, may be reflected in a second assessor assessment of educator performance behavior. These theories have a potential application to this experimental investigation.

There is some evidence that "public" feedback does indeed have a positive effect on an educator's performance, although the evidence is far from conclusive, particularly in higher education. Centra (10) cited the study by Tuckman and Oliver in which 286 teachers of vocational subjects in high school and technical institutes were used. They found that educators who received student feedback demonstrated greater "gains" in student ratings, as measured by changes in those ratings after a twelve-week interval, than did educators who received no feedback. In this study the expectation that less experienced educators were expected to change more than experienced educators was not supported. Changes in ratings of teaching were also reported by Bryan and Gage, Runkel and Chatterjee who experimented with sixth-grade teachers, according to Centra (10).

The results in higher education, however, have been far less positive. Miller (32) reported that end-of-semester student ratings for teaching assistants who received mid-semester feedback did not differ from end-of-semester ratings for teaching assistants who did not receive the feedback. But because of the small and limited sample (thirty-six teaching assistants), the results of the Miller study are inconclusive.

The preceding studies did not include a number of relevant variables nor did they consider the variation between different groups of assessors or the effect of multi-assessor feedback on modification of educator performance behavior. None of the studies investigated the educator's assessment of his own performance indicated by self-assessment. On the basis of equilibrium theory, one could hypothesize that the greater the variance between multi-assessor assessment and educator self-assessment, the greater the likelihood that there would be modification in performance behavior, since great variation would create the greatest imbalance in an educator.

Need for the Study

Educator assessment has, since the middle 1960s, been an issue of much controversy. The demand for accountability has been the primary catalyst in initiating assessment procedures. While there has been a wealth of study and literature on these topics, few people have considered or studied the effect of assessed performance as feedback on possible modification of the educators' performance. Centra (10, p. 1) reported:

> There is some evidence that student feedback does indeed have a positive effect on teaching performance, although the evidence is far from conclusive. . . . Changes in rating of teaching were also reported by Bryan (1963), . . . and by Gage, Runkel, and Chatterjee (1963).
>
> The results at the college level, however, have thus far been less positive. Miller (1971) reported that end-of-semester student ratings for teaching assistants who had received mid-semester feedback did not differ from end-of-semester ratings for teaching assistants who did not receive the feedback.

This need is increased as the expressed purpose of educator assessment becomes the improvement in instruction. Furthermore, an increasing number of educators are being dismissed from their positions because of inadequate performance behavior as determined by student assessment. Frequently the educator is not informed of inadequate performances and consequently has no opportunity to modify performance behavior. On the basis of equilibrium theory, one could hypothesize that the greater the gap between assessment and educator self-concept, the greater the likelihood that there would be change in instruction.

As Hidlebaugh (21, p. 7) reported, "several states have enacted legislation requiring accountability in education. The most notable is the law passed in California in 1971." While this legislation is aimed at secondary education, legislators in passing bills and appropriations for higher education are greatly concerned about educational costs.

Few, if any, studies or faculty assessment schemes have measured the possible effect of evaluative feedback on modifying educator performance. Furthermore, none of the studies in the review of literature considered the use of multi-assessor group's evaluative feedback on modifying performance testing equilibrium theory.

In the research completed about educator evaluation, the unit of analysis has usually been the student's ability to measure educator "effectiveness." There appears to be a lack of experimentally supportive evidence relating to the measurement of change in educator performance behavior in higher education.

Statement of the Problem

The problem was to investigate the effects and relationships of multi-assessor evaluative feedback on modifying educator performance behavior. One purpose for the investigation was to test experimentally the equilibrium theory. Another purpose was to investigate the relationships between student group assessors and peer group assessors.

Hypotheses Tested

The following hypotheses are stated in general form. They were modified for any specific test for a given assessor group on the seventeen item Educator Performance Instrument.

Hypothesis 1: There are no significant differences between the experimental group and the control group posttest mean scores as perceived by the consensual student assessment as measured by the Iowa State University Educator Performance Instrument.

Hypothesis 2: There are no significant differences between the experimental group and the control group posttest mean scores as perceived by the consensual peer assessment as measured by the Iowa State University Educator Performance Instrument.

Potential Value of this Investigation

The primary purpose of the study was to investigate the effect of multi-assessor evaluative feedback on modifying educator performance behavior. The use of evaluative feedback has potential value for the improvement of instruction. The use of multi-assessors assessment upon educator's performance, creating a condition of imbalance with the self concept of performance, may serve as a model for future strategies in formal educator assessment.

Definition of Terms

The following definitions of terms, as defined by Good (19), are presented to give clarity to their use and meaning.

1. Assessment - to set an estimated value on criteria in assessing an educator's performance inside and outside the classroom.

2. Assessing - to set an estimated value, made according to some systematic procedure, of the degree to which an individual possesses any given characteristic.

3. Accountability - holding the educational system and/or professionals responsible for results in student learning proportionate or greater than the input resources (money).

4. Behavior - an educator's manner of behaving, i.e., actions, conduct, and achievements in performance of educator responsibility.

5. Educator - a person whose chief tasks are to educate, one who is involved in the formal process of education.

6. Educator assessment - the consideration of evidence in the light of value standards, and in terms of the goals which the individual or group is striving to attain.

7. Feedback - the return of compiled data of the output to the input, i.e., to render assessment to the performer.

8. Peer - a person of the same rank, value, quality, ability, or status, etc.: equal; specifically equal before the law.

9. Performance - the act of performing; execution; accomplishment, an exhibition of skill and talent, i.e., the behaviors of the educator inside and outside the classroom.

Delimitations of the Study

This study was limited to the problem of investigating the effects and relationships of multi-assessor evaluative feedback on modifying educator performance behavior. In so doing, several evaluation schemes for faculty educators in higher education were reviewed and a search of the literature was conducted in the area of educator performance behavior modification. Selected studies were considered because of their relevancy to the problem.

Measurement of educator performance was limited to student assessors and peer assessors. The study was limited to fifty faculty educators as experimental units. These faculty educators were full-time members of the

College for Human Resources Development, University of North Dakota. The College for Human Resources Development contains seven departments, namely; Counseling and Guidance, Health Physical Education and Recreation, Home Economics, Industrial Technology, Media Education, Occupational Therapy, and Social Work. There were 850 student assessors and 150 peer assessors.

Assessment of educator performance behavior was measured by the use of the Iowa State University Student Rating Instrument as perceived by student assessor groups and peer assessor groups. These assessors were from the College for Human Resources Development in a pretest-posttest control group design with random assignment of experimental subjects.

The treatment for the experimental group was limited to the pretest data analysis of each seventeen variable educator performance characteristics of the Educator Performance Instrument and a personal conference with this investigator. The analysis was in the form of comparative and normative data, namely; the mean, standard deviation, and general discussion concerning the weaknesses and strengths on each educator performance behavior variable as perceived by student and peer group assessors. The treatment for the control group was no feedback. They were limited to only the pretest-posttest assessment by student and peer assessors.

Organization of the Study

The remainder of this investigation is organized into five chapters in the following manner. Chapter II contains a review of the literature relating to educator assessment, particularly to the measurement of educator performance behavior. Chapter III delineates the limitations of the study, research design, selection of the sample, instrumentation, treatment, data collection, and statistical methods utilized in the investigation. Chapter IV reports the findings of the statistical analysis resulting from the investigation. Chapter V consists of the discussion and, Chapter VI is the summary.

CHAPTER II. REVIEW OF LITERATURE

The review of literature was made considering the terms of educator performance versus educator effectiveness. The literature reveals that most authors do not discriminate between these two terms, but accept them as synonymous. The terms can be differentiated as described by Menne (30) He states that:

> If the behaviors of the teachers are measured in some way (e.g., by observations made by administrators, peers, or students), then the teacher's performance is being evaluated. However, if the incremental knowledge gained by the students as a consequence of the contact with a particular teacher is measured, then the teacher's effectiveness is being evaluated.

The search of literature has shown that there has been a multitude of investigations regarding teacher evaluation. In particular, it is replete concerning student ratings of teacher effectiveness. The literature that describes investigations regarding the relationships and effects of administrative, peer, and self assessment of educator performance is not as extensive.

Historical Background

Since the first formal educational setting, evaluation of the educator (teacher) has been an evident process. This process has been conducted by self, students, peers, and supervisors. It is the process that has marked the philosophy of the existentialist who asks the questions: Who am I? What am I doing? Where am I going?

The turn of the century marks the beginning of serious empirical methods of assessing educator performance. Master educators were selected to observe an educator's performance, and submit their evidence to appropriate authorities. The evidence served a two-fold purpose. The first purpose was to enable the educator to review the assessment and make the appropriate modifications to bring his performance in line with expected behavior patterns. The second purpose was to submit the educator's supervisor with sufficient evidence for incremental salary, promotion and retention, or to build a case of sufficient grounds for dismissal. Kartz (27) departed from the procedure of using master teachers and sought students' opinions by using a course evaluation questionnaire to gain information about the performance of the best teachers.

Ryans (36, p. 416), reviewed research on teacher behavior and noted that there had been a large number of research reports during the five year period preceding his review. One trend he noted was "a lessening of attention to the topic of teacher effectiveness." Isaacson, McKeachie, Milholland, Lin, Hofeller, Baerwaldt and Zinn (24, p. 344) wrote:

> "During the last 40 years many scales have been devised for rating characteristics of teaching. These scales include hundreds of different items, many of which are closely related."

Educational literature is replete with discussions of investigations that seek ways of assessing teacher performance, of predicting effectiveness, and of using various course and student questionnaires in improving instruction. Stimart and Taylor (43, p. 74), by contrast, have studied the use of a vector algebra approach of educator performance. They suggested that a basic problem with predicting excellent teachers at all educational levels has been one of the lack of a clearly defined criterion for excellence.

They proclaim that "college teaching varies by the situation and the identification of an excellent college teacher is ad hoc . . . ."

Their declaration is that "at the college level, as well as other educational levels, we can recognize which teachers conform most perfectly to any given definition." They then suggest that "in order to predict those college teachers that will do a good job, it has to be in terms of a comparison to those already in the same college." With current interest in the accountability and quality of performance in higher education, the times demand novel and innovative approaches for assessment of educator performance.

The question arises as to how this prediction can be done. Their method can be summarized as follows:

1. Explicitly define the ideal college teacher for a given college situation.

2. Identify the college teacher that most closely approximates this ideal.

3. Collect all possible descriptive data on the ideal regardless of the type of scale, i.e., ratio, nominal, or ordinal data, that it is recorded on.

4. Define a n-dimension space where n is the number of descriptive variables collected.

5. Normalize all variables and transform all scales so a score of 1 is the score the ideal received for each.

6. Identify a second ideal college teacher and eliminate any variables causing a difference between the factors for the two ideals.

7. Describe all perspective college teachers as vectors in the adjusted n-dimensional space.

8. Rank the perspective college teachers in terms of least deviation from the ideal.

The procedure for predicting excellence in an educator's performance via a vector algebra approach is different from a majority of other applications thus far.

Educators and researchers have been attempting to assess the quality of performance from the very beginning with great efforts and vast amounts of money allocated to determine the most formal and objective methods.

Researchers have produced enumerable ways of assessing educator performance. Biddle and Ellena (7, p. 6) stated that:

> Recent summaries have revealed that literally thousands of studies have been conducted on teacher excellence since the beginning of the twentieth century. Investigators have looked at teacher training, traits, behaviors, attitudes, values, abilities, sex, weight, voice quality, and many other characteristics. Teacher effects have been judged by investigators themselves, by pupils, by administrators and parents, by master teachers, by practice teachers, and by teachers themselves. The apparent result of teaching have been studied, including pupil learning, adjustment, classroom performance, sociometric status, attitudes, liking for school, and later achievement. And yet, with all this research activity, results have been modest and often contradictory. Few, if any, facts are now deemed established about teacher effectiveness, and many former "findings" have been repudiated.

Because of the voluminous quantity of research reported on teacher evaluation, this review has been limited to the following major areas: relationships among multi-assessor groups, i.e., administrators, peers, self, and students. It also includes a review of the effects of multi-assessor evaluative feedback on educator performance behavior.

Jenkins and Bausell (26, p. 572) after noting the emphases on the accountability movement suggested that "discrepancies of teacher effectiveness may be the root of the strong feelings raised by the accountability issue." Their investigation attempted to uncover some conceptions that accountability advocates might modify their approach to teacher performance.

To provide some structure for such an inquiry, they developed a survey instrument based upon categories employed by Harold Mitzel in his contribution to the 1960 edition of the Encyclopedia of Educational Research. Mitzel, after examining the kinds of criteria that numerous investigators had identified to study teaching effectiveness, perceived three categories which he labeled presage, process, and product.

Jenkins and Bausell (26, p. 572) gave a denotation of each category and they are as follows:

> Presage Criteria. When teacher evaluation is based upon one's personality or intellectual attributes, . . . his performance in training, his knowledge or achievement, . . . or his inservice status characteristics.
>
> Process Criteria. When teacher evaluation is based upon classroom behavior, either the teacher's behavior, his students' behavior, or the interplay of teacher/student behavior.
>
> Product Criteria. When teachers are judged by their effectiveness in changing student behavior, in Mitzel's scheme, product criteria. The teacher is judged on the basis of a measurable change in what is viewed as his product, student behavior. What constitutes acceptable products, or changes, has never been made altogether clear. But it would seem that measures of growth in skills, knowledge of subject matter, and attitude which could be logically or empirically attributed to the teacher's influence constitute acceptable data in the product category.

Assessors

Students

The past several years have seen a marked increase in attempts to investigate evaluation procedures of educators in higher education. Included in this increase has been an overriding study and use of student assessment of courses and educators. An analysis of the increase and decrease of student ratings was found in a report by Costin, Greenough, and Menges (11, p. 511) which summarizes two investigations by Gustad in 1961 and again in 1967. They reported that an extensive survey by Gustad, into the methods of teacher evaluation used by 584 colleges and universities, revealed that "student ratings were cited most often." More recently, however, Gustad, in 1967, reported a substantial decline in the systematic use of student ratings. He suggested that the decline in the use of student ratings was due to the lack of "convincing validity data." He further stated that "perceived threat to faculty may also be an important cause, since in recent years a strong impetus to use student ratings has come from the students themselves."

Frey (14), writing in Change Magazine, reports that significant new forces in higher education have wrought tremendous change in faculty composition, activities, and attitudes. This change is due, in part, to the growing use of formalized evaluation procedure, to assessed teaching performance and increased involvement of students in decision making.

A very recent study was reported by Greenwood, Bridges, Ware and McLean (20, p. 141) in the summer of 1974, of a new instrument called the "Student Evàluation of College Teaching Behaviors (SECTB)," developed in the College of Education at the University of Florida. SECTB represents at attempt to develop a student evàluation of teaching instrument that:

(1) is empirically derived but which reflects a broad conception of college instruction;
(2) focuses on specific teaching behaviors; and
(3) permits the students to rate only those items which they consider to be relevant.

The authors conclude that the SECTB was representative of an effort toward an empirical assessment to the appraisal of teaching behaviors.

Spencer and Aleamoni (41, p. 209) describe an instrument known as the Illinois Course Evaluation Questionnaire (CEQ). This instrument elicits student opinions about a standard set of statements relative to certain standardized aspects of an instructional program, and the norms which enable an educator to compare results of other educators. The questionnaire is

> "made up of 50 short statements. The student is asked to respond to these statements by indicating his agreement or disagreement on a four-point scale: strongly agree (SA), agree (A), disagree (D), and strongly disagree (SD). The items range from specific statements such as:
>
> 47. The instructor exhibited professional dignity and bearing in the classroom.
>
> to
>
> 42. Generally, the course was well organized."

The development of the questionnaire includes six subscales by factor analyzing the CEQ's fifty items covering basic course elements. The subscales are labeled as follows:

> (a) General Course Attitude, (b) Method of Instruction, (c) Course Content, (d) Interest and Attention, (e) Instructor, and (f) Other. Each of the subscales contains eight unique items except for Other which contains ten items.

A response set score was developed to handle careless student responses. This was done by constructing twenty-two negatively stated items that expressed approximately the same concepts as twenty-two corresponding positively stated items. The authors report that "the response set score is . . . helpful in explaining score unreliability resulting from the failure of students to know their true opinions or to express them honestly." The normative data identified by Spencer and Aleamoni were established on more than 100,000 students, 2,000 course sections, and 400 different courses. The correlation between the 22 negative and the 22 positive items for a sample of 297 CEQ's was +.849 according to Aleamoni and Spencer. They also reported that "a split-half reliability was computed with the negative and positive items in each group; thus twenty-five items in each half. The correlation result for the sample of 297 was .93."

Among the authors' conclusions they stated that "there appears to be no widely used instrument for student evaluation of courses." This conclusion is interesting since numerous attempts to investigate methods of student assessment of course and educator performance apparently has not yielded sufficient evidence to support systematic methods of assessment.

