TEACHER EVALUATION AS A FUNCTION OF LEADERSHIP STYLE:
A MULTIPLE CORRELATIONAL APPROACH

DISSERTATION

Presented to the Graduate Council of the North Texas State University in Partial Fulfillment of the Requirements

For the Degree of

DOCTOR OF PHILOSOPHY

By

Ronald G. Swanson, B.S., M.S.
Denton, Texas
May, 1974
Swanson, Ronald G., Teacher Evaluation as a Function of Leadership Style: A Multiple Correlational Approach. Doctor of Philosophy (Educational Psychology), May, 1974, 156 pp., 14 tables, bibliography, 165 titles.

One of the most persistent issues in contemporary organizations has been how to evaluate individual performance. Basically, the problem is who should evaluate whom and against what productivity criterion. Educational institutions have been the organizations most concerned with this dilemma in recent years. As recently as September, 1973, teachers went on strike over accountability procedures.

This study was conducted to identify which mode of teacher evaluation was most efficient, based on fairly objective performance criterion, and to establish a basis for viewing teaching style as leadership style. In existing research, superior ratings were the most used evaluation measure, student ratings were a rapidly growing mode of evaluation, self-ratings were considered biased, and peer ratings were used very little. Hence, who should do the evaluating was an unsolved problem. All four evaluation modes were employed in the present study for comparison.

A number of current teacher evaluation questionnaires were examined to ascertain what educators felt constituted good teaching. Two major factors emerged: teacher-centered and student-centered behaviors. These factors
corresponded to the two orthogonal dimensions of the Supervisory Behavior Description (SBD) along which leader behavior varies: Consideration and Structure. Every teacher evaluation questionnaire item was parallel to one or more SBD items. Since educators essentially had been after description of a teacher's leadership style, it appeared that a more efficient approach would be direct use of leadership data.

Most educational institutions do not have or maintain performance data. Therefore, an educational setting was selected which had relatively objective performance data and extensive records available--the U.S. Air Force pilot training system. Five different performance aspects were weighted and combined to form a multiple-input criterion which included an instructor's own performance as well as the performance of his students. The two SBD dimensions and a forced-ranking were used as three predictor variables. They were combined through the use of a multiple linear regression model.

The Ss were eighty military instructor pilots. Four separate evaluations of each were obtained: superior, student, self, and peer ratings. It was hypothesized that (1) there would be significant differences between the superior, student, self, and peer Rs, and (2) each predictor variable would contribute significantly to each R. No technique could be found for comparing related Rs, so a new statistical test was devised to test differences in
prediction error residuals rather than the Rs themselves. A .05 significance level was used for all tests.

The results indicated that peers ($R = .83$) had significantly less prediction error than all other groups. Superiors were next ($R = .52$), followed by self ($R = .43$), and students ($R = .23$). Hypothesis 1 was fully supported by peers, but only partially by the other groups. Hypothesis 2 was supported by forced-rankings, but only partially by the SBD dimensions.

It was concluded that the multiple correlational approach to evaluation, via peer rankings and ratings of leadership style, is inexpensive, available, better justified than any other evaluation technique encountered, and provides empirical data to an area previously characterized by conjecture and opinion.
# TABLE OF CONTENTS

LIST OF TABLES ........................................... iv

Chapter

I. INTRODUCTION ...................................... 1
   Statement of the Problem
   Purpose of the Study
   Hypotheses

II. REVIEW OF THE LITERATURE ......................... 4
   Introduction
   The Accountability Problem
   The Specific Evaluation Criteria Used
      in Modern Education
   The Teacher as a Leader
   Productivity

III. METHOD ........................................... 79
   Subjects
   Instrument
   Procedure
   Analysis of Data

IV. RESULTS ........................................... 100

V. DISCUSSION ........................................ 104

VI. SUMMARY AND CONCLUSIONS ....................... 123

APPENDIX ............................................ 133

BIBLIOGRAPHY ......................................... 143
<table>
<thead>
<tr>
<th>Table</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>I. Sample Items from Thomas' Questionnaire</td>
<td>59</td>
</tr>
<tr>
<td>with Corresponding SBD Dimensions and Parallel SBD Items</td>
<td></td>
</tr>
<tr>
<td>II. Sample Items from Hunter's Questionnaire</td>
<td>60</td>
</tr>
<tr>
<td>with Corresponding SBD Dimensions and Parallel SBD Items</td>
<td></td>
</tr>
<tr>
<td>III. Sample Items from Jones' Questionnaire</td>
<td>61</td>
</tr>
<tr>
<td>with Corresponding SBD Dimensions and Parallel SBD Items</td>
<td></td>
</tr>
<tr>
<td>IV. Sample Terms and Definitions from</td>
<td>62</td>
</tr>
<tr>
<td>Roberson's Questionnaire with Corresponding SBD Dimensions and Parallel SBD</td>
<td></td>
</tr>
<tr>
<td>Items</td>
<td></td>
</tr>
<tr>
<td>V. Categories from Gromish, et al. Questionnaire</td>
<td>63</td>
</tr>
<tr>
<td>with Corresponding SWD Dimensions and Parallel Items</td>
<td></td>
</tr>
<tr>
<td>VI. Factors from Watman's Questionnaire</td>
<td>64</td>
</tr>
<tr>
<td>Informal Questionnaire with Corresponding SBD Dimensions and Parallel SBD</td>
<td></td>
</tr>
<tr>
<td>VII. Sample Items from North Texas State Informal Questionnaire</td>
<td>65</td>
</tr>
<tr>
<td>with Corresponding SBD Dimensions and Parallel SBD Items</td>
<td></td>
</tr>
<tr>
<td>VIII. Weights Assigned Relative to Total Errors Assessed</td>
<td>92</td>
</tr>
<tr>
<td>IX. Sample Calculations of Highest and Lowest Overall Productivity Scores</td>
<td>94</td>
</tr>
<tr>
<td>X. Obtained Multiple Regression Equations and Multiple Correlation Coefficients for the Four Sets of Predictors</td>
<td>100</td>
</tr>
<tr>
<td>XI. Summary of Repeated Measures Analysis of Variance for the Differences Between Predicted and Obtained Productivity Scores</td>
<td>101</td>
</tr>
<tr>
<td>Table</td>
<td>Page</td>
</tr>
<tr>
<td>---------------</td>
<td>------</td>
</tr>
<tr>
<td>XII. Summary of Newman-Keuls Studentized Range Tests of Significance for the Mean Difference Scores</td>
<td>101</td>
</tr>
<tr>
<td>XIII. Step-Wise Multiple Regression Analysis for the Four Multiple Correlations</td>
<td>102</td>
</tr>
<tr>
<td>XIV. Simple Correlations</td>
<td>103</td>
</tr>
</tbody>
</table>
CHAPTER I

INTRODUCTION

One of the most pressing and persistent issues in contemporary education at all levels is the teacher evaluation dilemma. Accountability, in one form or another, has touched practically every educator, and the appraisal of teacher performance has become a vital issue for teachers' organizations, school administrators, and school boards.

Legislators and an economy-minded public are demanding that some sort of evaluation system be used; however, there has been no definitive work done on who can best evaluate teachers, or what constitutes productivity in an educational setting. In addition, it appears that the modern teacher is no longer thought of as a leader, but rather as a mere purveyor of knowledge.

A review of the recent literature on the teacher evaluation problem revealed that the disagreement and confusion is intensifying instead of subsiding. Opinions flourish with little or no objective data to support them. The preferred approach to teacher evaluation is difficult to determine.

Statement of the Problem

Though numerous articles and studies have been completed using every conceivable mode of teacher evaluation
(administrative ratings, student ratings, self-ratings, and peer ratings), few have arrived at or used any objective productivity measure as a criterion. The major difficulty in this area is that fairly objective measures of productivity are unavailable in most educational settings. Without such a criterion, it is impossible to reach any plausible conclusion about which mode of teacher evaluation is most efficient.