The academic community has been increasingly concerned with the accountability, effectiveness, and performance of those most immediately responsible for the education of students in higher education. Frequent attempts have been made to measure "good teaching." With a myriad of student evaluation instruments developed, none of which according to Aleamoni and Spencer (5) appear to have gained wide acceptance, "good teaching" has then been defined as good scores on the teacher evaluation form. Faculty in higher education have very little confidence in these assessments because of the varied purposes for which the assessed information has been put to use. Educators have been fired, denied promotion, and were not awarded an increment in salary solely because they received poor scores on "good teaching" evaluation forms.

Frey (15) states that he believes there is generally a positive relationship between student ratings and good teaching, however, the strength of this relationship depends critically on the technical sophistication of the rating questionnaire. He discussed some concepts of using students as assessors of educator's performance. First, he believes that it is useful to consider the student as an information source rather than as an evaluator. Secondly, he deems it necessary to treat the teaching situation as one having many dimensions that can be rated separately. Thirdly, he emphasizes the importance to take into account that students' perceptions are a product of their own personalities as well as of the educator's behavior. He concludes that "any analysis which assumes that educator assessments depend entirely on the target and are independent of their source is woefully inadequate."

Swanson and Sisson (44) investigated the use of a theoretical model for the appraisal of university faculty which identified three dimensions of performance. The model was designed to allow for assessment of teaching, scholarly productivity, and service. Their conclusion was that students are best qualified to rate the performance of the faculty but are not able to assess the research and service dimensions of an educator's performance.

Other authors agree with Swanson and Sisson, particularly Aleamoni and Yimer (6, p. 277), who found that "teachers and students differ in the basis of their rating since instructors appear to take into consideration academic rank of the instructor in their rating, while this does not appear to be the case for students."

This relationship, explains the authors, may be explainable in terms of reputation. The educator who is at an institution longer than others is apt to be known more by his colleagues. On the other hand, students are assessing the actual classroom performance they observe and are not considering the educator's overall dimensional performance behavior.

Reliability, validity, and usefulness of student assessment

There has been a great deal of research directed toward the reliability, validity, and usefulness of student assessment in assessing the activities, effectiveness, and teaching of educators. The concept of performance, as stated by Menne, i.e., performance as related to behaviors of the educator which are measured, has not received wide investigation. The methods of developing and utilizing student assessment forms have varied considerably. Frequently, questionnaire forms were developed by ill-prepared groups such as students, departmental committees, or by individuals who were attempting to define discretely the "good" performers from the "bad." Only occasionally were these instruments developed under the auspices of a group, committee, or individuals whose members were well qualified in educational measurement.

Costin, Greenough, and Menges (11, p. 511) note that:

> . . . some faculty members will frequently challenge the administration and potential use of student ratings of instruction no matter who prepares the forms. Typically, they claim that student ratings are unreliable, that the ratings favor an entertainer over the instructor who gets his material across effectively, that ratings are highly correlated with expected grades (a harder grader would thus get poor ratings), and that students are not competent judges of instruction since long-term benefits of a course may not be clear at the time it is rated.

They also contend, where criteria for salary and promotion are considered, that since "good teaching" and "good research" go hand in hand, it will suffice to reward the "good researcher."

Rodin (35, p. 67) claims that reliability of student assessment can be gained in various ways. First, investigators can ask students to assess the educator at various times during the semester, and then have the successive assessments correlated. The difficulty with this approach, she claims, is that when assessments are obtained within short duration "there is a memory factor that must be taken into account." Data in the psychological literature strongly suggest that once people commit themselves to a position, they adhere to that position regardless of subsequent evidence. Secondly, the way in which reliability has been assessed has been by an examination of the internal stability of the rating scales. The typical technique is to compare the mean score on the odd items with the mean score on the even items. Correlations obtained by this procedure yielded very high reliability and is interpreted that the instrument is a good one. Rodin investigating student evaluations concludes: "In sum, none of the standard methods for measuring the reliability of student evaluations is completely satisfactory."

Aleamoni (1, p. 1) in his several years of study and numerous investigations approaching the reliability, validity, and usefulness question has reported a wealth of data in support of the issue. He claims that "the Illinois Course Evaluation Questionnaire (CEQ) has perhaps the most extensive reliability and validity data to support it as well as the most extensive norm data base." His CEQ is used to collect data on student attitudes towards a course and educator and its purpose is to enable educators to gain evaluative feedback information about their efforts. He also sees the CEQ instrument as a source for information to be used to provide feedback to administrators if it is couched in an instructional scheme consisting of not only student assessment, but also peer and supervisor assessment.

As data is accumulated and compiled over repeated offerings of a course by an educator, it becomes possible to obtain a relatively stable indication of differences between courses he claims. Furthermore, he asserts, this enables the interpretation of the actual differences between an obtained class score for a particular educator and the average scores for all the courses taught by the educator.

It would seem, on the basis of three reliability studies by Aleamoni (2), Costin, Greenough and Menges (11), and Rodin (35), that the face validity of CEQ's and their high reliability, is ample evidence that extremely low scores on a particular course perhaps indicates some problem areas in an educator's performance as viewed by students. It is important to recognize that student opinions are in existence and they do provide a source of quite reliable and valid data relative to the effectiveness of an educator's performance.

That students should serve as the "experts" in assessing the performance of their educators, is a relatively new and revolutionary idea in the field of higher education. No one has doubted that students have opinions about an educator's performance, but only within recent times have these ideas been systematically collected. Numerous investigations have been made with conclusions for and against the use of student ratings.

Costin, Greenough, and Menges (11) state that a review of empirical studies indicates that students' ratings can provide reliable and valid information on the quality of courses and instruction. They also note that where criteria for educator performance exists, i.e., supervisor and peer ratings and measures of postinstruction student performance, student ratings tended to show a low positive correlation, suggesting that assessment does make its contribution. They claim that there was also some evidence that feedback in the form of student assessment may improve an educator's performance.

Frey (15, p. 84), considering student evaluations, makes the distinction between students as evaluators and students as information sources. He states that: "When a student makes an evaluative judgment about his teacher, he is likely to weight the specific teaching traits somewhat differently than would a faculty member or an administrator." When care is taken to develop a sound measuring instrument, instructional ratings can provide a documented record of faculty performance which is valuable to all concerned, argues Frey.

Zelenak and Snider (50, p. 570) suggest that evaluation philosophies are usually separated into two distinct beliefs. One emphasizes that the intent of evaluation is for administrative purposes, whereas the other suggests that it is for instructional purposes. In their investigation of these assumptions, their study shows rather conclusively that educators who feel student evaluation is for instructional purposes are in favor of evaluation. However, those educators who feel student evaluation is utilized for administrative purposes "(teacher tenure, promotion, dismissal, assignment, salary, and permanent record file . . .)" regard the educator evaluation process negatively.

McKeachie (29) in offering suggestion and comments regarding the utilization of evaluative procedures presents this principle of learning. He points out that, in spite of spotty evidence on the validity of student assessment of an educator's performance, feedback or knowledge of results aids learning is a psychological principle of long standing. Feedback of student assessment, coupled with other information, may be of great value to us as educators, he argues.

Student assessment and grade point average

Blum (8, p. 217) investigating the relationship existing between students' grades and their ratings of the instructor's ability to teach stated that: "In the extensive bibliography on ratings or estimations, one fails to find a reference to the problem: Are students influenced by their standing in the course in rating instructors?" He concluded following his study of two classes over an eight-week summer session that, (1) if a statistical basis for grading is used, students can estimate their grades correctly, provided they are not lower than a C, and (2) students are not influenced by their actual standing or estimated standing in the course in rating the instructor on his ability to teach the course. He further concluded that "regardless of whether a group of students receive an A, B, C, or D in the course the estimation of the instructor's ability remains essentially the same and closely resembles the average estimation of the group."

Voeks and French (46, p. 330) studied the question, Are Student Ratings of Teachers Affected by Grades? They stated: "At present we have only a few clues concerning what relationship exists between a teacher's rating by students and the grades he has assigned the raters." Their major find was that grades and student-ratings had no reliable relationship and teachers with the highest student-ratings seldom had given higher grades than teachers with the lowest ratings.

Apparently, high ratings cannot be bought by giving high grades, nor are they lost by giving low grades. They concluded:

> Both when judging their instructor's over-all value as a teacher and when rating his skill in specific respects, such as clarity of presentation and development of interest, the students rarely, if ever, were influenced by the grades which they had received from that teacher. College students appear to have greater objectivity and less superficial value systems than we have realized. Were we to heed their preceptions of our teaching abilities, we might find a rich source of clues which would enable us to increase our skills.

Students do indeed make judgments about their teachers. These judgments may be based on false or questionable criteria, but they do judge. When large numbers of students share a particular judgment about a teacher's behavior, that judgment should not be ignored. The disadvantages to student evaluation are: teachers may resent criticism; they may attempt to gain favor with the students; students may blackmail the teachers; students may not have insights into what constitutes good instruction so that they can aid a teacher in improvement.

Peer

Vielhaber and Gottheil (45) investigated first impressions and subsequent ratings of performance. These very brief observations were related to later independent ratings of performance.

Webb (47), investigating peer ratings, concluded that one of the most effective ways of evaluating complex behavior characteristics is the use of peer ratings. The procedure requires that the individual be rated by the immediate members of a group in which he is an active member. This technique has been widely used in leadership studies and is being increasingly used as a measure of job performance. His investigation of peer ratings is based upon six sections of Naval Cadets, constituting a total of one hundred seventeen (117) cases.

A biserial correlation was run between the derived standard score and the population dichotomized on the basis of having received one or more high nominations from the group versus having no nominations as a high. The resultant biserial correlation between the algebraic sum ratings and a dichotomy based on receiving or not receiving a positive nomination was +.87.

McCarter (28), writing in the American Vocational Journal, depicts an instructional evaluation system. The system admittedly comes from several subjective sources, namely: students, peers, and administrators. His consensus is that input from several sources yields sufficient information to support a manageable faculty evaluation plan for faculty assessment.

The essence of peer assessment lies in the matter of the so-called friendship factor in peer assessment. Implicit in this issue is presumably the contaminating influence of which several assumptions are noteworthy: first, that peers will be more inclined to favor friends; second, that this bias toward friends will operate independently of the people to be assessed; and third, that peer assessees' scores consequently will be weighted with popularity.

Hollander (22, p. 435) investigated these assumptions and he concluded that friendship operates as an adversely biasing and an invalidating factor in peer assessment. One of the more intriguing outgrowths of his investigation is the suggestion offered for a redirection of emphasis. "Perhaps this apparent favoring of friends does not serve to literally create status so much as it reflects a desire to have as friends those who are already manifestly high on valued status continua," states Hollander. He further concludes that the results demonstrate that "while friends appear to be favored for higher scores, the validity of peer assessment scores are not adversely affected by considerations of friendship."

Howsam (23, p. 16) discussed four types of rating scales. With respect to peer ratings he stated, "peer ratings are of limited value, due to the fact that teachers have little opportunity for observation of the work of another." Smart (39, p. 10) disagrees; he stated:

> Evaluation. . . may be done by the college administration, but is better done by colleagues, who are in a better position to judge the dignity, courtesy and temperateness of language, the patience, considerateness and pedagogical wisdom employed.

Frey (14, p. 47), commenting on teaching competencies, contended that: "The university teacher is one of few professionals whose work is seldom observed by his peers. His teaching reputation is often based more on hearsay than on substantive evidence."

It is for this reason that promotion committees frequently make final tenure decisions without seriously considering information about an educator's teaching performance. To counter this tendency, many campuses have recently instituted a system of student instructional ratings.

Administrator

Assessment of educators performance in higher education has been and remains primarily the responsibility of the department chairperson. That person is charged with subjectively applying evaluative criteria such as: evident ability as a teacher, service to the academic community, engagement in scholarly research, and creative work.

Swanson and Sisson (44, p. 64) claim evaluations by chairmen have many sources of error and frames of reference among chairmen differ rather markedly. They also stated: "Chairmen's ratings may be affected by faculty members who differ greatly in age, teaching field, sex, years of experience, abilities, and other factors within and between departments and universities." Stanley and Weiley (42, p. 12) support these statements, but in addition state "chairmen's evaluations are at best unpredictable and in many cases without validity."

The chairperson may not be best qualified to assess all dimensions of an educators performance. While the chairman's ratings may constitute the best measure of performance in one or more dimensions of an educator's performance, he cannot be considered the only assessor. Menne (30, p. 5) points out that measures of teacher performance must meet three conditions in order to have evidence of measuring anything. They are:

(1) there must be more than one rater,
(2) the raters must closely agree in their ratings, and
(3) the ratings must indicate differences between teachers.

Although the reliability of chairpersons' ratings are not usually thought to be a problem, there is some question as to the ability of the chairperson to make valid assessment of total educator performance. Validity of educator performance assessment is enhanced when conditions are described by Menne are met.

Self

Educator self-assessments have been proposed as a possible source of information for performance modification and, to a lesser extent, as an input into performance assessment. Self assessment as a basis for decisions on salary or promotions are not likely to have much validity. However, it perhaps is plausible that some form of systematic self-assessment could foster improvement in performance, particularly if coupled with external assessment supplied by students, peers, or administrators.

There has been very little research on educator self-assessment. In particular, the effects and relationships between self-assessment and those supplied by multi-assessors. Centra (10) reporting his investigation of self-ratings of college teachers quoted Webb and Nolan finding a correlation of +.62 between instructor self-ratings and student ratings. He pointed out that Clark and Blackburn, in 1971, reported a correlation of +.19 between student ratings and faculty self-ratings, and a similar +.28 correlation between self-ratings and colleague ratings. It should be noted that in both of these preceding studies, overall teaching was rated as opposed to specific instructional practices.

Although private self-evaluation is more or less continuous, even if haphazard, systematic and planned self-evaluation of educator performance is rare. Formalized, conscious procedures for modifying and refining performance of self-perception are seldom developed and most people need assistance in using self-evaluation deliberately and constructively. Some contend that educator self-evaluation is a waste of time. Such is the contention of Simpson (37, p. 35) when stating: "They contend that any use of such ratings in performance evaluation will skew the results upward."

Miller (33, p. 35) states that "the validity of these points cannot be reputed by evidence. There is none. Self-evaluation, however, can fall back upon considerable research on sensitivity and human awareness." He proclaims that two uses of self-assessment are to be recommended. One is the early-term assessment which is for the educator's eyes only, primarily to assist him in modifying performance for the balance of the term. The student, and/or administrator and/or peer assessment can be used in comparison with the educator's self-assessment. The second use provides the educator with a basis for comparison of his perceptions with those of others. This assessment can also serve as a basis for annual educator performance modification of behavior because the instrument for self-assessment is identical to the student, and other appraisors' instruments providing a basis for comparison among many assessors.

The availability of educator self-evaluation tools are numerous. Simpson and Seidman (38) report a list of seventeen educator self-evaluation tools which were prepared and distributed to 487 representatives of the American Association of Colleges for Teacher Education.

Their conclusions are as follows:

1. The tools judged most successful for self-evaluation in terms of information gathering are teacher oriented rather than student oriented.

2. Lack of knowledge about the process of self-evaluation is a restraining factor.

3. The use of self-evaluative tools is dependent upon the subject matter field involved.

4. An extremely small fraction of college instructors react almost violently to any self-evaluation proposal.

Simpson (37) reports that self-evaluation tools appear regularly in such periodicals as The School Review, Harvard Educational Review, The Clearing House, Journal of Education Psychology, and Phi Delta Kappan.

Jarrett (25, p. 41) states that self-evaluation may be used by administrators for the purpose of making sound recommendation as to promotion and tenure. However, he says "until evaluation becomes self-evaluation, at least to the extent of internalizing someone else's criticism, nothing very important has happened." He continues:

> In these days of quantified measurement and statistical manipulation and objective, public, repeatable, reliable observation and description, it takes a certain nerve to say anything in defense of subjective . . . , qualitative judgment.

Centra, investigating the effectiveness of student feedback in modifying college instruction, states that underlying the intended use of evaluation, i.e., to improve teaching, is the assumption that the instructor will use the information to alter and improve his teaching. However, he claims it is an assumption open to question.

Contrary to Jarrett's proposal that administrators use an educator's self-evaluation for promotion, Centra (10, p. 33) argues: "As a basis for decisions on promotion or salary, self-evaluations are not likely to have much validity." He suggests that it is possible that some form of systematic self-evaluation could be helpful to improve instruction, particularly if combined with external evaluations provided by students and colleagues.

Summary

A plethora of research has been conducted in the area of teacher effectiveness. There are few studies that deal with the measurement of performance behavior modification.

In determining teacher effectiveness, researchers have used a variety of strategies. The use of student raters has been popular. However, there is some question whether they are observing the wholeness of an educator performance. Some researchers argue that the use of peers (colleagues) assessment, administrator (supervisor) assessment, and self-assessment is a viable means to assessing one's performance.

Peer assessment has been discussed more frequently in recent times. One primary reason that this is gaining popularity is that it is a reaction to supervisory assessment. The use of peers may not be threatening, depending upon procedures in selection and on the peer's responsibility in assessment. The use of peers may have a disadvantage in that this scheme could establish an advisory effect since assessment identifies possible weaknesses or shortcomings.

Administrative assessment over the years has been the most conventional scheme for assessing the educator's performance. There are two difficulties in administrative assessment in that (1) a thorough satisfactory assessment consumes more time than is available, and (2) administrator assessment can be threatening to the educator.

Self-assessment can be fruitful if, according to behavioral psychologists, people will make necessary changes in what they do. The most obvious disadvantage of self-assessment is that it may not be objective or accurate, thus skewing the results upward.

CHAPTER III. METHODS AND PROCEDURES

Introduction

The objective of this chapter is to delineate the methods and procedures of this investigation. This chapter is composed of the following topics: introduction, sources of information, limitations of the study, research design, selection of the sample, instrumentation, treatment, data collection, and statistical methods.

The primary purpose of this investigation was two-fold; (a) to investigate the effects of multi-assessor feedback on educator performance and, (b) to investigate the relationship between multi-assessor groups.

The topics: educator performance, use of assessors, assessment, and possible modification of behavior in the interest of instructional improvement provided the impetus for this experimental investigation.

Sources of Information

The first major task following the development of the problem was to conduct a search of the literature. This search involved the use of ERIC in selecting appropriate studies for review, and an exhaustive search of the literature in the Iowa State University Memorial Library. Additional literature was obtained from the Educational Testing Service, Princeton, New Jersey, and the Measurement and Research Division, University of Illinois.

Research Design

In order to investigate the problem, the Pretest-Posttest Control Group design was selected. The apparent merits of this design are rather obvious in recognition of the kinds of control this investigator has on the research. When trying to answer the question of what effects feedback has on modification of performance behavior and, what comparisons exist between assessor groups, it is appropriately accomplished by this design.

The Pretest-Posttest Control Group Design necessitates an appropriate duration of time in order to allow for certain factors to evolve. It was assumed that the semester span of time would serve that requirement. The eighteen week semester allowed this investigator to schedule the pretest at mid-semester, followed immediately with feedback, and the posttest at the end of the semester. The research was conducted the fall semester 1974.

Selection of the Sample

The population for this investigation consisted of the faculty educators within the College for Human Resources Development, University of North Dakota. All faculty educators within the College were asked to participate in the investigation. Prior to mid-semester an alphabetical list of educators was obtained from the office of the Dean for the College. Each educator was assigned a three-digit identifier from 001 to 102 starting at the beginning of the alphabetical list.

Random assignment of subjects (educators) to the experimental investigation was conducted following the collection of pretest measures. Randomization of subjects was accomplished via a table of random-numbers. Fifty subjects were assigned equally to the experimental and control group from the sixty-eight subjects returning pretest data.

Limitations of the Study

This study was limited to the problem of investigating the effects and relationships of multi-assessor feedback on educator performance behavior. Measurement of educator performance was limited to student assessors and peer assessors utilizing the Educator Peformance Instrument. This investigation was limited to fifty faculty educators as experimental units. The treatment was limited to pretest data analysis administered to the experimental feedback group. Treatment to the control group was limited to no-feedback. The duration of time for treatment effect was limited to the time between mid-semester and end of semester. The statistical methods were limited to the generation of mean scores, Pearson product-moment coefficients of correlations, and analysis of variance and analysis of covariance.

The institution

The University of North Dakota is a member institution of the Association of American Universities and has been accredited by the North Central Association of Colleges and Secondary Schools since the Association was organized in 1913. Individual colleges and schools are members of the various accrediting associations in their respective fields.

The College for Human Resources Development was approved by the Board of Higher Education in March 1972. The principal purpose of the college is to prepare students for professional careers in human service occupations. Several of the departments prepare elementary and secondary educators and other school service personnel in cooperation with the Center for Teaching and Learning.

The College awards a Bachelor of Science degree in the areas of Social Work, Home Economics, Occupational Therapy, Industrial Technology, and Health Physical Education and Recreation. Graduate degrees are awarded through the Graduate School.

The student enrollment for the college, fall semester 1974, was 4,397 with 102 full-time and part-time faculty members.

Instrumentation

The instrument used in this investigation was the Iowa State University of Applied Science and Technology Student Rating Instrument. It was an instrument developed primarily by Dr. John W. Menne, Assistant Director of the Student Counseling Service and Associate Professor of Psychology, Iowa State University.

This instrument contains seventeen educator performance behavior items. These seventeen items evolved from a pool of 104 items. The instrument items were found to be valid and, by analysis of variance procedures were found to discriminate between teachers. The items had a Cronbach Alpha reliability estimate of .86 when analyzed as a person-measuring device. These results have been interpreted as an indicator of a suitable set of procedures to use in assessing educator performance behavior.

The purpose of the instrument in this investigation was (using multi-assessor groups) to measure performance behavior characteristics. It was necessary, for this investigation, to utilize an instrument which could elicit responses about a standard set of statements relative to certain standardized aspects of performance behavior. Moreover, it was important to develop feedback data which would be administered as treatment. This feedback would enable an educator to compare his assessed performance behavior with his self-concept.

For the purpose of this investigation, the name of the instrument was changed to read, Educator Performance Instrument. This name appears throughout this dissertation. The terms, item(s) and variable(s) are used interchangeably.

Treatment

Data reduction and analysis of pretest measures were conducted in the computer center at the University of North Dakota. The data from the pretest assessment measures were processed so as to yield a mean and standard deviation for each item on the educator performance inventory, from each assessor group, and on each of the fifty subjects. These statistics became the data used in the feedback, as treatment, to the experimental group.

Within one week, feedback sessions were scheduled for each experimental subject. The feedback sessions were thirty-minute personal conferences. In the conference it was the procedure to show the mean and standard deviation values for each of the seventeen educator performance variables for (1) the student assessment and (2) the peer assessment. This procedure enabled the experimental subject to view each variable. Additional treatment was administered by discussing the apparent weaknesses and strengths as perceived by the multi-assessor groups. While no ranking of subjects was conducted, subjects also had the opportunity to view the data for other subjects in the study. Complete confidentiality was maintained because numeral identification appeared on the compiled data. Subjects were allowed to make anecdotal records of their data and discussion. Rather obviously, treatment was withheld from the control group.

Data Collection

Prior to mid-semester, faculty names were placed on large manila envelopes which contained a cover letter, thirty optical scanning forms, thirty instruments, and three business size envelopes. Number two pencils were left with departmental secretaries for use in recording responses. The cover letter (see Appendix B) gave directions concerning the procedures in administering the mid-semester data collection. The faculty educators were not told the full details of the study, in particular, that assessor feedback would be purposely withheld from some of them.

Faculty educators were assured that only they would have access to their individual assessment results from student and peer assessors. This assurance undoubtedly contributed to the cooperation from the faculty educators who participated in the mid-semester data collection.

The faculty educators were instructed to select one course that they were teaching for the full duration of the semester. This would enable this investigator to collect the data both at mid-semester and at the end of the semester utilizing the identical assessor groups. This requirement eliminated some potential subjects because they were involved in teaching "mini-courses."

The method of recording the multi-assessor responses was by each assessor blackening in an appropriate space on the IBM H 95025 optical scan form. Their responses arose from reacting to each of the seventeen variables on the Educator Performance Instrument. Each assessor assessed the subject on a five-point Likert measurement scale with one, the lowest possible assessment, to five, the highest possible assessment.

Following data collection both at the mid-semester and end-of-semester, the raw data was transferred from the optical scan forms to computer card form by machine. The data was examined to correct any erroneous data transfer. Each computer card contained subject identification, assessor group, pretest or posttest, and experimental group.

Statistical Methods

The statistical methods used in this investigation were: *Pearson product-moment correlation* (r_{xy}), *Analysis of variance* (ANOVA), and *Analysis of covariance* (ANCOVA). These statistical methods enabled this investigator to analyze, describe, and draw inferences from this investigation.

Units of Statistical Analysis and Experimental Units

In this experimental investigation, a distinction was made between the *unit of statistical analysis* and the *experimental unit*. This distinction is made as described by Glass and Stanley (18).

The units of statistical analysis are the data means that were considered to be the outcomes of independent multi-assessor group responses for each of seventeen variables on each educator. If you will, the units of statistical analysis are the numbers counted for degrees of freedom "within" or for replications. Hence, the educator was the unit of statistical analysis for each of the two experimental groups on each of 17 variables. For the sake of analysis, each educator was considered to be a replication of this experiment; the experimental group was replicated 25 times, and the control group was replicated 25 times.

The experimental units are the experimental subjects (educators) that have been randomly assigned to the two experimental groups and that have responded independently of each other for the duration of this investigation's treatment. The experimental subjects are the experimental units: hence, the educator's seventeen consensual multi-assessor group mean scores are the units of statistical analysis. The means were computed upon the number of assessors for each group, consequently the student assessor groups' independent responses were greater than the peer assessor group which was limited to three independent responses. Hence, the means based on a greater number of assessors was a more accurate mean than means based on three assessors.

Generation of means

Since the Variable mean scores for each of the seventeen educator performance variables as perceived by (1) the consensual student assessor group, and (2) the consensual peer assessor group, were the units of statistical analysis used in the analyses for this investigation, the method to which these mean scores were computed, is explained. The mean scores were computed by summing the consensual assessor group's response score, for a given experimental subject and for a given variable, and dividing by the number of scores summed, i.e.,

$$\bar{X}.1 = \frac{\sum_{i=1}^{n} X_i}{N} .$$

N was equal to the number of assessors making assessments on a given educator and for a given variable. The grand means, as reported in Chapter IV, for the experimental and control group were computed by summing all mean scores for a given group and for a given variable and divided by the number of mean scores summed, i.e.,

$$\bar{X}.. = \frac{\sum_{j=1}^{n} \sum_{i=1}^{n} X_{ij}}{N} .$$

Pearson product-moment correlation

The purpose in the use of the *Pearson product-moment correlation* coefficients was to express the degree of relationship between the student group assessors and peer group assessors. This method was utilized on both the pretest and posttest means of the seventeen item Educator Assessment Instrument. The correlation coefficients of the multi-assessor groups are presented in Tables 3 to 6.

Analysis of variance

The objective in using the *Analysis of variance* (ANOVA) technique was to demonstrate the item discrimination power of the 17 variables on the Educator Assessment Instrument. The measurement units of concern were the consensual responses made to the items by members of the assessor groups in the investigation. A brief discussion and illustration of the use of analysis of variance for *variable discrimination power* follows to make its purpose clear.

The analysis of variance pattern of between group and within group variance was used to determine which variables on the Educator Performance Instrument discriminated among educators, Menne and Tolsma (31). To discriminate, a certain percentage of the total sum of squares must be due to between group variance. Since the ratio of between to within group mean squares, under the usual analysis of variance, varies as the F statistic and is also influenced by the size of the sample, it is more pragmatic to use the percentage of total sum of squares due to between groups as an appropriate index. This percentage is independent of sample size and, therefore, is an advantageous procedure.

The percentage of the *total sum of squares* (SS_{tot}) is partitioned (analyzed) into two components, the *sum of squares between groups* (SS_{bet}) and the *sum of squares within groups* (SS_{within}). Thus the ratio of between to total sum of squares, or its percentage, is an appropriate index of variable discrimination.

Characteristics of one educator can be distinguished from those of another, provided the consensual responses made by the members of the respective groups are different. In other words, the variables selected must be capable of (a) eliciting similar responses from members of the same group, and (b) eliciting different responses from members belonging to a different group when the groups in the investigation have perceived dissimilar conditions. Thus, whether or not the variables are discriminating can be inferred from the pattern of between group and within group variances. For discrimination, the within group variance should be low in relationship to the between group variance.

The following example illustrates the rationale underlying the use of a percentage of the total sum of squares (SS_{tot}) due to between groups variance as a discrimination index for a group size of sixteen members. This investigator used the group size of sixteen members, for this illustration, because this was the average number of assessors per student group for fifty groups. Moreover, sixteen members per group demonstrates an approximate minimum percentage that was used in order to have an F statistic value at .01 level of significance when measuring two educators.

Table 1. Analysis of variance for two groups with sixteen members per group

Source	DF	% SS Index	(Relative) MS	F
Between groups	2-1 = 1	21%	21	7.99**
Within groups	2(16-1) = 30	79%	2.63	
Total	31	100%		

**The critical F value with 1 and 30 degrees of freedom at .01 level of significance is 7.56.

The power to produce effects or intended results of the measuring instrument using group responses can be improved in two ways: (1) to inform the user of the minimum number per group for which the instrument was developed or, (2) to adopt a variable selection criterion (percentage) which will allow the instrument to be used effectively in the minimum practical situation for which its use was intended. In the foregoing example, the instrument was used to measure educators' performance. A reasonable minimum criterion was that approximately 21% (P less than .01) of the total sum of squares be assigned to between groups. In order to be useful in the peer assessor group, the criterion for variable selection should be that the between groups sum of squares approximate 85% (P less than .01) of the total sum of squares for the two, six-member groups.

Thus, if one knows the percentage sum of squares between, then one can tell the degrees of freedom required to make the F statistic significant. If one knows the degrees of freedom he also will be able to determine the group size on which the items will "work".

Thus, on the seventeen variable Educator Performance Instrument, items were determined to discriminate between two educators with sixteen student assessors if, as a minimum, 21 percent of the sum of squares was due to educator variance at the .01 level of significance.

The following table is used to illustrate the percentage of sum of squares required for the peer group to reach the .01 level of significance. This percentage criterion was 85% using two, three-member groups of peer assessors.

Table 2. Analysis of variance for two groups with six members per group

Source	DF	% SS Index	Relative MS	F
Between Groups	2-1 = 1	85	85	22.40**
Within Groups	2(3-1) 4	15	3.75	
Total	5	100		

**The critical F value with 1 and 4 degrees of freedom at the .01 level of significance is 21.20.

Because of the limited number of peer assessors per group, a larger percentage of the total variance must be due to between groups' variance, i.e., between educators, for variables to be judged discriminating. (F ratio significant at .01 level). It is important to point out that the actual percentage of the sum of squares due to groups be derived from a larger body of data to insure some stability to the percentage value.

The analysis of variance technique reported by Menne and Tolsma (31) was used to determine the level of significance at the .01 level. The variable discrimination analysis of the multi-assessor groups are given in Tables 7 through 14.

Analysis of covariance

The aim in the use of the *Analysis of covariance* (ANCOVA) was to test the null hypotheses to determine the level of significance for the specific test and for a given assessor group on each of the seventeen items on the

Educator Assessment Instrument. Subsequently, 34 specific hypotheses were tested.

The use of *Analysis of covariance* (ANCOVA) allows for statistical control of the pretest means. The effect of analysis of covariance is to make the two experiment groups adjusted for pretest differences by using the pretest means as the covariate with respect to each *Item variable*. This procedure gave this investigator the opportunity to view the pretest-posttest design as a measure of change.

The full model for analysis of covariance is (40)

$$y_{ij} = \mu + \alpha_i + \beta(X_{ij} - \overline{X}..) + e_{ij}$$

where

y_{ij} = observed posttest,

μ = grand mean,

α_i = treatment,

β = common pooled slope,

X_{ij} = pretest score,

$\overline{X}..$ = pretest mean, and

e_{ij} = random deviation.

The analysis of covariance was a test of the significance for the unique contribution of the group membership variables to the prediction of the y variable in the prsence of the covariate, the pretest mean score. The null hypothesis was accepted or there was insufficient evidence to reject the hypotheses at the .05 level of significance; for each of the seventeen variables as a result of (1) student assessment on the experimental subjects, and (2) peer assessment on the experimental subjects.

It should be pointed out that the analysis of measurement accuracy (item discrimination) appears in Chapter IV. However, it is not an integral part of this investigation. This analysis was conducted and reported in the interest of whether the multi-assessor groups measured accurately the experimental educator subjects.

The standard error of the mean for the Educator Assessment Instrument reveals, in an indirect way, that it is comparable with other measuring instruments such as the course evaluation questionnaire (CEQ) developed by Aleamoni (1).

CHAPTER IV. FINDINGS

Introduction

This investigation was conducted to investigate the effects and relationships of multi-assessor group's evaluative feedback on modifying educator performance behavior. The data collected and analyzed as a result of this investigation are presented in this chapter.

The chapter is divided into four subdivisions, each reporting specific aspects of the data analysis. The first subdivision reports the Pearson product-moment coefficients of correlations between student and peer assessor groups pretest variable means, posttest variable means; and between pretest and posttest variable means for each of the assessor groups. The second subdivision reports the analysis of variance (ANOVA) variable discrimination results for the student and peer assessor groups on the pretest and posttest mean scores. The third subdivision reports the means analysis, including the standard deviations and standard error of the means. The fourth subdivision reports the analysis of covariance (ANCOVA) results for the adjusted posttest differences between the experimental (feedback) group and the control (no-feedback) group.

Coefficients of Correlations

The Pearson product-moment correlation statistical procedure was used to ascertain whether the mean responses of the multi-assessor groups were correlated or independent. The hypotheses tested was whether the means correlated were equal to zero (Ho: $\rho = 0$).