In addition, the teacher's style of teaching seems to have been the target of many evaluation methods even though it has been admitted that style of teaching is a nebulous and difficult to measure attribute. Commonly used teacher evaluation forms involve numerous dimensions with no clear underlying idea of what the construct being measured represents.

Purpose of the Study

The purpose of the study is to identify which mode of teacher evaluation is most efficient when a fairly objective productivity criterion is obtained, and to establish a basis for examining the teacher's teaching style as style of leadership. Since current research indicates that all four modes of teacher evaluation (administrative ratings, student ratings, self-ratings, and peer ratings) are commonly used, all modes are retained for comparison in this study. Because the typical educational institution provides little relatively objective performance or productivity data for use as a
criterion, the present study was conducted in an educational setting that does provide such data—a military training program. The leadership style approach to teacher evaluation is incorporated by using leadership dimensions as inputs to each mode of evaluation.

Since there is little doubt that teacher evaluation is here to stay, the present study is a step toward reducing the confusion and lack of agreement about who should evaluate, and against what criterion. Due to the presence of four groups of multiple predictors and a measurable criterion, it was decided that a multiple linear regression approach to the problem would best serve the purpose of the study.

Hypotheses

Using two leadership dimensions and an overall ranking as predictors, and a composite teacher/student productivity measure as the criterion, the following hypotheses were tested:

1. It was hypothesized that there would be significant differences between administrator, student, self, and peer multiple correlation coefficients used to predict productivity when leadership style was an input to the prediction equations.

2. It was hypothesized that each input variable to the regression equations would contribute significantly to their respective multiple correlation coefficients.
CHAPTER II

REVIEW OF THE LITERATURE

Introduction

The review of the literature for this study could occupy many volumes of text, but thorough research revealed that the evaluation problems of today were also the evaluation problems of yesteryear. For this reason, the most recent three-year period is examined meticulously for pertinent information. This approach seemed logical when it appeared that in recent years teacher evaluation has become a truly cogent issue in education. Studies from other time periods are included when they are precursors for or supplements to more recent works.

Many of the sources used are cited primarily to show how little empirical data has been gathered in relation to the large amounts of verbiage, opinion, and unsupported conjecture. Since much of the literature review is of this nature, criticisms pertaining to the present study are rendered as the review proceeds.

In order to portray the current state of the teacher evaluation problem, and to bring in the leadership aspects of teachers, it is necessary to review the literature in several specific areas simultaneously. These areas are (a) the accountability problem, (b) the specific evaluation
criteria used in modern education, (c) the teacher as a leader, and (d) some approaches used in measuring productivity.

The Accountability Problem

According to Callahan (5), the emphasis on efficiency and accountability in education stemmed from four major causes: (a) the reform movement of 1910, which stimulated muckraking journalists to expose the inefficiency of all public institutions, including schools; (b) the emergence of scientific management in business, exemplified by the efficiency expert and time-and-motion studies; (c) the growing prestige of businessmen, which initiated the trend of applying business methods to education, a trend reinforced by the election of businessmen to school boards and boards of regents; and (d) the emergence of a school-management profession made up of administrators who were at the mercy of the school boards and who grasped desperately for a way to win respect or at least survive.

School administrators and teachers were vulnerable to public opinion and to all kinds of pressure from school board members and citizens. The demand for more efficiency (lower costs) became a fixture of American education. Callahan suggested that the cult of efficiency had led to a vicious circle; the demand for more economy had pushed men who thought of school as a business into administrative positions where pupils were considered raw
materials to be converted into products. These men bowed to public pressure in order to keep their jobs. Over time, such leadership produced poorly educated citizens who were ignorant of many aspects of schooling and who demanded economy in the schools. Their ignorance plus their concern about high taxes sustained the pressure on school administrators and resulted in more efficiency-minded people being put into office.

Popham (87) pointed out that citizens were not elated with their perceptions of the quality of education. They wanted dramatic improvements in the schools and unless they got them, there was real doubt as to whether much increased financial support for educational endeavors could be expected. Popham said,

The public is in no mood to be assauged by promises. Deliver the results, we are being told. No longer will lofty language suffice, and yesteryear's assurances that "only we professionals know what we're doing" must seem laughable to today's informed layman (87, p. 27).

Whether the details on the evaluation of accountability in education were accurate or not was not really important. What was important was that accountability had become an accepted fact by many educational leaders, and monetary-economic factors were the reason. Young (115) concluded that the age of accountability was here for all educators. Effectiveness and efficiency of education would increasingly be in demand as costs rose. Direct competition from private industry was threatening traditional educators in terms of performance/accountability.
House (52) pointed out that as much as 10% of a school's budget may be set aside for evaluation purposes, while Popham (87) found that California legislators had very recently enacted a teacher evaluation law requiring each teacher in the state to be evaluated by locally devised teacher appraisal systems. These local evaluations were required to include certain state-stipulated elements.

Kossoff (62) saw bleakness in the future of education at all levels due to the demand by financial backers of education for a turn from the search for new knowledge to the problems of communicating the present store of knowledge to the population at large. Similarly, McNally (72) postulated that accountability was increasingly being asked of educators to account for their stewardship.

Egnatoff put it most bluntly when he said, "It is not a question of whether teacher performance should be appraised. It is rather a question of for what purpose? By whom? How? When?" (15, p. 19). Wilson (112) agreed that educational administrators could no longer side-step the issue of faculty evaluation.

Thomas (104) proposed that the evaluation of teachers was an inevitable task of the educational administrator for three reasons: (a) the public employed professionally trained people to run its schools and felt entitled to demand an accounting of their efforts; (b) teaching had to be evaluated for professional improvement; and (c) evaluations
were used privately to determine salary and promotion of the
teacher. Pulley (88) fully agreed with Thomas and added that
retention, transfer, demotion, and dismissal had also been
determined by evaluation.

Fantini (18) acknowledged accountability as a result of
overtaxed citizens rebelling against school costs, demanding
the most from their educational dollar, and questioning the
school's leadership patterns. It appears that the age of
accountability is upon us, and as Esler (17) implied, the
effects of accountability upon the schools is likely to be
with us for some time.

Having established the background and reasons for teacher
evaluation, the next logical step is to examine the myriad
difficulties inherent in the accountability procedures that
are being used. Faculty appraisal is a subject capable of
producing violent disagreements and much rhetoric in a short
period of time. A review of recent literature on the sub-
ject produced few solid conclusions except that there is
confusion and dissension among those concerned.

Popham (86) found that since the turn of the century,
how to measure a teacher's instructional skill had perplexed
educational researchers and evaluators. He emphasized that
"The most widely used measures--ratings, classroom observa-
tions, and pupil performance on standardized tests--all had
proved dismally inadequate" (86, p. 59). Most techniques
failed to take into account the fact that different teachers
pursued very different goals.
Accountability had as many definitions as it had definers according to Selden (94). Education was becoming more expensive, and yet there was no way to show that the public was getting more for their money. He advocated shared policymaking and seriously doubted the value of accountability per se.

Some of the confusion generated by the accountability conflict stemmed from the fact that most people believed that identification of good teaching was a simple matter (104). Communities, students, and others were more than prepared to render judgments regarding the competence of teachers. All one had to do was ask them.

McNally (72) observed that the question as to how to evaluate in education was still unanswered—particularly where teacher evaluation was concerned. The major problem as he saw it was that educators had not been able to get together on what should be evaluated or why. McNally conjectured that most efforts at teacher evaluation were too narrowly focused and oversimplified. He divided the purposes of teacher evaluation into two categories: administrative and instructional. The administrative aspect was undertaken to decide, for example, whether or not a teacher should be employed, transferred, tenured, promoted, or given a raise. These so-called merit plans had almost always turned out to be unsatisfactory. The instructional aspect was to be for professional improvement of weak areas.
McNally made one important point which concerned the widespread acceptance of evaluation and accountability for teachers, but made little or no mention of such measures for educational administrators. While this was a cogent point worthy of further consideration, the remainder of the article contained no objective recommendations on how to improve evaluations.