Correlations were conducted for the following paired mean scores and are reported in this chapter. The paired mean scores are: student and peer assessor groups pretest variable means responses, student and peer assessor groups posttest variable means responses, pretest and posttest student assessor group variable means responses, and pretest and posttest peer assessor group variable means responses. The experimental groups pretest and posttest variable mean scores were correlated in the analysis of covariance procedure. This indicates that a percent of variance in the two measures that have been correlated is common to both. These correlations were used in the testing of the specific variable hypothesis and are reported in that subdivision.

In this investigation, the primary interest was in the possible behavior modification of an educator over the mid-semester to end-of-semester time period. Because two measures of educator performance took place by two multi-assessor groups simultaneously, at mid-semester and at the end-of-semester, the correlations of the two-groups consensual responses and, pretest and posttest group consensual responses could be found.

The correlation of assessor mean responses included approximately 800 student assessors and 150 peer assessors on 50 randomly drawn samples (educators). Fifty paired student and peer assessor group mean responses were used in this correlational analysis. The coefficients of correlation are examined on the basis of whether they reached the .05 level of significance and the .01 level of highly significant differences from zero in computing r. Table 3 presents the means, coefficients of correlation, rank order, percent of variance common to both mean scores, and the significance level.

Student-peer pretest correlations

Examination of the coefficients of correlations, given in Table 3, between the student and peer groups on the pretest variable mean scores, reveal that variables 9, 14, and 15, even though they were positive, did not reach the .05 level of significance. Results indicate that there were 14 of 17 variables with significantly high positive correlations at the .05 level. Moreover, there were 8 of the 14 variables with highly significant positive correlations at the .01 level.

Variable 4, Interest, had the highest coefficient of correlation at 0.644, disclosing a 42 percent common variance between the two variable means correlated. Variable 15, Relevance of Work, had the lowest coefficient of correlation at 0.207, indicating only 4 percent common variance.

Student-peer posttest correlations

Inspection of the coefficients of correlations, given in Table 4, between student and peer groups on posttest variable mean scores, disclose that variables 1, 5, 7, 8, 9, 10, 11, 13, 14, 16, and 17, while positive, failed to reach the .05 level of significance. One variable, number 15, was negative with a value of -.044, or nearly zero correlation. The data indicates that variables 2, 3, 4, 6, and 12 were significant at the .05 level while only variable number 3 reached the .01 level of significance with a value of 0.375, or a modest correlation. Inspection of the rank order of the coefficients of correlations with the highest positive

Table 3. Coefficients of correlation, rank order, percent of variance, and level of significance for between student and peer assessor consensual variable mean responses, pretest measures

Variable		Pretest means		r	Rank	% of
		Student	Peer		order	variance
1	Organization/planning	3.77	4.11	0.500**	(3)	0.25
2	Class time efficiency	3.65	3.82	0.570**	(2)	0.33
3	Preparedness	3.93	4.12	0.450**	(5)	0.20
4	Interest	4.11	4.14	0.644**	(1)	0.42
5	Oral presentation	3.96	3.85	0.322*	(12)	0.10
6	Written presentation	3.60	3.83	0.346*	(9)	0.12
7	Explanations	3.73	3.94	0.321*	(13)	0.10
8	Relevance	3.80	3.90	0.395**	(6)	0.17
9	Respect	3.96	4.20	0.276	(15)	0.08
10	Tolerance	3.80	4.03	0.394**	(7)	0.16
11	Fairness	3.87	4.22	0.364**	(8)	0.13
12	Availability	3.70	4.12	0.324*	(11)	0.11
13	Expectations	3.68	3.96	0.300*	(14)	0.09
14	Amount of work	3.69	3.95	0.255	(16)	0.07
15	Relevance of work	3.71	4.11	0.207	(17)	0.04
16	Evaluation	3.59	3.96	0.334*	(10)	0.11
17	Overall rating	3.92	3.99	0.476**	(4)	0.23

*,**Values of r at the .05 and .01 percent level of significance from zero is .279 and .361, respectively, N = 50 paired means.

Table 4. Coefficients of correlation, rank order, percent of variance, and level of significance for between student and peer assessor consensual variable mean responses, posttest measures

Variable		Posttest means		r	Rank order	% of variance
		Student	Peer			
1	Organization/planning	3.66	4.39	0.202	(9)	0.04
2	Class time efficiency	3.60	4.08	0.334*	(2)	0.11
3	Preparedness	3.78	4.33	0.375**	(1)	0.14
4	Interest	3.97	4.42	0.321*	(3)	0.10
5	Oral presentation	3.83	4.16	0.253	(6)	0.06
6	Written presentation	3.48	4.07	0.297*	(5)	0.09
7	Explanations	3.62	4.17	0.026	(15.5)	0.00
8	Relevance	3.71	4.18	0.150	(12)	0.02
9	Respect	3.84	4.44	0.216	(7)	0.05
10	Tolerance	3.74	4.20	0.212	(8)	0.05
11	Fairness	3.78	4.32	0.026	(15.5)	0.00
12	Availability	3.65	4.16	0.318*	(4)	0.10
13	Expectations	3.66	4.08	0.139	(13)	0.02
14	Amount of work	3.62	4.16	0.196	(10)	0.04
15	Relevance of work	3.66	4.25	-0.044	(17)	0.00
16	Evaluation	3.62	4.16	0.034	(14)	0.00
17	Overall rating	3.87	4.25	0.182	(11)	0.03

*,**Values of r at the .05 and .01 percent level of significance from zero is .279 and .361, respectively, N = 50 paired means.

significance from zero was number 3, Preparedness. This highest r of 0.375 reveals a 14 percent common variance between the two means. Four variables were without a percentage of common variance.

Pretest-posttest student correlations

The coefficients of correlation presented in Table 5, were found to be highly significant beyond the .01 level. Two variables had r values which exceeded .80 when compared with the student pretest and posttest mean scores. These variables, in rank order, were numbers 17 and 5, Overall Rating and Oral Presentation, respectively. Six of the variables r value exceeded 0.70 when compared. These variables, in rank order, were numbers 4, 10, 11, 7, 3, and 2, Interest, Tolerance, Fairness, Explanations, Preparedness, and Class Time Efficiency, respectively.

All of these r values between the pretest-posttest variable mean scores for the student assessor groups were positively high, indicating significant agreement in the way they perceived the educators' performance from pretest (mid-semester) to posttest (end-of-semester), within a very small margin of error. Ten variables had correlations ranging from 0.65 to 0.85 disclosing a percentage of common variance from 42 72 percent.

Pretest-posttest peer correlations

Examination of the coefficients of correlation presented in Table 6, were found to be highly significant beyond the .01 level except for variable number 6. This variable was significant at the .05 level with an r value of .32. Three variables had r values which exceeded .70 when compared

Table 5. Coefficients of correlation, level of significance, rank order, and percent of variance, for between pretest and posttest variable means, student assessment

Variable		Student assessment Pretest	 Posttest	r	Rank order	% of variance
1	Organization/planning	3.77	3.66	0.67**	(10)	0.45
2	Class time efficiency	3.65	3.60	0.70**	(8)	0.49
3	Preparedness	3.93	3.78	0.71**	(7)	0.50
4	Interest	4.11	3.97	0.75**	(3)	0.56
5	Oral presentation	3.96	3.83	0.81**	(2)	0.66
6	Written presentation	3.60	3.48	0.59**	(12)	0.35
7	Explanations	3.73	3.62	0.72**	(6)	0.52
8	Relevance	3.80	3.71	0.58**	(13)	0.34
9	Respect	3.96	3.84	0.69**	(9)	0.48
10	Tolerance	3.80	3.74	0.73**	(4.5)	0.53
11	Fairness	3.87	3.78	0.73**	(4.5)	0.53
12	Availability	3.70	3.65	0.55**	(14.5)	0.30
13	Expectations	3.68	3.66	0.55**	(14.5)	0.30
14	Amount of work	3.67	3.62	0.44**	(16.5)	0.19
15	Relevance of work	3.71	3.66	0.44**	(16.5)	0.19
16	Evaluation	3.59	3.62	0.61**	(11)	0.37
17	Overall rating	3.92	3.87	0.82**	(1)	0.67

**Values of r at the .05 and .01 percent level of significance from zero is .279 and .361, respectively, N = 50 paired means.

Table 6. Coefficients of correlation, level of significance, rank order, and percent of variance for between pretest and posttest variable means, peer assessment

Variable		Peer assessment Pretest	 Posttest	r	Rank order	% of variance
1	Organization/planning	4.17	4.39	0.68**	(4)	0.46
2	Class time efficiency	4.98	4.08	0.79**	(1)	0.62
3	Preparedness	4.20	4.33	0.61**	(9)	0.37
4	Interest	4.23	4.42	0.64**	(7.5)	0.41
5	Oral presentation	4.02	4.16	0.43**	(16)	0.19
6	Written presentation	3.90	4.07	0.32**	(17)	0.10
7	Explanations	3.97	4.17	0.57**	(12)	0.33
8	Relevance	4.03	4.18	0.66**	(5)	0.44
9	Respect	4.27	4.44	0.56**	(13)	0.31
10	Tolerance	4.03	4.20	0.64**	(7.5)	0.41
11	Fairness	4.22	4.32	0.71**	(2)	0.50
12	Availability	4.12	4.16	0.55**	(14)	0.30
13	Expectations	3.96	4.03	0.59**	(10.5)	0.35
14	Amount of work	3.95	4.16	0.59**	(10.5)	0.35
15	Relevance of work	4.11	4.25	0.52**	(14)	0.27
16	Evaluation	3.96	4.16	0.65**	(6)	0.42
17	Overall rating	3.97	4.25	0.70**	(3)	0.49

**Values of r at the .05 and .01 percent level of significance from zero is .279 and .361, respectively, N = 50 paired means.

with peer pretest and posttest mean scores. These three variables, in rank order, were numbers 2, 17, and 11, Class Time Efficiency, Overall Rating, and Fairness, respectively. Six variables had r values which exceeded 0.60 when compared, disclosing a marked correlation. These variables, in rank order, were numbers 1, 8, 16, 4, 10, and 3, Organization/ Planning, Relevance, Evaluation, Interest, Tolerance, and Preparedness, respectively.

Generally speaking, these r values between the peer pretest and posttest variable mean scores were positively high, denoting significant agreement in the way they perceived the educators' performance from pretest (mid-semester) to posttest (end-of-semester). Six variables had correlations ranging from 0.65 to 0.85 revealing a percentage of common variance from 42 to 72 percent.

Analysis of Measurement Accuracy

The statistical method utilized to determine item discrimination power of the seventeen variables on the Educator Performance Instrument was analysis of variance (ANOVA), Menne and Tolsma (31). To discriminate, a percentage of the total sum of squares must be due to between group variance as an index. In Menne and Tolsma's (31) scheme for item discrimination, this percentage is independent of sample size and has been used as an advantageous procedure to determine item discrimination power. A more thorough discussion of this procedure appeared in Chapter III. Methods and Procedures.

For this analysis, 21% was determined to be the minimum criterion for the student group assessors. The significance of the 21% was that, with two, 16-member assessment groups, this percentage was the minimum criterion of between group variance for differences to be significant at the .01 level.

For the purpose of expressing the measurement accuracy, the standard error of the average student assessor's pretest rating on Table 8, item 1, is explained and used as an example.

With a between group sum of squares of 140.96 and within group sum of squares of 443.30 with 49 and 762 degrees of freedom (812-1-49 = 762) yields a mean square between of 2.876 and within of 0.5817. The F value for this mean square ratio is 4.94 (highly significant). Since the pooled within group variance is 0.58, with typically 16 student assessors, the standard error of the mean is 0.19 ($\sqrt{.58/16}$). Observing Table 7, item 1, for the measurement accuracy between two educators with 24% between group sum of squares yields a significant $F_{1,30}$ value of 9.60 with 16 student assessors. Of interest, is that if the between group sum of squares is equal to or exceeds 24% for two groups of 16 per group is used, the F value will be significant.

The standard error of the average peer assessor pretest rating on Table 12, item 1, is explained. With a between group sum of squares of 59.48 and within group sum of squares of 44.42, with 49 and 92 degrees of freedom (142-1-49 = 92) computed yields a mean square of between 1.21 and within of 0.49. The F value for this mean square ratio is 2.47, again highly significant. Therefore, the pooled within group variance is 0.49, so with typically 3 peer assessors, the standard error of the mean is 0.40 ($\sqrt{0.49/3}$). Observing Table 11, item 1, for the measurement accuracy between two educators with 57% between group sum of squares does not yield a significant $F_{1,4}$ value of 5.28

Note that the pooled within group variance for peer assessors at 0.49 is slightly lower than for student assessors at 0.58. However, the standard error of the mean for peer assessors is 0.40 or over twice the error for student assessors at 0.19.

The analysis of variance procedures was conducted in four parts: (1) student assessors pretest scores, (2) student assessors posttest scores, (3) peer assessor pretest scores, and (4) peer assessors posttest scores.

Student assessors pretest scores

Inspection of Table 7 reveals the significance level reached on each of the seventeen variables for student assessors' pretest scores. All seventeen variables have an F value exceeding the .05 level of significance with 1 and 30 degrees of freedom and an F value of 4.17. The anlyysis of data yields a highly significant F value for eleven of the seventeen variables, thus reaching the critical F value of 7.36. As a consequence of these F values, there is sufficient evidence to show that the student assessors discriminated between educators on the pretest measures.

An inspection of the summary of student assessors, given in Table 8, discloses the total, within, and between sum of squares for each pretest variable. The item discrimination percentage indices are also given. The item summary analysis indicated that twelve of seventeen variables reached or exceeded the 21% minimum criterion of between group variance. Five of the item discrimination percentages ranged from 17% to 20%. Thus, the analysis of data indicated that, by and large, the items discriminated between educators.

Table 7. Summary of item discrimination analysis of results of pretest mean scores by student assessors on seventeen educator performance variables

Item	Source	DF	% of SS Index	Relative MS	F
1	Between Groups	1	24%	24	9.60**
	Within Groups	30	76%	2.5	
2	Between Groups	1	20%	20	7.41*
	Within Groups	30	80%	2.7	
3	Between Groups	1	23%	23	8.85**
	Within Groups	30	77%	2.6	
4	Between Groups	1	21%	21	8.08**
	Within Groups	30	79%	2.6	
5	Between Groups	1	25%	25	10.00**
	Within Groups	30	75%	2.5	
6	Between Groups	1	21%	21	7.5*
	Within Groups	28	79%	2.8	
7	Between Groups	1	22%	22	8.46**
	Within Groups	30	78%	2.6	
8	Between Groups	1	17%	17	6.07*
	Within Groups	30	83%	2.8	
9	Between Groups	1	26%	26	10.40**
	Within Groups	30	74%	2.5	
10	Between Groups	1	29%	29	12.08**
	Within Groups	30	71%	2.4	
11	Between Groups	1	22%	22	8.46**
	Within Groups	30	78%	2.6	

*,** The critical F ratio values with 1 and 30 degrees of freedom at the .05 level is 4.17, and at the .01 level is 7.36.

Table 7 (Continued)

Item	Source	DF	% of SS Index	Relative MS	F
12	Between Groups	1	26%	26	10.00**
	Within Groups	28	74%	2.6	
13	Between Groups	1	18%	18	6.67*
	Within Groups	30	82%	2.7	
14	Between Groups	1	24%	24	9.60**
	Within Groups	30	76%	2.5	
15	Between Groups	1	16%	16	5.71*
	Within Groups	30	84%	2.8	
16	Between Groups	1	19%	19	7.04*
	Within Groups	30	81%	2.7	
17	Between Groups	1	30%	30	13.04**
	Within Groups	30	70%	2.3	

Student assessors posttest scores

Examination of Table 9 discloses the significance level reached on each of the seventeen variables for student assessors posttest scores. Each of the seventeen variables possess F values exceeding the .05 level of significance with 1 and 28 degrees of freedom and an F of 4.20. The analysis yields ten F values exceeding the .01 level of significance at 7.64. Consequently, there is sufficient evidence to demonstrate that the student assessors discriminated between educators on the posttest measures.

Table 8. Summary of student assessors discrimination analysis on seventeen educator performance pretest variables

Item	N	SS total[a]	SS within[a]	SS between[a]	Item discrimination percentage
1	812	584.25	443.30	140.96	24
2	810	626.10	499.08	127.02	20
3	812	552.79	424.10	128.69	23
4	812	634.31	499.03	135.27	21
5	811	691.61	515.42	176.18	25
6	740	557.89	443.09	114.80	21
7	810	658.46	511.04	147.42	22
8	803	672.52	556.94	115.58	17
9	810	685.56	507.39	178.17	26
10	796	742.74	527.27	215.47	29
11	804	611.56	478.16	133.39	22
12	757	641.93	477.16	164.77	26
13	802	573.01	469.67	103.34	18
14	788	719.28	544.04	175.24	24
15	775	650.77	549.87	100.90	16
16	784	679.63	553.34	126.29	19
17	801	633.97	442.15	191.82	30

[a] Figures in these columns of the table are rounded off to hundredths.