Watman (107) recommended that when constructing a school evaluation system, an attempt should be made to insure that the system was basically a helping one for teachers, and that the alarm and fear of teachers toward evaluation systems was largely unnecessary. It is almost paradoxical, however, that the same author a few lines later acknowledged the use of evaluations for promoting, transferring, retaining, and terminating teacher contracts.

The remainder of Watman's article consisted of an elaborate plan for teacher evaluation which included the development of a school philosophy, classroom visitations by three different supervisors, follow-up conferences with written reports, teacher written replies to the reports, and the right of the teacher to appeal to higher authority in case of unfairness or personal bias. This proposed system is very complicated, expensive, and time-consuming for all individuals involved, and is nothing more than a system of subjective opinions.
Jones (55) maintained that dissatisfaction with existing procedures for the evaluation of teachers was widespread. His research indicated that most evaluation programs were so superficial that neither administration nor faculty gave much credence to the whole process. Most teachers saw the evaluative process as documentary evidence of a teacher's less favorable qualities, and not much else. The use of evaluation forms as dismissal and demotion devices had practically done away with teacher evaluations as a positive force in improving instruction.

Jones' conclusions were (a) there was little agreement in all the research as to what an effective teacher was; (b) evaluations pointed out individual teacher deficiencies, but provided no clear path for self-improvement; and (c) evaluation forms were vague and irrelevant. Being told that one had deficiencies did not necessarily mean that one believed it or knew what to do about it. The research conducted in this study was adequate, but the recommendations for evaluation improvement were weak. For example, rating by principals, the teachers themselves, and by students were advocated to be combined into an amalgamation of ratings. This proposed program is time consuming, expensive, and purely subjective. In addition, recommendations were made that were vague and added little to objectiveness; i.e., (a) evaluation items should speak to teacher characteristics desirable in a particular system at a particular
time, and (b) vagueness and irrelevancy in items should be totally eliminated.

Esler (17) had conclusions similar to several previous works which insisted that teachers must be made to view accountability not as something restrictive and fearful, but as an opportunity to objectively assess their performance and improve their level of instruction. Once again, the nemesis of using evaluations for aversive reasons was acknowledged, but not seen as serious. It is difficult to see how any given teacher can resolve this paradox.

It was a rare circumstance to find a school utilizing teacher evaluation procedures that were not in need of revision according to Kult (63). The most common oppositions from teachers concerning evaluations in this study were (a) the large majority of persons observing and evaluating teachers were lacking in proportional academic competencies to be considered credible evaluators; (b) the forms used by evaluators were seen as highly subjective and entirely useless for teacher improvement purposes; (c) teachers felt unduly discriminated against since they comprised the only segment of education that was constantly evaluated and found wanting; and (d) teachers felt that a formal evaluation should be administered to all levels of educational strata. Administrators commonly exempted themselves from such procedures.

Kult also made some recommendations for the improvement of evaluatory processes such as the initiation of a cooperative school evaluation exchange program via video tape, and
incorporated evaluations of teacher to teacher, administrator to administrator, teacher to administrator, administrator to teacher, students to teacher, and students to administrator. Besides the time and money that such a program would involve, Kult added another device which was to develop teacher/administrator improvement through the writing of units, reports on innovations, data gathering in one's field, and others. Even if such a costly endeavor were possible, it would still be left with no objective performance data as a basis.

Swanson and Weaver (100) observed that no definitive study describing faculty appraisal systems had been cited by the American Association of University Professors. When these authors queried a number of institutions about evaluation practices, a large number of respondents expressed dissatisfaction with the current state of the art and requested copies of any usable findings. While there was some agreement as to which elements of faculty performance should be appraised, there was little similarity among methods used. For example, of the institutions responding, the methods used included student opinions, years of teaching experience, research and publications, outside academic work, professional recognition, community activity, attitudes, appraisal by department chairmen, peer ratings, tenure, self-evaluations, and a number of others.
Other researchers (33) found that committee work, student advisement, and service were scrutinized along with teaching ability for evaluative purposes. Their major question was whether to use self-evaluations, administrative observances, student ratings, classroom visitations, or a combination of all these. Their results showed that no objective criteria had been established to test which ratings reflected good teaching. They found that students and the chairmen of departments did not rate faculty members similarly. This indicated a dire need to define evaluation criteria better, and to devise more precise methods by which to measure them. These authors did not, however, have any ideas about how this could be accomplished.

Dyer (13) offered a system of evaluation which included performance indicators that combined various kinds of information into a set of numbers that could describe the performance of an educational system while avoiding the pitfalls that had characterized assessment in the past. The variables used were called input, educational process, surrounding conditions, and output. From the description of each of these variables, it is difficult to see how traditional pitfalls are going to be avoided, but it was suggested that no single index of performance could adequately characterize any teacher or system. The proposed system was of such complexity that the author admitted that it was "unlikely to get off the ground until 1976" (13, p. 105).
A number of studies defined and outlined the problems of evaluation nicely, but had little of substance for recommending improvement. Kaufman (57) suggested that a generic systems approach to evaluation problem-solving in education might bring about an integration of accountability with the quantitative improvement of instruction. Here education was viewed as the identification and resolution of problems, e.g., what was to be accomplished and how it was to be accomplished. It was suggested that this generic approach would assist educational practitioners in improving the products and processes of education, and that it would aid them in achieving accountability.

Similarly, Hickcox and Scott (45) offered a four-aspect conference evaluation plan which included a post-conference approach, a qualified observer approach, a measurement of objectives approach, and a planning committee approach. The authors suggested that such evaluation efforts should attempt to avoid halo effect, should take objectives into consideration, and should provide for subsequent analysis. Such rhetoric offered very little in the way of real solutions to the accountability problem.

A few valid points came out of current research, such as the evidence that no one satisfactory way has been found for describing effectiveness in the classroom (82) and that it is extremely difficult or impossible to compare pupils, schools, or procedures across the country (74).
One interesting technique, which gained some measure of popularity, was the appraisal-by-results program (42), in which teachers were rated by superiors on the basis of predetermined objectives met within an agreed-upon period of time. This method supposedly emphasized performance rather than personality, but appeared rather narrow in scope.

Tumin (106) thought that many negative consequences could arise from introducing structured formal evaluation too early and too uniformly in an unstructured and variegated field like education. He believed that it was most important to treat evaluation as a transactive process. This belief was based on the results of a two-year effort at devising a model for measuring and comparing the effectiveness of educational systems which indicated significant, but instructive failure.

Perhaps one of the most disputed areas, in attempts at educational evaluation, has been the question of who should rate whom. There were many arguments for and against nearly every proposal in this area. For example, Owens (85) found that administrators, teachers, and college supervisors perceived many areas of teacher competence similarly, while Thomas (105) observed that teachers and administrators differed greatly on what they felt constituted good teaching performance. Egnatoff (15) concluded that although considerable progress had been made in developing interaction analysis models, more research was needed to
determine who should do the appraising: superiors, peers, subordinates, external observers, or the teacher himself. Research in this area was incomplete and inconclusive.

Stones and Morris, in a study of colleges of education and university departments of education, raised the question "How can you tell a good teacher when you see one?" (99, p. 110). Their results showed that bases of criteria used for the assessment of teachers varied enormously, and it was revealed that about half of the institutions sampled did not even have a formal list of criteria in written form. This indicated grave weaknesses in the present system of teaching assessment, particularly in the diversity of the methods of assessment.

House felt strongly that a teacher had little to gain from having his or her work examined. His thoughts on evaluation were "Since there are no punishments for not exposing one's behavior and many dangers in so doing, the prudent teacher gives lip-service to the idea and drags both feet" (52, p. 406).

According to Popham, "The accountability gang is beginning to fire some pretty potent pistols at the embattled teachers. One is labeled 'teacher tenure.' Another is called 'teacher evaluations.' A third simply says 'taxpayers revolt'" (87, p. 25). Since evaluations had often been used for promotions, salary adjustments, and termination, Popham saw the stakes as being very high and felt
that most teachers would do a better job for their pupils if they only knew how. His research showed evidence that for the purposes of measuring the results of instruction, standardized tests were usually inappropriate since they were designed to permit distinction between different learners, not teachers, and would typically lead to misleading information.

Kossoff (62) advocated a halt to the search for objective criteria, and wondered why educational researchers insisted on forcing qualitative traits into a quantitative mold. She wondered why questionnaires were after statistical rather than verbal data. She saw nothing wrong with mere descriptive profiles on teachers or evaluation of teachers by human observers. In her attack on quantitative methods she maintained that

A rating scale assumes a continuum divisible into discrete parts, but are all human qualities continuous or divisible? If a teacher collapses in front of a classroom, we could, I suppose, survey the five hundred students in the lecture hall to see if they believe he is "thoroughly dead," "moderately dead," "a little bit dead," or "slightly alive," but all we really need is the opinion of one doctor who can tell us whether to call for the ambulance or the mortician (62, p. 82).

The fault in her reasoning is that there are no "doctors" in educational evaluation who can render an ultimate opinion as to the validity of evaluation as the doctor in her example did.
The American Educational Research Association (2) reported that the condition of research on teacher effectiveness held little promise of results commensurate with the needs of American education. After forty years of research on teacher effectiveness, they cited few outcomes that a superintendent of schools would be able to employ in hiring a teacher or granting tenure. Remmers (90) assured readers that the statement required no amendment. Review of recent literature gave no reason to believe that much progress had occurred in over sixty years.

Watman (107) favored a new checklist system to aid teachers in objectively evaluating what went on in their classroom. The checklist contained headings such as physical conditions in the room (clean, neat, well ventilated), the teacher's classroom organization (materials present, activities organized), discipline or class control, long-range planning, pupil participation, and administrative assignments. While Watman may have considered such evaluations new, they had a strange familiarity from the past.

What conclusions could be offered when the past and present research on teacher evaluations were compiled? Egnatoff (15) felt that researchers and evaluation systems developers were still just scratching the surface. It was apparent to Wilson (112) that other means and criteria were needed to make decisions about instructional personnel, and that most current systems had to be operated for two
years or more before useful data could be gathered. It was also apparent that faculty members could make or break an instructional program and that the extremely sensitive activity of faculty evaluation could obviously affect teacher morale, attitude, and effort. Since most evaluation changes had been made on the basis of hunch, bias, and personal convenience, the most frustrating aspect of teaching was seen as the lack of clearly stated and objectively defined and accepted criteria of performance.

Taylor, DeCorte, and Swinne (103) viewed the difficulty of establishing criteria for judging outcomes of instruction as a very important and urgent problem in educational evaluation. While they realized the weaknesses involved in subjective judgments, the authors still insisted on the use of pupil achievement standards, as defined by expert judges, for evaluation criteria of instructional effectiveness.

Mueller (79) concluded that a program for testing the performance of teachers was desperately needed, and suggested that the best way to evaluate performance was through thorough testing of student achievement. He saw the ultimate test of a teacher's ability in the achievement test scores of their students, and openly advocated that teachers teach the test.

Kellough (58) saw teacher evaluation by means of student scores as a circumvention of the real problem. He thought that all that was needed to solve educational problems was
increased time for the teacher to freely communicate with fewer numbers of students. Such changes would certainly help, but were they possible? Popham (87) thought that focusing on teaching performance tests assessed only one competency of a teacher—what of the others? His conclusion fell in line with the majority of other modern educators; while everyone had been talking about the merits of educational accountability, few had mentioned the fact that practical procedures for making accountability work had not been devised.

McNeil (73) thought current practices in evaluating teachers were shocking. He cited two national surveys, by the National Education Association (80, 81), which revealed that only about half the school systems in the United States followed any formal procedures in evaluating teachers, and that the procedures used were most inadequate. Though three out of four superintendents and principals expressed confidence in their programs of evaluation, over half of their teachers did not.

These two studies revealed a number of other findings pertinent to the evaluation conflict. The procedures used as a basis for evaluation were considered unfair; pupil achievement records did not always take into account the fact that a teacher had had a particularly unruly class or that he or she had written books, published articles, or conducted research. Reports from parents were entirely
too subjective, as were student and principal ratings. Classroom visitations were considered poor since they were rarely conducted on any regular schedule, were only done on about one-third of all teachers surveyed, and were followed by a conference with the observer less than one-half the time. One of the most mentioned criticisms of classroom visitation was that few teachers (34%) were observed for a period of five minutes or longer, yet subsequent rating forms stated in detail not only the rater’s estimate of the teacher’s competence, but also his or her personal qualifications.

McNeil also described the problem caused by discontent on the part of school board members. There were a number of explanations for the actions of school board members where teacher evaluation was concerned, but basically they all included the idea that board members wanted to be assured that the superintendent was differentiating among teachers so that the district would select and retain those teachers who could best help pupils attain the expectations of the community. They wanted administrators to identify differences among teachers so that monies for salaries (which made up about 80% of the system’s budget) could be wisely spent. College and university professors were also becoming increasingly involved in the evaluation struggle as administrators and boards of regents came under increased public pressure for efficiency and differentiation.
It was evident, from the preceding review of recent literature, that the issue of teacher accountability was wrought with disagreement, confusion, and a definite lack of objective criteria for performance. Having established the need for more substantive work in this area, it was considered essential that the next topic for review should be that of the specific evaluation criteria used in modern education.

The Specific Evaluation Criteria Used in Modern Education

Basically, there are four categories of criteria: administrative ratings, student ratings, self-ratings, and peer ratings. Few works had compiled a meaningful list of criteria used for identifying effective instruction, but those that did (14, 73, 78, 107) agreed closely on the categories involved. Since Morsh and Wilder's review (78) contained over 900 references which looked at the popularity and worth of certain criteria, and since it had the most representative list available, their outline is used as a basis for the following review.

Administrative Ratings

The first category examined was administrative or superior ratings which constituted the most widely used measure of instructional competence. For the most part, these ratings did not produce very high correlations with measures
of student gain, and were seen as being subject to profound halo effect.

Pulley (88) indicated that a majority of theorists and practitioners agreed that the school principal or department chairman had primary responsibility for teacher appraisal, but suggested that such should not be the case. He cited research which indicated that teachers preferred to be evaluated by a close superior, not someone several administrative steps removed from them in the hierarchy. He recommended that information flow freely between teachers and supervisors who were responsible for teacher appraisal, but did not offer any specific means by which this could be accomplished.

Thomas (104) studied principals' evaluation of teachers for classroom and administrative promotion and found that although a common body of criteria was used, principals placed emphasis on different aspects depending upon whether the teacher was being considered for classroom promotion or administrative promotion. It was also found that certain biographical, personality, experience, and age variables affected which criteria were emphasized.

Swanson and Sisson (101) found that the evaluation of university faculty personnel was, in most instances, the responsibility of the department chairman, but that the chairman's evaluations were at best unpredictable and in many cases without validity. While the chairman's rating
was seen as the best measure of performance, in one or more dimensions of faculty performance, it could not be considered the sole criterion. They believed that a rating system should consider all the dimensions of faculty performance, as well as who might best assess performance in the dimensions. A university professor's duties were seen as including teaching, scholarly productivity, and service. They used student ratings to measure teaching ability, faculty or peer ratings to measure scholarly productivity, and chairman's ratings to measure service.