Table 9. Summary of item discrimination analysis results of posttest mean scores by student assessors on seventeen educator performance variables

Item	Source	DF	% SS Index	Relative MS	F
1	Between Groups	1	27%	27	10.39**
	Within Groups	28	73%	2.6	
2	Between Groups	1	25%	25	9.26**
	Within Groups	28	75%	2.7	
3	Between Groups	1	26%	26	10.00**
	Within Groups	28	74%	2.6	
4	Between Groups	1	23%	23	8.21**
	Within Groups	28	77%	2.8	
5	Between Groups	1	22%	21	7.50**
	Within Groups	28	78%	2.8	
6	Between Groups	1	21%	21	7.50*
	Within Groups	28	79%	2.8	
7	Between Groups	1	23%	23	8.21**
	Within Groups	28	77%	2.8	
8	Between Groups	1	23%	23	8.21**
	Within Groups	28	77%	2.8	
9	Between Groups	1	20%	20	6.90*
	Within Groups	28	80%	2.9	
10	Between Groups	1	22%	22	7.86**
	Within Groups	28	78%	2.8	
11	Between Groups	1	17%	17	5.67*
	Within Groups	28	83%	3.0	

*,**The critical F ratio values with 1 and 28 degrees of freedom at the .05 level is 4.20, and at the .01 level is 7.64.

Table 9 (Continued)

Item	Source	DF	%	Rela-	F
			SS	tive	
			Index	MS	
12	Between Groups	1	18%	18	6.20*
	Within Groups	28	82%	2.9	
13	Between Groups	1	17%	17	5.67*
	Within Groups	28	83%	3.0	
14	Between Groups	1	18%	18	6.21*
	Within Groups	28	82%	2.9	
15	Between Groups	1	20%	20	7.00*
	Within Groups	28	80%	2.9	
16	Between Groups	1	22%	22	7.86**
	Within Groups	28	78%	2.8	
17	Between Groups	1	30%	30	12.00**
	Within Groups	28	70%	2.5	

An examination of the summary of student assessors posttest measures are presented in Table 10. This summary reveals the total, within, and between sum of squares for each variable. The item discrimination percentage indices are also presented. The item discrimination analysis indicated that twelve of seventeen variables reached or exceeded the 21% minimum criterion. Five of the item discrimination percentages ranged from 17% to 20%. Thus, the analysis of data indicated that, for the majority of items, the items discriminated between educators.

Table 10. Summary of student assessors discrimination analysis on seventeen educator performance posttest variables

Item	N	SS total[a]	SS within[a]	SS between[a]	Item discrimination percentage
1	735	594.55	433.59	160.96	27
2	739	596.42	449.36	147.06	25
3	735	603.41	446.51	156.89	26
4	738	653.75	504.21	149.53	23
5	734	635.91	494.67	141.24	22
6	681	495.69	389.87	105.82	21
7	734	651.84	503.19	148.65	23
8	733	678.02	521.38	156.64	23
9	732	682.73	544.78	137.95	20
10	727	708.23	555.75	152.48	22
11	732	627.72	519.35	108.36	17
12	706	598.51	491.89	106.62	18
13	732	580.47	481.23	99.24	17
14	720	600.60	492.23	108.37	18
15	715	632.34	507.86	124.49	20
16	733	630.58	488.94	141.65	22
17	711	634.41	446.72	187.69	30

[a] Figures in these columns of the table are rounded off to hundredths.

For this analysis of peer assessors, 85% was determined to be the minimum criterion for item discrimination. The significance of the 85% was that with two, three-member assessment groups, this percentage was the minimum criterion of between group variance for differences to be significant at the .01 level.

Peer assessor pretest scores

Examination of Table 11 reveals that none of the variables had F values reaching the significance level required to determine statistically whether items discriminated. The critical F ratio values with 1 and 4 degrees of freedom at the .05 level and .01 level is 7.71 and 21.20, respectively. As a consequence of these F values, there is some evidence to show that the peer assessors did not discriminate accurately between educators on the pretest measures.

An examination of the summary of peer assessors pretest scores are presented in Table 12. This summary reveals the total, within, and between sum of squares for each pretest variable. The item discrimination percentage indices are also presented. The item discrimination indices disclose that none of the variable between group variances reached the minimum 85% criterion.

Table 11. Summary of item discrimination analysis results of pretest mean scores by peer assessors on seventeen educator performance variables

Item	Source	DF	% SS Index	Relative MS	F*,**
1	Between Groups	1	57%	57	5.28
	Within Groups	4	43%	10.8	
2	Between Groups	1	54%	54	4.70
	Within Groups	4	46%	11.5	
3	Between Groups	1	45%	45	3.27
	Within Groups	4	55%	55	
4	Between Groups	1	48%	48	3.69
	Within Groups	4	52%	13.0	
5	Between Groups	1	50%	50	4.00
	Within Groups	4	50%	12.5	
6	Between Groups	1	45%	45	3.27
	Within Groups	4	55%	13.75	
7	Between Groups	1	42%	46	3.17
	Within Groups	4	58%	14.5	
8	Between Groups	1	46%	46	3.41
	Within Groups	4	54%	13.5	
9	Between Groups	1	37%	37	2.35
	Within Groups	4	63%	15.75	
10	Between Groups	1	39%	37	2.43
	Within Groups	4	61%	15.25	
11	Between Groups	1	37%	37	2.35
	Within Groups	4	63%	15.75	

*,** The critical F ratio values with 1 and 4 degrees of freedom at the .05 level if 7.71, and at the .01 level is 21.20.

Table 11 (Continued)

Item	Source	DF	%		F
			SS		
			Index		
12	Between Groups	1	48%	48	3.69
	Within Groups	4	53%	13.0	
13	Between Groups	1	47%	47	3.55
	Within Groups	4	53%	13.25	
14	Between Groups	1	46%	46	3.41
	Within Groups	4	54%	13.5	
15	Between Groups	1	45%	45	3.27
	Within Groups	4	55%	13.75	
16	Between Groups	1	41%	41	2.78
	Within Groups	4	59%	14.75	
17	Between Groups	1	51%	51	4.16
	Within Groups	4	49%	12.25	

Peer assessors posttest scores

Inspection of Table 13 discloses that only one of the variables had significant F values required to determine statistically whether items discriminated. The critical F ratio at .05 and .01 levels is 7.71 and 21.20, respectively. Variable 4 nearly reached the F of 7.71, falling short by a few points. Consequently, there is some statistical evidence to illustrate that the peer assessors did not discriminate between educators using the 85% minimum criterion.

Table 12. Summary of peer assessors discrimination analysis on seventeen educator performance pretest variables

Item	N	SS total[a]	SS within[a]	SS between[a]	Item discrimination percentage
1	142	103.89	44.42	59.48	57
2	140	111.74	51.33	60.41	54
3	141	129.92	72.08	57.83	45
4	142	125.08	65.17	59.91	48
5	139	128.00	64.08	63.92	50
6	122	103.18	57.00	46.18	45
7	138	108.94	63.00	45.94	42
8	138	103.88	56.58	47.30	46
9	143	108.94	68.50	40.44	37
10	141	120.00	73.08	46.92	39
11	142	119.87	75.00	44.87	37
12	141	133.40	68.75	64.65	48
13	142	120.66	64.42	56.24	47
14	138	129.54	70.33	59.20	46
15	140	106.14	58.75	47.39	45
16	140	121.54	71.58	49.96	41
17	138	106.94	52.33	54.60	51

[a] Figures in these columns of the table are rounded off to hundredths.

Table 13. Summary of item discrimination analysis results of posttest mean scores by peer assessors on seventeen educator performance variables

Item	Source	DF	% SS Index	Relative MS	$F^{*,**}$
1	Between Groups	1	58%	58	5.52
	Within Groups	4	42%	10.5	
2	Between Groups	1	60%	60	6.00
	Within Groups	4	40%	10.0	
3	Between Groups	1	52%	52	4.33
	Within Groups	4	48%	12.0	
4	Between Groups	1	64%	64	7.11
	Within Groups	4	36%	9.0	
5	Between Groups	1	50%	50	4.00
	Within Groups	4	50%	12.5	
6	Between Groups	1	63%	63	6.77
	Within Groups	4	37%	9.3	
7	Between Groups	1	56%	56	5.09
	Within Groups	4	44%	11.0	
8	Between Groups	1	58%	58	5.52
	Within Groups	4	42%	10.5	
9	Between Groups	1	59%	59	5.76
	Within Groups	4	41%	10.25	
10	Between Groups	1	51%	51	4.16
	Within Groups	4	49%	12.25	
11	Between Groups	1	53%	53	4.51
	Within Groups	4	47%	11.75	

*,**The critical F ratio values with 1 and 4 degrees of freedom at the .05 level is 7.71 and at the .01 level is 21.20.

Table 13 (Continued)

Item	Source	DF	%	Rela-	F
			SS	tive	
			Index	MS	
12	Between Groups	1	66%	66	7.77*
	Within Groups	4	34%	8.5	
13	Between Groups	1	54%	54	4.70
	Within Groups	4	46%	11.5	
14	Between Groups	1	53%	53	4.49
	Within Groups	4	47%	11.8	
15	Between Groups	1	53%	53	4.49
	Within Groups	4	47%	11.8	
16	Between Groups	1	50%	50	4.00
	Within Groups	4	50%	12.5	
17	Between Groups	1	59%	59	5.73
	Within Groups	4	41%	10.3	

An inspection of the summary of peer assessors posttest scores are presented in Table 14. This summary discloses the total, within, and between sum of squares for each posttest variable. The item discrimination percentage indices are also presented. The item discrimination indices reveal that none of the variable between group variances were sufficient to reach the 85% criterion.

Table 14. Summary of peer assessors discrimination analysis on seventeen educator performance posttest variables

Item	N	SS total[a]	SS within[a]	SS between[a]	Item discrimination percentage
1	148	65.27	27.17	38.10	58
2	147	109.85	43.58	66.27	60
3	146	80.22	38.50	41.72	52
4	147	75.69	27.50	48.19	64
5	147	84.08	42.17	41.92	50
6	139	64.42	23.58	40.83	63
7	148	81.43	35.83	45.60	56
8	148	87.32	36.83	50.48	58
9	148	74.45	30.58	43.87	59
10	148	113.51	46.67	57.84	51
11	148	81.11	38.42	42.69	53
12	147	112.08	37.75	74.33	66
13	148	84.27	38.50	45.77	54
14	144	85.31	39.83	45.47	53
15	146	74.11	35.08	39.03	53
16	148	90.11	44.67	45.44	50
17	144	85.49	35.00	50.49	59

[a] Figures in these columns of the table are founded off to hundredths.

Means Analysis

Analysis of the means was conducted to show the standard deviation as a measure of the sample mean scores variability. The standard error of the mean was also computed to disclose the error of measurement in the sampling distribution of the means.

The means, standard deviations, and standard error of the means are given in eight subdivisions. The subdivisions are: student assessment of experimental and control group pretest and posttest means, and peer assessment of experimental and control group pretest and posttest means.

Student Assessments

Pretest experimental group

A survey of the means, as reported in Table 15, reveal that the highest mean assessment among the 17 variables is variable seventeen, Overall Rating, with a mean of 4.057, quite close to the top 10 percent of all other educators compared by this multi-assessor group. Variables 4 and 5 show the largest standard deviations of 0.912 and 0.905, respectively. Variables 6 through 14 disclose standard deviations above 0.800. These same variables disclose a standard error of the mean within a range of 0.163 to 0.182.

Pretest control group

Observation of the data analysis, in Table 16, discloses that the highest mean assessment among the 17 variables is variable four, Interest, with a mean of 4.077. This score was related to the top 20 percent of all other educators compared by this group on the experimental subjects.

Table 15. Means, standard deviations, standard error of the mean for seventeen experimental group pretest scores by student assessors

Variable	Pretest mean	Standard deviation	Number of observations	Standard error of mean
1	3.810	0.439	25	0.088
2	3.771	0.386	25	0.077
3	3.967	0.424	25	0.085
4	3.991	0.912	25	0.182
5	3.953	0.905	25	0.181
6	3.470	0.817	25	0.163
7	3.687	0.878	25	0.176
8	3.739	0.875	25	0.175
9	3.877	0.886	25	0.177
10	3.724	0.866	25	0.173
11	3.786	0.852	25	0.170
12	3.602	0.881	25	0.176
13	3.613	0.827	25	0.165
14	3.585	0.868	25	0.174
15	3.759	0.382	25	0.076
16	3.654	0.369	25	0.074
17	4.057	0.422	25	0.084

Table 16. Means, standard deviations, standard error of the mean for seventeen control group pretest scores by student assessors

Variable	Pretest mean	Standard deviation	Number of observations	Standard error of mean
1	3.700	0.474	25	0.095
2	3.522	0.455	25	0.091
3	3.870	0.442	25	0.088
4	4.077	0.478	25	0.096
5	3.829	0.492	25	0.098
6	3.594	0.488	25	0.098
7	3.622	0.478	25	0.096
8	3.714	0.376	25	0.075
9	3.912	0.559	25	0.112
10	3.738	0.633	25	0.127
11	3.817	0.480	25	0.096
12	3.679	0.374	25	0.075
13	3.615	0.369	25	0.074
14	3.642	0.476	25	0.095
15	3.687	0.343	25	0.069
16	3.523	0.408	25	0.082
17	3.788	0.555	25	0.111

Variables 9 and 17 reveal a standard deviation of 0.559 and 0.555, respectively. Variables 1, 2, 3, 4, 5, 6, 7, 11, 14, and 16 show a standard deviation range of 0.408 to 0.492. The variable with the largest standard deviation was 10, Tolerance, with a value of 0.633. The standard error of the mean values disclose that fourteen variables had a score below 0.100 with the three remaining scores nearly the same value.

Posttest experimental group

Table 17 presents the findings of the means, standard deviations, and standard error of the mean. Inspection of the means show only one variable with a high mean score of 4.030; that variable is number 4, Interest. The standard deviations reveal a range of from 0.277 to 0.451 for the seventeen variables. The standard error of the mean scores were for all the variables, below 0.100.

Posttest control group

Inspection of Table 18, reveals that all the mean scores are within a range of from 3.420 to 3.907, scores related to among the top 20 percent of all educators. The standard deviations disclose that there are nine variables with values at 0.500 or just exceeding this point. Variable 17, Overall Rating, has a standard deviation of 0.657, the largest for all the variables. The standard error of the mean values all hover around 0.100.

Pretest experimental group

Reviewing the data in Table 19, discloses mean scores for all variables well above the 4.000 score. These scores range from 4.167 to

Table 17. Means, standard deviations, standard error of the mean for seventeen experimental group posttest scores by student assessors

Variable	Posttest mean	Standard deviation	Number of observations	Standard error of mean
1	3.733	0.436	25	0.087
2	3.683	0.383	25	0.077
3	3.829	0.430	25	0.086
4	4.030	0.389	25	0.078
5	3.908	0.365	25	0.073
6	3.535	0.388	25	0.078
7	3.723	0.451	25	0.090
8	3.751	0.444	25	0.089
9	3.907	0.407	25	0.081
10	3.816	0.371	25	0.074
11	3.835	0.277	25	0.055
12	3.687	0.394	25	0.079
13	3.700	0.355	25	0.071
14	3.671	0.401	25	0.080
15	3.718	0.296	25	0.059
16	3.637	0.351	25	0.070
17	3.973	0.409	25	0.082

Table 18. Means, standard deviations, standard error of the mean for seventeen control group posttest scores by student assessors

Variable	Posttest mean	Standard deviation	Number of observations	Standard error of mean
1	3.594	0.493	25	0.099
2	3.455	0.483	25	0.097
3	3.739	0.512	25	0.102
4	3.907	0.529	25	0.106
5	3.759	0.538	25	0.108
6	3.420	0.443	25	0.089
7	3.521	0.486	25	0.097
8	3.668	0.534	25	0.107
9	3.774	0.537	25	0.107
10	3.666	0.579	25	0.116
11	3.732	0.502	25	0.100
12	3.612	0.411	25	0.082
13	3.623	0.460	25	0.092
14	3.565	0.411	25	0.082
15	3.609	0.542	25	0.108
16	3.596	0.585	25	0.117
17	3.772	0.657	25	0.131

Table 19. Means, standard deviations, standard error of the mean for seventeen experimental group pretest scores by peer assessors

Variable	Pretest mean	Standard deviation	Number of observations	Standard error of mean
1	4.433	0.511	25	0.102
2	4.253	0.626	25	0.125
3	4.433	0.567	25	0.113
4	4.447	0.575	25	0.115
5	4.373	0.609	25	0.122
6	4.167	0.620	25	0.129
7	4.180	0.603	25	0.121
8	4.287	0.545	25	0.109
9	4.447	0.533	25	0.107
10	4.367	0.481	25	0.096
11	4.513	0.481	25	0.096
12	4.327	0.721	25	0.144
13	4.200	0.625	25	0.125
14	4.167	0.609	25	0.122
15	4.327	0.610	25	0.122
16	4.253	0.586	25	0.117
17	4.213	0.636	25	0.127

4.447, a difference of only 0.280 over the seventeen variables. One standard deviation appears at 0.721 for variable 12, Availability. Eight variables have standard deviation values at 0.600. These variables are: 2, 5, 6, 7, 13, 14, 15, and 17 indicating that more than half of the variables with a standard deviation of approximately 0.650. The standard errors of the means show that 15 of 17 variables have a range of 0.102 to 0.144 values, while variable means for 11 and 12, disclose values below 0.100 at 0.096 and 0.096, respectively.