The biggest weakness in this study is that mean rank scores were used as the index of faculty performance. No objective criteria were attempted or obtained, and the results established that the three groups rated faculty members on independent criterion. For example, peer and chairman ratings correlated .82, peer and student ratings .70, and student and chairman ratings .50. These were interesting results even though the number of subjects was quite small ($N = 12$).

In the course of their study, Swanson and Sisson discovered that some faculty members did not regard the evaluation of teaching performance as necessary. They met resistance to the use of some of their instruments, especially the student and peer ratings portions. They also reminded their readers that decisions on retention, promotion, tenure, and salary had, were, and would continue to
be made. The question was, would they be made via a single input or based on multiple inputs? Their study was original in many respects, but a serious weakness cropped up which the authors themselves had to acknowledge. It was conceivable that even though the faculty were ranked from highest to lowest in performance, they all could fall into one performance level (all "A's" even though the scale ran A, B, C, D, F). Such a procedure would be conducive to inflationary ratings to keep from inhibiting anyone's career or future.

An effort to probe the assumption that evaluation acted as a barrier to communication between teachers and supervisors was conducted by Lindemann (67). The findings indicated that teachers who perceived that they were subject to evaluation by their supervisors communicated more often with them than those teachers who perceived that they were not being evaluated by their supervisors. The more formal the evaluation, the more the trust, and the more frequent the teacher-supervisor interaction. Lindemann recommended that supervisors ought to evaluate teachers, and that supervision should be shared responsibility.

Masling and Stern (69) studied the effect of an observer on the classroom behavior of the teacher and students. Since superior ratings were often obtained via classroom visitation, this study seemed very relevant. The results revealed that the effects of the observer were extremely complex and affected various aspects of classroom behavior differentially.
The important point was that the presence of an observer did alter the natural classroom behavior.

McNally (72) reviewed a multitude of studies about teacher evaluation and found that there was no question that the principal or department chairman was the key person in the program. It was a rare school system where the superior was not the person most responsible for teacher evaluation even though the validity of superior ratings remained to be demonstrated.

The conclusions garnered from the research on evaluation by superiors (principals, superintendents, and department chairmen) left vagueness and confusion. Though widely used, superior evaluation had little substantial validity data to support it, and appears to be subject to considerable halo effect. There is doubt as to the knowledge of teacher activity possessed by superiors, and sparse evidence that enough time is put into the evaluations themselves. In spite of all this, teachers seem to prefer superior evaluation to peer or student evaluations even though none of these have been related by any substantial degree to objective measures of performance.

Student Ratings

The second category investigated was that of student ratings, which appeared to be growing rapidly. Such ratings had shown fair consistency, but were subject to considerable halo effect. Results indicated that if the instructor
favored brighter students, he would be approved by them. If
the weaker students were favored, the teacher would be dis-
approved by the brighter students (73).

The methods of developing and using student rating forms
varied considerably. Originally, student ratings were thought
to provide feedback which the instructor might not be able to
get from students on a face-to-face basis. They were also
seen as capable of setting departmental and college-wide norms
against which individual faculty ratings could be judged. The
most current aspect of student ratings appeared to be their
use as a source of information to aid other students in selec-
tion of courses.

Though student ratings are currently the most used form
of evaluation in certain areas (100), they at one time were
a declining method. Gustad (34) cited this decline and at-
tributed it to the lack of convincing validity data. However,
perceived threat to faculty might also have been an important
cause since a strong impetus to use student ratings had more
recently come from the students themselves (110).

Faculty members often claimed that student ratings were
unreliable, and that the ratings would favor an entertainer
over the instructor who got his material across effectively.
Students were also not considered competent judges of in-
struction since long-term benefits of a course may not have
been clear at the time it was rated.
Costin, Greenough, and Menges (10) conducted a massive review of the literature concerning student ratings of college teaching, and found that students could rate classroom instruction with a reasonable degree of reliability and were at least partially capable of distinguishing certain qualities of instruction which increased their knowledge or motivation. They found no significant relationship between years of experience and student ratings but did find evidence that feedback from student ratings might improve a teacher's performance (as evaluated by students).

They found mixed relationships between a teacher's research productivity and student ratings of their teaching effectiveness, and cited a long series of articles, dating from 1930 through 1968, which indicated that student ratings nearly always had reflected personality or charisma qualities such as warmth, humor, sincerity, helpfulness, agreeableness, emotional stability, fairness, enthusiasm, interest in students, and pleasing personality. Results were highly mixed on the relationship of student ratings and expected or actual grades received.

Other correlates of student ratings were majors tended to rate courses more highly than non-majors; students that were required to take a course often rated it lower than those for whom it was an elective; upper class and graduate students tended to give higher ratings than underclassmen.
Though it had often been claimed that teaching effectiveness and research productivity went hand in hand, the empirical findings indicated that, at best, the relationship was a weak one in the eyes of students.

Costin, Greenough, and Menges concluded that student ratings fell far short of a complete assessment of an instructor's teaching contribution, but could hardly be ignored as a systematic measure of teacher performance. The most damaging evidence against student ratings was that teaching effectiveness was positively correlated with a teacher's agreeableness, enthusiasm, humor, warmth, and personality. Though these qualities are desirable, they can hardly be considered an objective aspect of teacher productivity.

Holmes (50, 51) found that although differences in actual grades given did not affect student rating evaluations, if the students' grades disconfirmed their expectancies, the students tended to deprecate the instructor's teaching performance. Due to the differential effect of instructors on different types of students, it was clear that the grade/expectancy relationship did not distort the evaluation process extremely, and did not pose a serious threat to the evaluation system.

Levinson (66) strongly recommended that superiors be appraised by their subordinates in a variety of situations that ranged from business enterprises to education.
Jones (55) agreed with this and suggested that students were, on the average, more accurate in rating teachers than supervisors and peers. Jones, however, provided no empirical data to support this argument.

Field, Simpkins, Browne, and Rich (19) saw student evaluation of teaching as being an increasingly accepted part of higher education. Using the method of discriminant analysis, they found that students discriminated quite clearly between teachers in their perceptions, and that they could identify distinctive configurations of teaching behaviors. This suggested the possible shortcomings of considering separately the various elements and dimensions of teaching behavior. For example, the value of active student participation achieved proper significance only when seen in conjunction with skillful teaching, and not as an isolated element of instructional practice.

Watman (108) decided that while it was true that some students would not reply on an evaluation in an honest and serious manner, experience had shown that most would. He also concluded that the specific method used to obtain the student evaluations made a difference in the results. He did not make any recommendations as to which methods were best.

Thomas (105) contended that although students tended to rate teachers as being all good or all bad, depending on their feelings about the teacher, student ratings still
appeared to be a good measure of the actual performance of a teacher. Even in the face of highly diverse findings about student ratings, Thomas maintained that student ratings were highly reliable and stable. Reliability and stability are necessary, but certainly not adequate where validity is concerned.

Several very recent studies were excellent examples of experimental design and control, and had some interesting results. Coats, Swierenga, and Wickert (8) attempted to determine and number the nature of factors which accounted for student perceptions of teacher effectiveness. They factor-analyzed the popular Teacher Image Questionnaire and found that a single factor accounted for 61.5% of the variance in test items. The remaining 40% of the variance was evenly distributed among five other factors. The single large factor was identified as teacher charisma or popularity. Hence, students did not respond directly to specific questions regarding teacher effectiveness. Rather, a kind of halo effect based on teacher charisma or popularity determined to a large extent how students reacted to questions about their teacher.

Coats et al. did not imply that student ratings of teachers were not important or meaningful since teacher charisma was probably a function of teacher effectiveness. The evidence did show, however, that student ratings were more a measure of popularity than of actual teacher effectiveness, and should be used as one part of an evaluation
package consisting of measures on several additional variables such as peer ratings, superior ratings, and the like.