Pretest control group

Examination of Table 20, reveals that the two variables, 4 and 9, Interest and Respect, had the highest mean scores of 4.007 and 4.100, respectively. All other variables have mean scores in the 3.653 to 3.963 range. The mean scores represent an overall assessment well above average and close to the Overall Rating of 3.760. The standard deviations are small, ranging from 0.492 to 0.682 standard deviation. The values for the standard error of the means are all very close to 0.100 with only variable 16 below that point.

Posttest experimental group

Upon study of Table 21, all seventeen variable means are extremely high indicating that the peer assessment perceived the experimental subjects performing within the top 10 percent of all other educators. No one mean score is extreme from the others, showing a range of 4.129 to 4.570. Variable 12, Availability, shows a standard deviation of 0.600. Eight variable means have a standard deviation value of 0.505 to 0.585

Table 20. Means, standard deviations, standard error of the mean for seventeen control group pretest scores by peer assessors

Variable	Pretest mean	Standard deviation	Number of observations	Standard error of mean
1	3.910	0.680	25	0.136
2	3.700	0.591	25	0.118
3	3.963	0.651	25	0.130
4	4.007	0.682	25	0.136
5	3.663	0.615	25	0.123
6	3.653	0.682	25	0.136
7	3.793	0.503	25	0.101
8	3.777	0.547	25	0.109
9	4.100	0.517	25	0.103
10	3.683	0.528	25	0.106
11	3.933	0.525	25	0.105
12	3.917	0.607	25	0.121
13	3.723	0.605	25	0.121
14	3.740	0.661	25	0.132
15	3.883	0.502	25	0.100
16	3.670	0.492	25	0.098
17	3.760	0.576	25	0.115

Table 21. Means, standard deviations, standard error of the mean for seventeen experimental group posttest scores by peer assessors

Variable	Posttest mean	Standard deviation	Number of observations	Standard error of mean
1	4.480	0.452	25	0.090
2	4.217	0.564	25	0.113
3	4.380	0.445	25	0.089
4	4.560	0.525	25	0.105
5	4.287	0.485	25	0.097
6	4.129	0.585	25	0.119
7	4.260	0.532	25	0.106
8	4.340	0.463	25	0.093
9	4.570	0.466	25	0.093
10	4.440	0.441	25	0.088
11	4.437	0.505	25	0.101
12	4.310	0.600	25	0.120
13	4.267	0.511	25	0.102
14	4.380	0.533	25	0.107
15	4.377	0.473	25	0.095
16	4.267	0.538	25	0.108
17	4.373	0.472	25	0.094

with seven variables near 0.430. All seventeen standard error of the mean values are below 0.120 with a majority of them below 0.100.

Posttest control group

An observation of Table 22, shows that 13 of 17 variables have a mean score ranging from 4.007 to 4.300 with variables 1 and 9, Class Time Efficiency and Respect having the highest assessment. The remaining four variables have range of 3.893 to 3.953 disclosing a rather high mean assessment for all variables. Eleven variables have a standard deviation value in the area of 0.530 with four variables at nearly 0.630, and the highest value at 0.826 for variable 12, Availability. All of the standard errors of the means are essentially at 0.100 to 0.165.

Analysis of Covariance

Analysis of covariance was utilized to test the mean posttest differences between the experimental group and the control groups. The group's pretest mean scores were employed as the covariate to allow for the adjustment of initial differences between the two groups with respect to the independent variables that were related to the posttest mean scores.

The general model for the analysis of covariance as described by Snedecor and Cochran (40) with two covariates is as follows:

$$Y_{ij} = \mu + \alpha_i + \beta_1 X_1 + \beta_2 X_2 + e_{ij}$$

where

Y_{ij} = observed posttest mean score, $i = 1,2,\ j = 1,\ldots n$

μ = overall grand mean

Table 22. Means, standard deviations, standard error of the mean for seventeen control group posttest scores by peer assessors

Variable	Posttest mean	Standard deviation	Number of observations	Standard error of mean
1	4.300	0.561	25	0.112
2	3.933	0.728	25	0.146
3	4.280	0.614	25	0.123
4	4.280	0.597	25	0.119
5	4.040	0.562	25	0.112
6	4.013	0.516	25	0.103
7	4.080	0.538	25	0.108
8	4.013	0.633	25	0.127
9	4.300	0.573	25	0.115
10	3.953	0.680	25	0.136
11	4.200	0.551	25	0.110
12	4.007	0.826	25	0.165
13	3.893	0.531	25	0.106
14	3.947	0.504	25	0.101
15	4.127	0.512	25	0.102
16	4.047	0.535	25	0.107
17	4.133	0.646	25	0.129

α_i = treatment effect

β_1 = partial regression coefficient of Y on X_1

X_1 = the deviation of X_{1ij} from the overall mean of X_1

β_2 = partial regression coefficient of Y on X_2

X_2 = the deviation of X_{2ij} from the overall mean of X_2

e_{ij} = random deviation

The group variable mean scores were units of statistical analysis. The number of observations was 17 and analysis of the adjusted posttest means was obtained with one covariate in the analysis of covariance. The coefficient of correlation is included in the report of the finding for between pretest and posttest variable means. The critical F value with 1 and 47 degrees of freedom at .05 level of significance is 3.97 and at .01 level of significance is 6.99.

Hypotheses Tested, Experimental/Control by Student Assessment

Ho_1--There was no significant difference between the experimental and control group adjusted posttest mean scores as perceived by student assessors on variable 1. (Organization/Planning)

Inspection of the data analysis for this variable for difference between the two experimental groups yielded a nonsignificant F value. This F value of 0.15, as reported in Table 23, was computed by analysis of covariance using the pretest mean as the covariate. This analysis failed to yield sufficient evidence to reject the null hypothesis. Consequently, there was no statistically significant difference between the two experimental groups.

Table 23. Analysis of covariance for experimental and control groups as perceived by student assessors, variable 1

Source of variation	d.f.	Residuals S.S.	Residuals M.S.	F
Experiment	1	0.018	0.018	0.15
Error	47	5.734	0.122	
Corrected total	48	5.752		

Ho_2--There was no significant difference between the experimental and control groups adjusted posttest mean scores as perceived by student assessors on variable 2. (Class Time Efficiency)

Examination of the findings in Table 24 reveals that a nonsignificant F value of 0.20. This was computed by analysis of covariance using the pretest mean as the covariate. This analysis failed to yield sufficient evidence to reject the null hypothesis. Therefore, there was no statistically significantly difference between the two experimental groups.

Table 24. Analysis of covariance for experimental and control groups as perceived by student assessors, variable 2.

Source of variation	d.f.	Residuals S.S.	Residuals M.S.	F
Experiment	1	0.020	0.020	0.20
Error	47	4.706	0.100	
Corrected total	48	4.726		

Ho_3--There was no significant difference between the experimental and control groups adjusted posttest mean scores as perceived by student assessors on variable 3. (Preparedness)

Inspection of the findings in Table 25 indicates that a nonsignificant F value of 0.00 was computed. This indicates that there was lutely no difference between the two groups. This analysis fails to yield sufficient evidence to reject the null hypothesis after adjusting the posttest means.

Table 25. Analysis of covariance for experimental and control groups as perceived by student assessors, variable 3

Source of variation	d.f.	Residuals S.S.	Residuals M.S.	F
Experiment	1	0.000	0.000	0.00
Error	47	5.302	0.113	
Corrected total	48	5.302		

Ho_4--There was no significant difference between the experimental and control groups adjusted posttest mean scores as perceived by student assessors on variable 4. (Interest in Teaching)

Observing Table 26, discloses that a nonsignificant F value of 0.79 was found. This shows evidence that there was no posttest difference between the two groups tested. The results fail to show sufficient evidence to reject the null hypothesis after adjusting the posttest means.

Table 26. Analysis of covariance for experimental and control groups as perceived by student assessors, variable 4

Source of variation	d.f.	Residuals S.S.	M.S.	F
Experiment	1	0.075	0.075	0.79
Error	47	4.468	0.095	
Corrected total	48	4.543		

Ho_5--There was no significant difference between the experimental and control groups adjusted posttest mean scores as perceived by student assessors on variable 5. (Oral Presentation)

The analysis yielded an F value which was not significant between the two groups. The results of the analysis are presented in Table 27. The results indicate an F value of 0.67 which is insufficient evidence to reject the null hypothesis after adjusting the posttest means.

Table 27. Analysis of covariance for experimental and control groups as perceived by student assessors, variable 5

Source of variation	d.f.	Residuals S.S.	M.S.	F
Experiment	1	0.056	0.056	0.67
Error	47	3.960	0.081	
Corrected total	48	4.016		

Ho_6--There was no significant difference between the experimental and control groups adjusted posttest mean scores as perceived by student assessors on variable 6. (Written Presentation)

The results for the analysis of this variable are presented in Table 28. The findings failed to yield sufficient evidence to reject the null hypothesis with an F value of 1.23. This value would indicate that the difference between the two experimental groups, following the removal of the covariate and subsequent adjustment of the posttest mean scores, was short of reaching the critical F value to be significant.

Table 28. Analysis of covariance for experimental and control groups as perceived by student assessors, variable 6

Source of variation	d.f.	Residuals S.S.	Residuals M.S.	Residuals F
Experiment	1	0.141	0.141	1.23
Error	47	5.364	0.114	
Corrected total	48	5.505		

Ho_7--There was no significant difference between the experimental and control groups adjusted posttest mean scores as perceived by student assessors on variable 7. (Explains Material Clearly)

The analysis of covariance F value of 0.20 presented in Table 29 was not significant. Therefore, there was insufficient evidence to reject the null hypothesis. The null hypothesis implying no difference between the two groups after adjusting for initial differences would occur.

Table 29. Analysis of covariance for experimental and control groups as perceived by student assessors, variable 7

Source of variation	d.f.	Residuals S.S.	M.S.	F
Experiment	1	0.022	0.022	0.20
Error	47	5.205	0.111	
Corrected total	48	5.227		

Ho_8--There was no significant difference between the experimental and control groups adjusted posttest mean scores as perceived by student assessors on variable 8. (Relevance of material used in instruction)

Examination of the analysis of covariance yields an F value of 0.10 derived from the analysis and is presented in Table 30. This value indicates a nonsignificant F value for difference between the two groups. The null hypothesis was no rejected on the basis of insufficient analytical evidence falling short of the critical F value at the .05 level of significance.

Table 30. Analysis of covariance for experimental and control groups as perceived by student assessors, variable 8

Source of variation	d.f.	Residuals S.S.	M.S.	F
Experiment	1	0.017	0.017	0.10
Error	47	7.823	0.167	
Corrected total	48	7.840		

Ho_9--There was no significant difference between the experimental and control groups and adjusted posttest mean scores as perceived by student assessors on variable 9. (Respect for Students)

The data presented in Table 31 reveal an F value of 0.49. Consequently, the analysis failed to yield sufficient evidence to reject the null hypothesis. This indicates that there was no significant statistical difference between the two experimental groups adjusted posttest mean scores.

Table 31. Analysis of covariance for experimental and control groups as perceived by student assessors, variable 9

Source of variation	d.f.	Residuals S.S.	Residuals M.S.	F
Experiment	1	0.058	0.058	0.49
Error	47	5.594	0.119	
Corrected total	48	5.652		

Ho_{10}--There was no significant difference between the experimental and control groups adjusted posttest mean scores as perceived by student assessors on variable 10. (Tolerance of weak students and differing opinions)

The results, presented in Table 32, of the analysis yielded an F value of 0.45. This result fails to supply sufficient evidence to reject the null hypothesis. The result, however, favors the experimental group indicating a token difference as perceived by student assessors.

Table 32. Analysis of covariance for experimental and control groups as perceived by student assessors, variable 10

Source of variation	d.f.	Residuals S.S.	M.S.	F
Experiment	1	0.051	0.051	0.45
Error	47	5.247	0.112	
Corrected total	48	5.298		

Ho_{11}--There was no significant difference between the experimental and control groups adjusted posttest mean scores as perceived by student assessors on variable 11. (Fairness with Students)

The results of the analysis was found to be nonsignificant as displayed in Table 33. The calculated F value was 0.14. After the initial differences were adjusted, with respect to the pretest mean covariate, there was insufficient evidence to reject the null hypothesis.

Table 33. Analysis of covariance for experimental and control groups as perceived by student assessors, variable 11

Source of variation	d.f.	Residuals S.S.	M.S.	F
Experiment	1	0.011	0.011	0.14
Error	47	3.674	0.078	
Corrected total	48	3.685		

Ho_{12}--There was no significant difference between the experimental and control groups adjusted posttest mean scores as perceived by student assessors on variable 12. (Availability outside of classroom)

Examination of the results in Table 34 disclose a nonsignificant F value of 0.36. This value is insufficient evidence to reject the null hypothesis with respect to the adjusted means. The adjustment of the posttest mean scores are in favor of the experimental group.

Table 34. Analysis of covariance for experimental and control groups as perceived by student assessors, variable 12

Source of variation	d.f.	Residuals S.S.	Residuals M.S.	F
Experiment	1	0.042	0.042	0.36
Error	47	5.459	0.116	
Corrected total	48	5.501		

Ho_{13}--There was no significant difference between the experimental and control groups adjusted posttest mean scores as perceived by student assessors on variable 13. (Expectations)

Inspection of Table 35 reveals a nonsignificant F value. The calculated F value was 0.00, indicating no difference whatsoever. The results of this finding yields insufficient evidence to reject the null hypothesis. This indicates that after initial differences were removed, there was no difference between the two groups.

Table 35. Analysis of covariance for experimental and control groups as perceived by student assessors, variable 13

Source of variation	d.f.	Residuals S.S.	Residuals M.S.	F
Experiment	1	0.001	0.000	0.00
Error	47	5.736	0.122	
Corrected total	48	5.737		

Ho_{14}--There was no significant difference between the experimental and control groups adjusted posttest mean scores as perceived by student assessors on variable 14. (Amount of work required)

The analysis of covariance presented in Table 36 revealed a nonsignificant F value. The calculated F value was 0.46. The results of this analysis yields insufficient evidence to reject the null hypothesis. This implies that there was no difference between the two groups when initial differences between the two groups, with respect to the covariate, had been adjusted.

Table 36. Analysis of covariance for experimental and control groups as perceived by student assessors, variable 14

Source of variation	d.f.	Residuals S.S.	Residuals M.S.	F
Experiment	1	0.062	0.062	0.46
Error	47	6.381	0.136	
Corrected total	48	6.443		

Ho_{15}--There was no significant difference between the experimental and control groups adjusted posttest mean scores as perceived by student assessors on variable 15. (Relevance of Work)

The analysis of covariance results are presented in Table 37. These results yield a nonsignificant F value. The calculated F value was 0.62. Therefore, there was insufficient evidence to reject the null hypothesis. This indicates that whatever the difference was, it fell far short of reaching the .05 level of significance.

Table 37. Analysis of covariance for experimental and control groups as perceived by student assessors, variable 15

Source of variation	d.f.	Residuals S.S.	Residuals M.S.	F
Experimental	1	0.097	0.097	0.62
Error	47	7.342	0.156	
Corrected total	48	7.439		

Ho_{16}--There was no significant difference between the experimental and control groups adjusted posttest mean scores as perceived by the student assessors on variable 16. (Evaluation procedures appropriate)

The analysis of covariance results are presented in Table 38. These results disclose an F value which is not significant. The F value calculated was 0.62. This value indicates that the difference between the two groups failed to reach the .05 level of significance. Therefore, there was insufficient evidence to reject the null hypothesis following the

Table 38. Analysis of covariance for experimental and control groups as perceived by student assessors, variable 16

Source of variation	d.f.	Residuals S.S.	M.S.	F
Experiment	1	0.033	0.033	0.22
Error	47	7.058	0.150	
Corrected total	48	7.091		

adjustment of means in the presence of the pretest mean covariate.

Ho_{17}--There was no significant difference between the experimental and control groups adjusted posttest mean scores as perceived by the student assessors on variable 17. (Overall Rating)

The difference between the two groups was found to be a nonsignificant F value. The calculated F value was 0.06 as reported in Table 39. The results of this analysis reveal that the F value fell far short of reaching the .05 level of significance. Therefore, there was insufficient evidence to reject the null hypothesis following adjustment of posttest means in the presence of the pretest mean covariate.

Table 39. Analysis of covariance for experimental and control groups as perceived by student assessors, variable 17

Source of variation	d.f.	Residuals S.S.	M.S.	F
Experiment	1	0.006	0.006	0.06
Error	47	4.970	0.106	
Corrected total	48	4.976		

Hypotheses Tested, Experimental/Control by Peer Assessment

Ho_{18}--There was no significant difference between the experimental and control groups adjusted posttest mean scores as perceived by peer assessors on variable 1. (Organization/Planning)

The results of the analysis are given in Table 40. A nonsignificant F value resulted after the adjustment of the posttest means for differences between the two groups. The calculated F value is 0.65. This value yields insufficient evidence to reject the null hypothesis. As a consequence of this F test value, there was no significant difference between the two experimental groups.