McDaniel and Ravitz (70) found that if one wanted to be liked as a teacher, one should teach students in their major area of interest. Students preferred professors who helped them meet their professional and personal objectives. Engineering students picked engineering professors, commerce students picked commerce professors, etc. While this seemed logical, they also found that students were alert to the amount of effort, care, and planning that entered into class preparation, but only to the extent that it directly benefited students.

McKeachie, Lin, and Mann (71) studied the validity of student ratings and found that their results, taken together with earlier studies, did not invalidate the use of student ratings as one source of evidence about teacher effectiveness, but were far less convincing than they had hoped. They concluded that student ratings had some usefulness, but wondered why they were not better. Their best guess was that the major slippage in validity studies was in the differing goals of teachers and students.

Probably the best controlled and most carefully designed study on student evaluations was conducted by Rodin and Rodin (92). This study assessed the validity of student evaluations by means of a comparison between objective and subjective criteria of good teaching. The objective criterion of
teacher effectiveness was based on what the students had learned from the teacher--i.e., initial ability in calculus followed by the instruction and then a post-course calculus test. The subjective criterion was based on student evaluations of teacher effectiveness. The results showed that the instructors with the lowest subjective scores received the highest objective scores and vice versa. In other words, students did not wish so much to maximize the amount learned as to reach an equitable compromise between effort involved in learning and the perceived importance of what was learned. Perhaps students resented instructors who forced them to work too hard and to learn more than they wished.

The major reason for defining good teaching in terms of good scores on student evaluation forms had been based on the analogy between the student and the consumer. The student, as primary consumer of the teaching product, was seen as being in the best position to evaluate its worth (92). The preponderance of data, however, indicated that students were less than perfect judges of teaching effectiveness. If how much students learn is considered to be a major component of good teaching, it must be concluded that good teaching is not validly measured by student evaluations as they now exist.

The conclusions garnered from all of the research on student ratings were mixed at best. There is no doubt that student ratings of teacher effectiveness are definitely on
the rise. In 1960, 40% of all institutions of higher learning and 67% of all high schools were asking students to evaluate their teachers. Though no figures were available for 1970, it seems safe to assume that there has been substantial increases in the number of student ratings used. The evidence cited indicates that student evaluations, to a large extent, tend to reflect the personal and social qualities of instructors, who he is rather than what he does (60).

**Self-Ratings**

The third criteria category examined was that of self-ratings. McNeil (73) noted that overall, there was some tendency for teachers to overrate themselves, and that self-ratings showed negligible relationships with administrative ratings, student ratings, and measures of student gain. Intuitively, self-ratings appeared very biased.

Roberson (91) found that the results of intensive research and observation of the teaching-learning process in classrooms indicated that no one particular style of teaching could be declared the model for everyone. Rather, the need appeared to be for self-appraisal or the opportunity for each teacher to find and develop their own effective style. The many observation systems developed in recent years had not been designed for teacher self-appraisal, and only very recently was any effort given to involving the teacher in the analysis of the teaching act.
Roberson advocated the development of a teacher self-appraisal system which utilized videotape recording equipment that would record classroom happenings and allow the instructor the opportunity to view him or herself in action. Additionally, a rating form would be used on which the teachers would evaluate themselves on three aspects: methods, objectives, and expressions. He maintained that these ratings could be validated by simply testing the students at the cognitive level planned and comparing their responses with observed teacher intentions. The resulting correlations would help the teachers determine just how effective their teaching was in terms of student learning. While much of Roberson's plan was interesting and possibly useful, it is certain that such a program would be very expensive and time consuming for already busy teachers.

Egnatoff (15) perceived teaching as a recognized profession and felt that teachers were becoming increasingly active in improving their own professional competence. On the other hand, Wilson (112) implied that, at best, teachers had vague and questionable criteria of performance, and consequently suffered from an inability to really know how well they were doing.

Seldon (94) cited evidence which prompted a conclusion that nothing good happened in schools unless teachers made it happen. Productivity would not be improved by piling on more oppressive supervision. Rather, he felt that teachers
should be given the responsibility for working out their own best methods of teaching by evaluating themselves.

Data on self-evaluations were rather sparse. This was probably due to the obvious weakness of the inflationary ratings often associated with self-appraisal. Swanson and Weaver (100) found that in schools of business, business administration, commerce, and economics, self-evaluations were used by only 17% of the institutions sampled, and then self-ratings were most often used only as one of several inputs in the evaluation system. Self-evaluations were not used alone, and apparently not with a great deal of confidence in any case.

**Peer Ratings**

The fourth criteria category investigated was peer ratings, which McNeil described as "little used" (73, p. 26). He concluded that in using peer opinions, ranks would probably yield better results than ratings, and that halo influenced peer ratings just as it did administrative ratings. Swanson and Weaver (100) supported these conclusions.

Costin, Greenough, and Menges (10) reported that colleagues tended to equate good teaching with creativeness. Students tended to equate it with good personality.

Wolins (113) reported, in a factor-analytic study, that peer ratings made by 974 subjects had reliabilities in the high .80's. The peer technique had also shown...
results that correlated highly with a criterion such as success in leadership ability (48, 111).

Jardin (54) concluded that peer judge rating differences attributable to the sex of the rater were statistically insignificant. Peer ratings were reliable and potentially useful feedback devices for use in education programs.

Engle and Betz (16) found that supervisors and peers had significantly higher positive agreement ($r = .77$) on who was a good counselor than that found among supervisors and students ($r = .46$) in a regular education program. They concluded that, in general, peer ratings showed more promise in evaluation than standardized tests, academic performance, and other measures of prediction. Meanwhile, Nunn (84) found a positive correlation between peer popularity and academic achievement in freshman nursing students.

The use of peer ratings in educational evaluation situations was sparse indeed. Little appeared to be known about the reliability of such appraisal devices. This is unusual in view of the fact that peer ratings are widely used in non-educational settings.

The current research on all four categories of criteria (administrative, student, self, and peer) provided no definite indicators as to which method of teacher evaluation is best. This is due, for the most part, to the lack of objective productivity measures in the educational system. One of the primary objectives of the present study is to ascertain
which mode of teacher evaluation is best when more objective productivity measures are utilized as the criteria. In other words, which of these rating groups, all of whom have so far been associated with questionable appraisal value, most accurately perceive the most productive teachers?

The Teacher as a Leader

The third area of concern in this study has to do with the teacher as a leader. It was evident, from the literature review, that very little had been said about the teacher as a leader. In fact, most articles deliberately avoided speaking about the teacher being in any position of decision or authority. Whether this is due to recent emphasis on humanism and openness in education cannot be determined, but it is clearly evident that the instructor's role as a leader has been severely played down.

It is a contention of the present study that the disregard of the leadership aspect of teaching has been a grave mistake in modern education, and that the teacher's leadership behavior is an objective way of getting at the teacher appraisal problem. While a majority of schools of education give little or no training or instruction in leadership principles and styles, it is difficult to visualize the teacher in the classroom as being any different from many positions of leadership in industry, business, and the military. In addition, many teachers eventually make their way up the promotional ladder, where they find themselves in unquestionable
positions of authority and leadership such as department heads, principals, superintendents, and committee chairmen. A quick glance at the current educational bureaucracy of almost any school district or university gives evidence of a vast administrative hierarchy not very different from non-educational organizations.

Modern schools of education have been graduating teachers in the image of a purveyor of knowledge. They tend not to look at themselves as leaders of people. Education has more or less assumed that each teacher graduated possesses very similar qualifications and approaches to the teaching/learning process. However, they differ widely in their individual approaches to the learning process. The contention of this study is that these individual differences reflect differences of leadership style, and that each style can, in turn, be related to productivity and improvement.

Cheong (6) offered the opinion that the teacher was the key figure in the teaching/learning situation, and influenced the pupil's philosophical frame of reference toward education. Field, Simpkins, Browne, and Rich (19) had similar conclusions and determined that certain discriminant functions helped to identify sets of interrelated behaviors that defined their teaching styles. It did not require much further analysis to interpret teaching styles as leadership styles.
Gagné (28) viewed the teacher as a manager of the conditions of learning, and cited others who used terms such as learning technician, instructional manager, learning engineer, and director of learning. He saw the teacher functioning as a decision maker in the areas of motivation, transfer of knowledge, assessment, conditions of learning, and the structuring of knowledge to be learned. In short, Gagné felt that the teacher's traditional role as an imparter or purveyor of knowledge should be de-emphasized, and that there would be a trend toward the teacher becoming more a manager of learning. The teacher would be planning, organizing, controlling, and coordinating learning in the educational enterprise. The role of the teacher in the future would require an increasing amount of supervisory skills so that he would be able to manage his educational team more effectively.

It was interesting to note that Gagné referred to the teacher as a leader and manager, and that he spoke of education as an enterprise. Obviously, the leadership aspect of teaching was seen as being even more important in the future than it had been, making the playing-down of leadership style highly illogical and inefficient.

Mitten (77) pointed out how important it was for the student or emerging manager to understand his leadership style in order to achieve the highest level of performance. He claimed that thorough knowledge of one's leadership style
could answer the question, "Have you shaped the circumstance or has it shaped you?" (77, p. 49).

Goldman and Frass (30) found that every group begets a leader. If one was not appointed, one emerged from the group itself with its approval. Apparently groups functioned best with a leader, and would more readily follow a proven leader.

Cooper (9) and Fleishman and Harris (25) discovered that men who worked for highly competent leaders were absent less, made fewer grievances, were late less often, and had less turnover of personnel. Both studies concluded that leadership style and competence were important factors in productivity and motivation. It did not require much imagination to visualize the teacher with similar problems (absenteeism, tardiness, etc.) if they were lacking in leadership skills.

Fantini (18) cited an unpublished study by Stanford University economist Henry Levin where it was suggested that administrators, who were under pressure from the public to account for both teacher productivity and the cost/effectiveness of the educational dollar, would do best by recruiting verbally able teachers. He claimed that experience and a number of other variables should be disregarded in favor of verbal ability scores. Not once was leadership style mentioned, nor was the possible association of verbal ability and leadership qualities discussed.

Halpin (36) compared educational administrators and military officers on their styles of leadership; he found that military leaders were more efficient and productive.
Educational administrators lacked an important dimension of leadership. The basis for Halpin's findings will be discussed in the next section, but at face value these results indicated that educational administrators were not as effective, as leaders, as they could have been.

Hills (46) found that a leader's style would manifest itself in the manner with which the leader dealt with persons outside as well as inside his work group. A given person's leadership style was found to be very stable over time.

Holloman (49) found that the role of military supervisors was no more clearly defined or supported by social custom than the role of civilian supervisors. This indicated a great degree of communality for leaders in all areas. The perception of supervision was constant across agencies, but different between hierarchical levels. Similar results were obtained by Mitchell and Porter (76).

Coats, Swierenga, and Wickert (8) described how social scientists had generally viewed the behavior of group leaders as loading on essentially two factors. These factors were given a number of different labels by researchers, but basically had been thought of as person-centered and organization centered. The early work of Halpin (39) and Stogdill (96), which used the Leader Behavior Description Questionnaire (LBDQ), suggested that educational administrators also were perceived along two basic dimensions: Structure and
Consideration. As a result of this research, there was some reason to believe that two basic factors, which might be labeled teacher-centered and student-centered, accounted for most of the variance in student-teacher relationships. Since these two dimensions, and the instrument used to measure them, are an important part of the present study, the rationale, development, and use of the questionnaire was necessary.

The Development of the Supervisory Behavior Description Questionnaire

The need for a questionnaire to measure leadership style became evident when a large amount of leadership work began in the late 1940's and early 1950's. This period of time found our nation just ending one war and beginning another. Military strength levels were quite high and personnel were dispersed around the world, making them accessible to researchers in many locations.

One of the first areas to be examined was the role and function of the leader in a group. An early study in this area sprang from research on the authoritarian personality. Adorno, Frenkel-Brunswick, Levinson, and Sanford (1) suggested that the authoritarian as a leader would be more likely accepted by his superiors than the equalitarian leader. Based upon this, Rohde (93) used B-29 bomber crews and tested the proposition that the authoritarian as a follower would be more likely accepted by the leader.
than the equalitarian. The results indicated that authoritarian people were not well accepted as either leaders or followers. Authoritarian persons were not desired as friends, did not inspire confidence, and were not looked upon as good risks in combat. Hollander (47) found similar results among naval preflight school cadets. The amount of authoritarianism began to appear as a cogent factor in leader behavior.

One of the largest and most significant research programs on leadership was conducted at Ohio State University during the years 1946-1956. This program was responsible for a variety of significant findings (98), but one of its most important contributions was related to the amount of authoritarianism possessed by a leader and its effect on the preferences and performance of subordinates.

Initially it was decided that there must be certain dimensions along which a leader's behavior varied, and that these dimensions must be labeled, quantified, and measured. Hemphill and Coons (43) tentatively designated nine dimensions of leader behavior which were integration, communication, production emphasis, representation, fraternization, organization, evaluation, initiation, and domination. A 150-item questionnaire, called the Leader Behavior Description Questionnaire (LBDQ), used the method of summated ratings with scaled multiple choice responses.

The number of dimensions measured, and the number of items on the LBDQ, were found to be prohibitive for research
where time and economy were at stake. For this reason, Halpin and Winer (40) reduced the number of LBDQ items to 130, administered the scale to fifty-two B-50 bomber crews, and factor analyzed the data obtained. The results showed that only two factors, Consideration and Structure, accounted for 83% of the total variance. Many of the LBDQ items were irrelevant to these two factors and were discarded to make the total questionnaire eighty items in length.

These two scales, Consideration and Structure, were found to be correlated to a low degree, and sufficiently independent to permit their use as separate measures of different kinds of behavior.

The Consideration dimension reflected the extent to which an individual was likely to have job relationships characterized by mutual trust, respect for subordinates' ideas, and consideration of their feelings. A high score was indicative of a climate of good rapport and two-way communication. A low score indicated that the supervisor was likely to be impersonal in his relations with group members.

The Structure dimension reflected the extent to which an individual was likely to define and structure his role and those of his subordinates toward goal attainment. A high score characterized individuals who played a more active role in directing group activities through planning, communicating information, scheduling, trying out new ideas, etc.
Other studies, using bomber crews as subjects, included Hemphill and Sechrest's (44) investigation of how superiors perceived the effectiveness of the aircrews under them. The somewhat paradoxical results of this study showed that a superior's ratings of crews under him was reliable, but was based on unreliable bombing accuracy data. Thus, there was more to a superior's perception of his crews than objective bombing data. Raters could agree in their knowledge of achievement records, but be quite divergent about the meaningfulness of such records. The effects of a superior's style and perception were seen as relevant topics for subsequent research.

Fleishman (20) continued the study of the LBDQ via more extensive factor analyses. A reorganization of the items into relatively more independent categories of leader behavior was accomplished. Two major factors or dimensions remained once again: Consideration and Structure. Based on item dimension loadings derived from an industrial population, two revised scoring keys were developed, one for Consideration and one for Structure. Twenty-eight items best met the criteria for Consideration, and twenty items for Structure were selected. Hence, the questionnaire now consisted of forty-eight items and was renamed the Supervisory Behavior Description (SBD) (23). The SBD was tested on an industrial population sample, and the results indicated (via orthogonal loading factors) that the two dimensions were quite independent.
The LBDQ and SBD questionnaires were usually given to subordinates, who were asked to fill in the items in reference to what their supervisor actually did. Subordinates rated supervisors. Fleishman (21) reworded the SBD, while keeping essentially the same items, and gave this form to supervisors, asking them to fill it out in terms of what they "should" do as leaders. This revision was called the Leadership Opinion Questionnaire (LOQ). By using this questionnaire, it was possible to compare the subordinates' perception of their leader with the leader's perception of how he thought he should behave as a leader. A pilot study revealed that industrial departments with high worker grievance rates contained foremen who perceived their own supervisors as expecting them to lead with a lower degree of Consideration and a higher degree of Structuring. These differential expectations were seen as a cause of role conflict for leaders and supervisors.