Table 40. Analysis of covariance for experimental and control groups as perceived by peer assessors, variable 1

Source of variation	d.f.	Residuals S.S.	 M.S.	 F
Experiment	1	0.104	0.104	0.65
Error	47	7.502	0.160	
Corrected total	48	7.606		

Ho_{19}--There was no significant difference between the experimental and control groups adjusted posttest mean scores as perceived by peer assessors on variable 2. (Class Time Efficiency)

Inspection of Table 41 reveals the data analysis results for this variable. A nonsignificant F value was found, indicating no significant difference between the two groups. The calculated F value is 1.21.

Table 41. Analysis of covariance for experimental and control groups as perceived by peer assessors, variable 2

Source of variation	d.f.	Residuals S.S.	Residuals M.S.	Residuals F
Experiment	1	0.241	0.241	1.21
Error	47	9.313	0.198	
Corrected total	48	9.354		

This finding yields insufficient results to reject the null hypothesis.

Ho_{20}--There was no significant difference between the experimental and control groups adjusted posttest mean scores as perceived by peer assessors on variable 3. (Preparedness)

Examination of the findings in Table 42 reveal that a nonsignificant F value was found following the analysis for this variable. The analysis of covariance calculated F value is 1.04. This value indicates that there is insufficient evidence to reject the null hypothesis for no significant differences between the two experimental groups.

Table 42. Analysis of covariance for experimental and control groups as perceived by peer assessors, variable 3

Source of variation	d.f.	Residuals S.S.	Residuals M.S.	Residuals F
Experiment	1	0.204	0.204	1.04
Error	47	9.238	0.197	
Corrected total	48	9.442		

Ho_{21}--There was no significant difference between the experimental and control groups adjusted posttest mean scores as perceived by peer assessors on variable 4. (Interest in Teaching)

The results of the data analysis for this variable are presented in Table 43. The results indicate that a nonsignificant F value was found. The calculated F value is 0.07 indicating virtually no difference between the two experimental groups following the adjustment of posttest scores in the presence of the covariate. Therefore, there is insufficient evidence to reject the null hypothesis for no difference between groups.

Table 43. Analysis of covariance for experimental and control groups as perceived by peer assessors, variable 4

Source of variation	d.f.	Residuals S.S.	M.S.	F
Experiment	1	0.014	0.014	0.07
Error	47	9.238	0.197	
Corrected total	48	9.252		

Ho_{22}--There was no significant difference between the experimental and control groups adjusted posttest mean scores as perceived by peer assessors on variable 5. (Oral Presentation)

Review and examination of the findings resulting from the analysis of the data for this variable indicate a nonsignificant F value. The calculated F value is 0.00 and found in Table 44. This value indicates that there was absolutely no difference, as perceived by peer assessors,

Table 44. Analysis of covariance for experimental and control groups as perceived by peer assessors, variable 5

Source of variation	d.f.	Residuals S.S.	M.S.	F
Experiment	1	0.115	0.115	0.00
Error	47	11.174	0.238	
Corrected total	48	11.289		

between the two experimental groups. Consequently, the results yield insufficient evidence to reject the null hypothesis.

Ho_{23}--There was no significant difference between the experimental and control groups adjusted posttest mean scores as perceived by peer assessors on variable 6. (Written Expression)

The findings for the analysis of the data for the difference between the two experimental groups is presented in Table 45. The results yield a nonsignificant F value. The analysis found an F value of 0.00 which is short of reaching the critical F value at the .05 level. Therefore, there is insufficient evidence to reject the null hypothesis for no difference between the two groups.

Table 45. Analysis of covariance for experimental and control groups as perceived by peer assessors, variable 6

Source of variation	d.f.	Residuals S.S.	M.S.	F
Experiment	1	0.001	0.001	0.00
Error	47	12.993	0.283	
Corrected total	48	12.994		

Ho_{24}--There was no significant difference between the experimental and control groups adjusted posttest mean scores as perceived by peer assessors on variable 7. (Explanation of Material Clearly)

The analysis of covariance results for the data on this variable are presented in Table 46. The analysis yields a nonsignificant F value. The calculated F value is 0.03 following the adjustment of posttest scores. These results suggest that there was insufficient evidence to reject the null hypothesis indicating no difference between the two experimental groups.

Table 46. Analysis of covariance for experimental and control groups as perceived by peer assessors, variable 7

Source of variation	d.f.	Residuals S.S.	Residuals M.S.	Residuals F
Experiment	1	0.006	0.006	0.03
Error	47	9.622	0.205	
Corrected total	48	9.628		

Ho_{25}--There was no significant difference between the experimental and control groups adjusted posttest mean scores as perceived by peer assessors on variable 8. (Relevance of Material Used in Instruction)

Inspection of the findings for the data on this variable, as perceived by peer assessors, are presented in Table 47. These findings yield a nonsignificant F value for difference between the two experimental groups. The calculated F value is 0.00 indicating absolutely no difference

Table 47. Analysis of covariance for experimental and control groups as perceived by peer assessors, variable 8

Source of variation	d.f.	Residuals S.S.	M.S.	F
Experiment	1	0.000	0.000	0.00
Error	47	8.913	0.190	
Corrected total	48	8.913		

following adjusted mean scores. As a consequence, there is insufficient evidence to reject the null hypothesis.

Ho_{26}--There was no significant difference between the experimental and control groups adjusted posttest mean scores as perceived by peer assessors on variable 9. (Respect for Students)

Examination of the results of the data analysis for this variable are presented in Table 48. These results yield an F value of 0.38 which is not significant. This F value indicates that there was only a very small difference between the two groups, falling far short of the critical F value at the .05 level of significance. Consequently, there is insufficient evidence to reject the null hypothesis for no difference between the two groups as perceived by peer assessors.

Table 48. Analysis of covariance for experimental and control groups as perceived by peer assessors, variable 9

Source of variation	d.f.	Residuals S.S.	M.S.	F
Experiment	1	0.733	0.733	0.38
Error	47	9.167	0.195	
Corrected total	48	9.900		

Ho_{27}--There was no significant difference between the experimental and control groups adjusted posttest mean scores as perceived by peer assessors on variable 10. (Tolerance of weak students and differing opinions)

The analysis of covariance results on this variable for difference between the experimental and control groups are presented in Table 49. These findings yield a nonsignificant F value at the .05 level of significance. The calculated F value is 0.07 falling far short of the .05 level. As a result of this analysis, there is insufficient evidence to reject the null hypothesis for no difference between the two groups.

Table 49. Analysis of covariance for experimental and control groups as perceived by peer assessors, variable 10

Source of variation	d.f.	Residuals S.S.	Residuals M.S.	F
Experiment	1	0.016	0.016	0.07
Error	47	10.574	0.225	
Corrected total	48	10.590		

Ho_{28}--There was no significant difference between the experimental and control groups adjusted posttest mean scores as perceived by peer assessors on variable 11. (Fairness with Students)

Examination of the findings, presented in Table 50, indicate a nonsignificant F value for no difference between the two experimental groups. The calculated F value is 1.10, falling short of the required value to be significant at the .05 level of significance. As a result of the data analysis for this variable, there is insufficient evidence to reject the

Table 50. Analysis of covariance for experimental and control groups as perceived by peer assessors, variable 11

Source of variation	d.f.	Residuals S.S.	M.S.	F
Experiment	1	0.191	0.191	1.10
Error	47	8.180	0.174	
Corrected total	48	8.371		

null hypothesis.

Ho$_{29}$--There was no significant difference between the experimental and control groups adjusted posttest mean scores as perceived by peer assessors on variable 12. (Availability Outside of Classroom)

The results of the analysis for the data concerning this test is presented in Table 51. The findings in this analysis yields a nonsignificant F value. This F value was calculated to be 0.14, far below the critical .05 level of significance required to be significant. As a result of this finding, it is therefore determined that there was insufficient evidence to reject the null hypothesis. Thus, there was no significant difference between the two experimental groups after the covariate was taken out.

Table 51. Analysis of covariance for experimental and control groups as perceived by peer assessors, variable 12

Source of variation	d.f.	Residuals S.S.	M.S.	F
Experiment	1	0.052	0.052	0.14
Error	47	17.940	0.382	
Corrected total	48	17.992		

Ho_{30}--There was no significant difference between the experimental and control groups adjusted posttest mean scores as perceived by peer assessors on variable 13. (Expectations)

The examination of the results from the analysis of data concerning group difference is presented in Table 52. The calculated F value is 1.12 falling far short of the required F value to be of significant difference at the .05 level. This nonsignificant F value indicates that there was no significant difference using the pretest mean as the covariate on the posttest. Consequently, the adjusted means difference does not yield sufficient evidence to reject the null hypothesis.

Table 52. Analysis of covariance for experimental and control groups as perceived by peer assessors, variable 13

Source of variation	d.f.	Residuals S.S.	Residuals M.S.	F
Experiment	1	0.205	0.205	1.12
Error	47	8.608	0.183	
Corrected total	48	8.813		

Ho_{31}--There was no significant difference between the experimental and control groups adjusted posttest mean score as perceived by peer assessors on variable 14. (Amount of Work Required)

Inspection of the results for the analysis of data regarding this variable is presented in Table 53. These results indicate a nonsignificant F value, thus, confirming that there was no difference between the two experimental groups. The F value calculated was 3.18 approaching the critical F value at .05 level of significance. Since there was

Table 53. Analysis of covariance for experimental and control groups as perceived by peer assessors, variable 14

Source of variation	d.f.	Residuals S.S.	Residuals M.S.	F
Experiment	1	0.549	0.549	3.18
Error	47	8.122	0.173	
Corrected total	48	8.671		

insufficient evidence, the null hypothesis is not rejected.

Ho_{32}--There was no significant difference between the experimental and control groups adjusted posttest mean scores as perceived by peer assessors on variable 15. (Relevance of Work)

Examination of the results, presented in Table 54, disclose that a nonsignificant F value was in evidence following the analysis of data for this variable. The calculated F value is 0.18, far short of the required value to be statistically significant. As a consequence of these findings, there shows insufficient evidence to reject the null hypothesis. Therefore, there was no significant difference between the two experimental groups.

Table 54. Analysis of covariance for experimental and control groups as perceived by peer assessors, variable 15

Source of variation	d.f.	Residuals S.S.	Residuals M.S.	F
Experiment	1	0.034	0.034	0.18
Error	47	8.813	0.188	
Corrected total	48	8.847		

Ho_{33}--There was no significant difference between the experimental and control groups adjusted posttest mean scores as perceived by peer assessors on variable 16. (Evaluation Procedure Appropriate)

A review of the results for the data analysis on this variable is presented in Table 55. These results indicate that there was insufficient evidence to reject the null hypothesis. An F value of 0.69 was calculated. This indicates that there was no statistically significant difference between the two experimental groups adjusted posttest means.

Table 55. Analysis of covariance for experimental and control groups as perceived by peer assessors, variable 16

Source of variation	d.f.	Residuals S.S.	Residuals M.S.	F
Experiment	1	0.134	0.134	0.69
Error	47	9.089	0.193	
Corrected total	48	9.223		

Ho_{34}--There was no significant difference between the experimental and control groups adjusted posttest mean scores as perceived by peer assessors on variable 17. (Overall Rating)

The examination of Table 56 shows that the analysis of data for difference between the two experimental groups yielded a nonsignificant F value. The F value calculated on the adjusted posttest means was 0.12. This value is far short of reaching the critical F value required if the two groups were statistically different. As a consequence of this data

Table 56. Analysis of covariance for experimental and control groups as perceived by peer assessors, variable 17

Source of variation	d.f.	Residuals		
		S.S.	M.S.	F
Experiment	1	0.021	0.021	0.12
Error	47	8.418	0.179	
Corrected total	48	8.439		

analysis, there is insufficient evidence to reject the null hypothesis.

Summary of Findings

To summarize the findings, the correlations between student and peer assessor groups assessment yielded significant agreement between the two group pretest and posttest mean scores. Also, there was significant agreement between pretest and posttest measures for each of the two assessor groups.

The analysis of variance for item discrimination yielded significant F values for student pretest and posttest assessments. Significant F values were not in evidence for the peer assessor groups.

The means analysis yielded standard deviations centered around 0.3 to 0.6. This implies that there was a small degree of variance in the assessments.

Moreover, end-of-semester assessment of experimental group subjects who were given mid-semester feedback did not differ, as hypothesized, from the control group. These findings were yielded by analysis of covariance procedures.

CHAPTER V. DISCUSSION

This investigation was concerned with the effects and relationships of feedback on educator performance. The major concerns in the investigation were: the effects of multi-assessor group evaluative feedback on modifying educator performance behavior, the relationships between student assessor groups and peer assessor groups, and the efficacy of the measuring instrument used in this investigation.

Analysis of Differences Between Experimental Groups

Presumably, on the basis of equilibrium theory, educators value assessor feedback enough to change their performance behavior. As a test of this theory the pretest and posttest measures were analyzed using analysis of covariance to ascertain the differences between the two experimental groups. In analyzing the possible effects of feedback on educator performance, experimental groups were analyzed on the adjusted posttest mean scores using the pretest means as the covariate. This analysis was conducted using measurements from student and peer assessor groups. If the equilibrium theory was operating, differences in the changes of educator behavior between the two experimental groups should have been observed.

Analysis of covariance (ANCOVA) results appear in Tables 23 through 56. In the first analyses, presented in Tables 23 through 39, adjusted posttest experimental and control group mean scores were analyzed as perceived by student assessors. The analysis of covariance F values presented in these tables provide a test of the differences between the two groups.

The adjusted posttest means between the two groups did not statistically differ significantly, as hypothesized, at the .05 level.

In the second analysis, presented in Tables 40 through 56, adjusted posttest experimental and control group mean scores were analyzed as perceived by peer assessors. The anlysis of covariance F values presented in these tables provide a test of the differences between the two groups at the .05 level of significance.

The adjusted posttest means, between the two groups, did not statistically differ significantly, as hypothesized. Therefore, as a consequence of the F values found, there is insufficient evidence to reject the null hypotheses.

The results of this analysis imply that neither of the two multi-assessor groups perceived statistically significant differences between experimental group's performance behavior from the mid-semester to the end-of-semester assessments. The differences between the adjusted posttest means of the two groups reveal that the experimental group had higher means than the control group on eleven and eight of the 17 means, as perceived by student and eight as perceived by peer assessors, respectively.

A point of interest, regarding pretest means, was that the experimental group means were higher than the control group. These higher pretest means lessened the possibility of significant differences because of the "ceiling effect." By and large, observation of pretest and posttest means were high on the 5-point assessment scale. The overall means were very close to 4.0 in both cases. Thus, lessening the chance of behavior modification to be assessed as a measure of change.

The findings of this investigation tend to agree with some of the evidence found by Centra (10). He summarized the findings of his study by pointing out that end-of-semester ratings of instructors who were given mid-semester feedback did not differ from either the no-feedback or the posttest groups. Moreover, the results parallel a study by Miller (32). His study reported that end-of-semester student ratings for teaching assistants who had received mid-semester feedback did not differ from end-of-semester ratings for teaching assistants who did not receive the feedback.

This investigator observed a positive attitude toward behavior modification within members of the experimental group when the educators received feedback not otherwise available to them. This attitude was observed from pretest to posttest time period. Such an observation leads one to surmise that when feedback is couched in a confidential manner, educators desire to know the assessment of their performance.

Analysis of Correlation Between Groups

The results of this investigation indicate that there was close agreement between student and peer group assessment on pretest measures. This agreement was not as significant on posttest measures. Consequently, these results are not in complete accord with those who contend that little agreement exists.

Analysis of Variance Among Educators

Using the Menne and Tolsma (31) adaptation of the analysis of variance F test, a minimum criterion of 21% of total means square due to between group variance appears to provide a valid cutoff point for use when large groups of students assess educators. From a practical standpoint, a much smaller group of peers are available to assess the educators. Therefore, the theoretical limit of 85% variance due to between group means square, based on three assessors, appears to provide a reasonable minimum criteria for identifying instrument variables which discriminate when used by a small group of peers to assess educator performance.

The results of this analysis demonstrate the power to which the items yielded values appropriate for measuring the specific performance behaviors of the educators in the experimental groups. The results of the student assessors pretest and posttest item analysis F values indicate that discrimination was made between educators. Such was not the case for peer assessor on either the pretest or posttest.

Recommendations for Further Study

The experience of conducting this investigation manifest a need for the following areas to be studied:

1. An investigation replicating this investigation in other colleges and universities.
2. Investigate the influence of factors such as age, sex, years of teaching experience, and professoral levels on educator performance modification.
3. Investigate the effect of educator self-assessment as a possible source of information on educator performance modification.
4. Investigate the effects of multi-assessor feedback as treatment in repeated sessions administering indepth discussion on possible ways of improving and modifying performance behavior in the

direction of assessor assessment.

5. Investigate the educator performance modification, following treatment over one and two full semesters.

6. Investigate independently the effects of student group assessors and peer group assessors feedback on modifying educator performance.

CHAPTER VI. SUMMARY

Purpose

The expressed purpose of this investigation was to study the effects and relationships of multi-assessor groups assessment as feedback on modifying educator performance.

Methodology

The instrument used in the collection of data was the Iowa State University 17-item Student Rating Instrument. This instrument was developed under the direction of Dr. John W. Menne (30).

Educators within the College for Human Resources Development were assigned randomly to a feedback (experimental) or no-feedback (control) group. The feedback and no-feedback groups used the 17-item assessment instrument in one of their classes at mid-semester during the fall 1974. Each member of the two groups requested an assessment from 3 peers. A summary of their students' and peers' responses were administered to each educator of the feedback group within one week, while results were withheld from the no-feedback group. Members of both groups used the assessment instruments in the same class and from the same peers at the end-of-semester.

Hypotheses Tested

The following general form of the hypotheses were tested.

Hypotheses 1. There are no significant differences between the experimental group and the control group on the adjusted posttest mean scores as perceived by student assessors as measured by the seventeen Educator Assessment Instrument variables.

Hypotheses 2. There are no significant differences between the experimental group and the control group on the adjusted posttest mean scores as perceived by peer assessors as measured by the seventeen Educator Assessment Instrument variables.

The findings of this investigation by analysis of covariance failed to yield F values which were statistically significant at the .05 level for any of the 34 specific hypotheses tested. Therefore, there was insufficient evidence to reject any of the 34 null hypotheses.

Research indicates that student assessments generally differ from peer assessments of the same educator. To examine this point of interest, correlation values between student and peer assessors, pretest mean responses indicated 14 of 17 variables were significant at the .05 level. On the student and peer assessors posttest, 5 of the 17 variables had r values significant at the .05 level. This study indicates that the low correlation between student and peer assessors, previously reported, may be due to the time of assessment.

Examination of the mean pretest scores revealed that the experimental group had higher pretest variable mean scores than the control group. However, when analysis of covariance was computed for each of the variables on (1) student assessment, and (2) peer assessment, no significant differences were found on the adjusted posttest means between the two groups.

Analysis of Measurement Accuracy

The analysis of variance technique used by Menne and Tolsma (31), was used in this study. This technique was used to determine item discrimination power for each of the seventeen items on the Educator Assessment

Instrument. The analysis of variance procedure was conducted using pretest and posttest student assessor means followed by pretest and posttest peer assessor means.

Menne and Tolsma (31) propose, "that the percentage of the total sum of squares (SS) due to "between groups" (i.e., between institutions, teachers, etc.) is an appropriate index of item discrimination."

The analysis of variance results for pretest and posttest student assessors indicated significant F ratios for 1 and 28 degrees of freedom at the .05 level of significance on all items. Eleven of 17 variables had F ratios that were highly significant at the .01 level on the student pretest. There were 10 of 17 variables with highly significant F values at .01 level. The results for pretest and postttest peer assessors indicated nonsignificant F ratios for 1 and 4 degrees of freedom at the .05 level of significance on all items.

Summary Statements

The findings of this investigation are condensed into the following summary statements:

1. There are no significant differences between the experimental group and the control group adjusted posttest mean scores as perceived by, (1) multi-assessor student group, and (2) multi-assessor peer group as measured by the Educator Assessment Instrument.
2. Correlation values for 50 paired student and peer, pretest mean responses indicated 14 of 17 variables were significant.

3. Correlation values for 50 paired student and peer posttest mean responses indicated 5 of 17 variables were significant.

4. The F values for student assessment pretest 17 variable means were significant for item discrimination analysis.

5. The F values for student assessment posttest 17 variable means were significant for item discrimination analysis.

6. The F values for peer assessment pretest 17 variable means were not significant for item discrimination.

7. The F value for peer assessment posttest 17 variable means were not significant for item discrimination.

The findings of this investigation indicate that multi-assessor feedback did not make its contribution such that educators modified their performance as perceived by either of the two multi-assessor groups.

Moreover, the finding reveal that there were significant r values between multi-assessor groups pretest and posttest variable mean responses. This implies that when the assessors saw high assessment on the pretest, they also saw high assessment on the posttest.

Of additional interest, was the significant F values resulting from analysis of variance of student pretest and posttest mean scores, disclosing that 17 variables discriminated among educators.

CHAPTER VI. BIBLIOGRAPHY

1. Aleamoni, L. M. The evaluation of instruction and the use of CEQ form 73. Research Memorandum No. 152. Urbana, Illinois: Office of Instructional Resources, University of Illinois, March 1974.

2. Aleamoni, Lawrence M. Evaluation by students to identify general instructional problems. Educational Resources Information Center. Bethesda, Md.: ERIC Document Reproduction Service, ED 076 166, 1973.

3. Aleamoni, Lawrence M. Being in love. Psychological Reports 31 (1972): 607-14.

4. Aleamoni, L. M., and Hexner, P. Z. The effect of different sets of instructions on student course and instructor evaluations. Research Report No. 339. Urbana, Ill.: Office of Instructional Resources, University of Illinois, May 1973.

5. Aleamoni, L. M., and Spencer, R. E. The Illinois course evaluation questionnaire: A description of its development and a report of some of its results. Educational and Psychological Measurement 33 (1973): 669-84.

6. Aleamoni, L. M., and Yimer, M. An investigation of the relationship between colleague rating, student rating, research productivity, and academic rank in rating instructional effectiveness. Journal of Educational Psychology 64 (1973): 274-77.

7. Biddle, Bruce J., and Ellena, William J., eds. Contemporary research on teacher effectiveness. New York: Holt, Rinehart, and Winston, 1964.

8. Blum, M. L. An investigation of the relation existing between students' grades and their rating of instructors' ability to teach. Journal of Educational Psychology 27 (1936): 217-21.

9. Brieter, Joan, and Menne, John W. Measuring teacher performance. Research paper, Dept. of Psychology, Iowa State University, 1974.

10. Centra, John A. Two studies on the utility of student ratings for improving teaching; I. The effectiveness of student feedback in modifying college instruction. II. Self-ratings of college teachers: A comparison with student ratings. Report No. 2. Princeton, N. J.: Educational Testing Service, 1973.

11. Costin, F., Greenough, W. T., and Menges, R. T. Student ratings of college teaching: Reliability, validity and usefulness. Review of Educational Research 41 (December 1971): 511-35.

12. Daw, Robert W., and Gage, N. L. Effect of feedback from teachers of principal. Journal of Educational Psychology 58 (1967): 181-88.

13. Festinger, Leon. A theory of cognitive dissonance. Stanford, California: Stanford University Press, 1962.

14. Frey, Peter W. The ongoing debate: Student evaluation of teaching. Change 6 (February 1974): 47-48, 64.

15. Frey, Peter W. Student ratings of teaching: Validity of several rating factors. Science 182 (1973): 83-85.

16. Gage, N. L., Runkel, P. J., and Chatterjee, B. B. Changing teacher behavior through feedback from pupils: An application of equilibrium theory. In. W. W. Charters and N. L. Gage, eds. Readings in the social psychology of education. Boston: Allyn and Bacon, 1963.

17. Gillmore, G. M. and Brandenburg, D. C. Would the proportion of students taking a class as a requirement affect student rating of the course. Educational Resources Information Center. Bethesda, Md.: ERIC Document Reproduction Service, ED 089 628, 1974.

18. Glass, Gene V. and Stanley, Julian C. Statistical methods in education and psychology. Englewood Cliffs, New Jersey: Prentice-Hall, Inc., 1970.

19. Good, Carter V. Dictionary of education. New York: McGraw-Hill, Inc., 1959.

20. Greenwood, Gordon E., Bridges, Charles M., Ware, William B., and McLean, James E. Student evaluation of college teaching behaviors (SECTB). Journal of Educational Measurement 11 (Summer 1974): 141-43.

21. Hidlebaugh, Everett J. A model for developing a teacher performance evaluation system: A multi-appraiser approach. Ph.D. dissertation, Iowa State University, 1973.

22. Hollander, E. P. The friendship factor in peer nomination. Personnel Psychology 9 (Winter 1956): 435-47.

23. Howsam, Robert B. Teacher evaluation: Facts and folklore. National Elementary School Principal 43 (1963): 2-18.

24. Isaacson, Robert L., McKeachie, Wilbert J., Milholland, John E., Lin, Yi G., Hofeller, Margaret, Baerwaldt, James W., and Zinn, Karl L. Dimensions of student evaluations of teaching. Journal of Educational Psychology (1964): 344-51.

25. Jarrett, James L. The self-evaluating college teacher. Today's Education 58 (January 1969): 40-41.

26. Jenkins, Joseph R., and Bausell, R. Barker. How teachers view the effective teacher: Student learning is not the top criterion. Phi Delta Kappan 55 (April 1974): 572-73.

27. Kartz, H. E. Characteristics of the best teachers are recognized by children. Pedagogical Seminary, 1896, pp. 413-18.

28. McCarter, Ronald W. Making the most of subjectivity in faculty evaluation. American Vocational Journal 49 (January 1974): 32-33.

29. McKeachie, Wilbert J. Student rating of faculty. American Association of University Professors Bulletin 35 (Winter 1969): 439-44.

30. Menne, John W. Teacher evaluation: Performance or effectiveness? Paper distributed at teacher evaluation conference at Iowa State University, 26-27 November 1972. (Mimeographed.)

31. Menne, J. W., and Tolsma, R. T. A discrimination index for items in instruments using group responses. Journal of Educational Measurement 8 (1971): 5-7.

32. Miller, Martin T. Instruction attitudes toward and their use of student ratings of teachers. Journal of Educational Psychology 62 (1971): 235-39.

33. Miller, Richard I. Evaluating faculty performance. Washington: Jossey-Bass Inc., 1972.

34. Oles, Henry J. Stability of student evaluations of instructors and their courses. Educational Resources Information Center, Bethesda, Md.: ERIC Document Reproduction Service, ED 091 409, 1973.

35. Rodin, Miram J. Can students evaluate good teaching? Change 5 (Summer 1973): 66-67, 80.

36. Ryans, David G. Assessment of teacher behavior and instruction. Review of Educational Research 33 (October 1963): 415-41.

37. Simpson, Ray H. Teacher self evaluation. New York: The Macmillan Company, 1966.

38. Simpson, Ray H., and Seidman, J. M. Use of teacher self-evaluative tools for the improvement of instruction. Improvement of instruction in higher education. Report to the American Association of Colleges for Teacher Education, 1962.

39. Smart, R. C. The evaluation of teaching performance from the point of view of the teaching profession. Chicago: American Psychological Association Meeting, 1965.

40. Snedecor, George W. and Cochran, William G. Statistical methods. Ames, Iowa: Iowa State University Press, 1967.

41. Spencer, R. E., and Aleamoni, L. M. A student course evaluation questionnaire. Journal of Educational Measurement 7 (1970): 209-210.

42. Stanley, J. C., and Weily, D. E. Development and analysis of experimental designs for ratings. Washington, D.C.: Cooperative Research Report No. 789, United States Office of Education, 1962.

43. Stimart, Reynold P., and Taylor, Alton L. Predicting excellence in university educators: A Vector algebra approach. Journal of Experimental Education 42 (Fall 1973): 74-76.

44. Swanson, Richard A., and Sisson, David J. The development, evaluation, and utilization of a department faculty appraisal system. Journal of Industrial Teacher Education 9 (January 1971): 64-79.

45. Vielhaber, D. P., and Gottheil, E. First impressions and subsequent ratings of performance. Psychological Reports 17 (December 1965): 916.

46. Voeks, Virginia W., and French, Grace M. Are student ratings of teachers affected by grades? Journal of Higher Education 31 (June 1960): 330-34.

47. Webb, Welse B. The problem of obtaining negative nominations in peer ratings. Personnel Psychology 8 (Spring 1955): 61-63.

48. Wert, James E., Neidt, Charles O., and Ahmann, J. Stanley, Statistical methods in educational and psychological research. New York: Appleton-Century-Crafts, Inc., 1954.

49. Williams, John D. Regression analysis in educational research. New York, N.Y.: MSS Information Corporation, 1974.

50. Zelenak, Mel J., and Snider, Bill C. Teachers don't resent evaluation--if it's for the improvement of instruction. Phi Delta Kappan 55 (April 1974): 570-71.

ACKNOWLEDGMENTS

Even though the acknowledgments are not a part of this investigation, it would be ludicrous to complete this dissertation without them.

I am extremely grateful to my wife, Avis, and our two children Mary and Stuart. Pursuit of this degree, unquestionably, would have been impossible without their everlasting love, help, and patience.

I am also most appreciative for the kindness and encouragement extended to me by my friends, relatives, and especially my mother.

Dr. Trevor Howe, Dr. Milton Brown, Dr. Ross Engel, Dr. George Kiser, Dr. John Menne, Dr. Rex Thomas, Dr. Richard Warren of Iowa State University, and Dr. Russell Peterson, of the University of North Dakota, are deserving of special recognition for their unrelenting efforts, faith, and considerations on my behalf.

APPENDIX A

EDUCATOR ASSESSMENT INSTRUMENT

Please assess the educator on the characteristics listed below in order to provide feedback which will enable him or her to improve their performance.

Instructions:

A) On the optical scan sheet place the full name of the educator.
B) Do not enter your name.
C) Use the number 2 pencil provided to make your response; do not use any other marking device.
D) Omit an assessment of an educator characteristic if you feel it would be inappropriate or that you do not have sufficient information to assess fairly.
E) Do not use the identification block on the optical scan sheet; start with item 1.

Please use the following five point scale to assess the educator performance. The assessment should indicate how this educator compares with all other educators you know in the College for Human Resources Development.

1/A Far Below Average (among the lowest 10%).
2/B Below Average (among the next 20%).
3/C Average (among the middle 40%).
4/D Above Average (among the next 20%).
5/E Far Above Average (among the top 10%).

Educator performance characteristics

1. Organization/Planning; organized and planned the course well.
2. Class Time Efficiency; used class time efficiently.
3. Preparedness; was well prepared for class.
4. Interest; was interested and enthusiastic about teaching this class.
5. Oral Presentation; spoke loudly enough and enunciated clearly.
6. Written Presentation; presented written material and blackboard work that was clearly legible.
7. Explanations; explained material clearly.
8. Relevance; showed the relevance of material.
9. Respect; showed respect for students.
10. Tolerance; was tolerant of weak students and differing opinions.
11. Fairness; was fair with students.
12. Availability; tried to be sufficiently available to students outside class.
13. Expectations; matched the level of the material to the ability of the class.

14. Amount of work; made sufficient and reasonable but not excessive assignments.

15. Relevance of work; made assignments which help in learning appropriate material.

16. Evaluation; presented clear, fair, and appropriate evaluation procedures for assessing performance.

17. Overall rating; compared to all other educators.

APPENDIX B

memorandum

TO: Faculty, College of Human Resources

DATE: October 18, 1974

FROM: Luvern R. Eickhoff Lre

RE: Research

The purpose of this communication is to inform you of the procedures in administering the instrument being used for my dissertation research.

The enclosed material deals with the collection of data for my research which Dean Tomasek has mentioned to you and which I described briefly at your faculty meeting.

The major purpose of the study is to investigate the effect of multi-assessor feedback on modifying educator performance. The study will hopefully cast some light on whether there is a difference between student and peer assessment; and whether assessment leads to modification in performance.

The research scheme is as follows:
1. The faculty will administer a brief assessment instrument in one of their classes and request that three colleaques (peers) who have some knowledge of their performance conduct the same assessment. The assessment will require less than ten minutes. Half of this group will receive an analysis of student and peer responses within a few days; the other half will not receive these results until after the end of the semester.

2. At the end of the semester I would like the same faculty to repeat the same procedures.
Thus the faculty will administer the instrument twice, this week and during the last week of the semester, and I will compare end-of-semester responses for those who received their results immediately vs. those who did not.

3. There are 30 copies of the instrument and optical scan response forms in a packet. These materials should be used in <u>one</u> of your classes - of your own choosing - this coming week. Should you need additional copies, call my office, 2249. After the instrument has been administered, put them back in the envelope (which has my name on it), seal it (to maintain confidentiality), and put it in inter campus mail. They should be returned no later than Friday.

4. The peer assessment should be returned by the assessor in the individual envelopes.

5. The optical scan forms will be processed, summarized, and returned in the manner previously stated. <u>Only you</u> will receive this analysis of assessor responses.

Thank you for your cooperation. You will receive the end-of-semester data collection materials around December 2.

memorandum

TO: Faculty, College of Human Resources Development DATE: December 2, 1974

FROM: Luvern R. Eickhoff Lre

RE: Research

To complete the second and final phase of my research, you will find enclosed optical scan sheets and instruments.

The instrument should be administered in the same class in which you administered the first phase. Also, you should request that the same peers respond to the instrument as previously performed. This procedure is imperative. I have consequently enclosed the same number of instruments and scan sheets as was required in the first phase (plus a couple extra). You should administer the assessment during your last class period of this semester. After they have been administered, put them in the envelope and seal it to maintain confidentiality. Return the envelope to your secretary.

Thank you for your assistance. I hope you will find your participation worthwhile.