During approximately the same time period, a number of other studies were revealing similar findings. Halpin (35) used the LBDQ on B-29 bomber crews and found that the crew's satisfaction with its commander was positively correlated with the Consideration dimension scores, and negatively related to the Structure scores. However, the superior's ratings were negatively related to the Consideration scores and positively related to the Structure scores. Thus, superiors and subordinates perceived the two dimensions
in opposite ways with subordinates preferring consideration in their leaders, and superiors preferring structure in their commanders. This difference confronted the leader with conflicting role expectations and constituted what the author called "the dilemma of leadership" (35, p. 22).

Halpin (38) performed a similar study which again showed that wing and squadron superiors rated favorably the performance of those commanders who showed high Structure behavior. The crews continued to prefer, as aircraft commanders, those leaders who were high in Consideration behavior. The critical finding in this study was that if leaders must satisfy both superiors and subordinates, who demand very different modes of behavior, it was best to choose an aircraft commander who was above average on both leader behavior dimensions.

In another study using B-29 bomber crews as subjects, Halpin (37) asked each crewmember to fill out the LBDQ on his own aircraft commander, and each commander was asked to fill out the LBDQ on himself. The results pointed out that the commander's belief about how he should behave as a leader was related in no way to how the crew perceived his leader behavior. This was true even though the commanders were well aware of the desirability of scoring high on both dimensions.

In an attempt to broaden the scope and use of the LBDQ, Halpin (36) undertook a study which compared educational administrators and aircraft commanders. A group of educational leaders and a group of B-29 and B-50 aircraft
commanders were rated on the LBDQ by their subordinates and by themselves. This resulted in a comparison of "ideal" and "real" leader behavior. Once again, for both samples of leaders, there was very little or no relation between the leader's belief in how he should behave and his behavior as described by the group members. The hypothesis that the educational administrators, both in their leadership beliefs and leader behavior, would display more Consideration and less Structure than the military leaders was supported. The administrators scored high on only the Consideration dimension, and it was assumed that educational leaders were not as effective as leaders as they might have been since being high on both dimensions had previously been found to be most productive (36, 38).

Other researchers were also becoming involved with the study of flying crews during this time. Haythorn (41) found that B-29 crew performance was related to how much the aircraft commander was liked, and the amount of liking depended on the extent to which the commander performed the informal leadership role expected by the crew members. It appeared that the formal group leader was expected to behave in certain ways during informal, off-duty hours. The extent to which the leader fulfilled the crews expectations on this matter was directly related to training mission scores. Another study (7) indicated that the development of desirable attitudes in newly formed aircrews, the evolution
of mutual confidence, and the willingness to go into combat were all related to the commander's amount of Consideration as measured by the LBDQ. Structure was strongly related to the rapid development of friendships and confidence. Thus, high scores on both dimensions were desirable for optimum B-29 crew effectiveness.

Meanwhile, Fleishman, Harris, and Burtt were continuing the use of the SBD in industrial settings. It appeared that "more Consideration and less initiating Structure on the part of foremen is a good thing for everybody" (24, p. 95). The authors decided to find out if this assumption was valid. They used four objective criteria to measure foreman effectiveness. The objective criteria were absenteeism, accidents, grievances, and turnover. The resultant correlations indicated that greater Consideration went with less absenteeism, while more Structure was accompanied by more absenteeism. There was also a tendency for more Structure to go with more grievances. Accidents and turnover had no appreciable relation to either dimension. As in an earlier study (35), the subordinate workers preferred the foremen that were high on the Consideration dimension and low on Structure. The most efficient foremen, in the eyes of their superiors, were those who showed more Structure and less Consideration.

Fleishman (22) pulled together all the available evidence on the SBD, and summarized both the military and industrial findings. The Consideration and Structure dimensions
were shown to be independent and reliable. Reliability was assessed in terms of internal consistency, interrater agreement, and stability of repeated measures. Validity was assessed through correlations with independent leadership measures such as objective group indices (absenteeism, turnover), productivity ratings, peer ratings, and leaderless group situation tests. The scales were found differentially predictive of a number of these criteria. Correlations with other measures revealed that the scores achieved were independent of measures of general intelligence. Low correlations between the scales and other psychometric and background measures, and substantial correlation with certain leadership criteria suggested that these scales had strong potential as useful additions in the field of leadership research and assessment.

In 1955 the military flying crew was suddenly cast aside as a mode of investigating leader behavior. Though several subsequent studies used flying crews (32, 59), these projects were concerned with the prediction of sociometric choices and group effectiveness respectively. The discontinuance of using flying crews was probably due to the termination of the Korean conflict. The military services were severely cut and researchers most likely found it very difficult to obtain sufficient samples in one location to warrant further research. Concomitantly, the phase-out of propeller-driven aircraft began in earnest, and the
experimental introduction of turbine-powered jet aircraft began. All of these changes added to the confusion of the times, and helped to stifle research on flying crews.

Industrial research continued, however, on the Consideration and Structure dimensions of the SBD and LOQ. Rambo (89) conducted six separate item analyses on the rating scales and found that the intercorrelation between scales was .02, suggesting that halo effect was negligible on these measures. Another study (4) reported a correlation of .69 between the two scales when used by superiors, and of .40 when used by subordinates. The credibility of these unusual results was questionable in that the supervisors used had very little opportunity to determine work structures in this sample. The authors admitted that this factor could very well have influenced the results.

Fleishman and Peters (26) expanded the use of the SBD in another industrial setting. The results indicated compatibility with previous military studies in that the manager's leadership attitudes toward the Consideration and Structure dimensions did not correspond with top management's effectiveness ratings. Higher-level managers tended to feel that they should initiate less Structure with subordinates. Top management tended to identify the effectiveness of subordinate managers with the effectiveness of their superiors; while subordinate managers could make their own bosses "look good," the reverse was also possible. Consideration and Structure
were positively related only in the fact that those leaders who were most effective tended to emphasize both dimensions.

In a related study, Fleishman and Harris (25) investigated the relationship between the leader behavior of industrial supervisors and the behavior of their group members. They found that low Consideration and high Structure went with high grievance and turnover rates. There appeared to be certain critical levels beyond which increased Consideration or decreased Structure had no effect on grievance or turnover rates. Foremen could compensate for high Structure by increased Consideration, but low-Consideration foremen could not compensate by decreasing their Structuring behavior. A most significant finding was that leader behavior tended to be quite stable over a period of years. This stability removed some of the possible variability in the measures and enhanced the results relative to previous findings.

Hills (46) determined that the Consideration and Structure dimensions were not solely concerned with problems of internal leadership. His findings showed that Consideration for group members could also be reflected in the manner with which a leader dealt with outsiders, and that a leader could initiate Structure upward as well as downward. Thus, the two dimensions were much more generalizable than had been previously thought, having external as well as internal applicabilities.
A series of other studies examined a variety of aspects pertaining to the questionnaires and dimensions previously discussed. Fleishman and Salter (27) found that a supervisor's empathy was significantly related to his Consideration behavior. Lennerlof (65) provided further evidence, from a Swedish population, that the Consideration and Structure dimensions were mutually independent and had satisfactory split-half reliability. Greenwood and McNamara (31) obtained similar results and found that the two dimensions measured by the LOQ were uncorrelated with various behavioral measures of supervisory performance, attitudes, and abilities. Stogdill (97) tested the validity of the subscores of the LBDQ by having observers rate movie presentations based on the LBDQ behaviors. The results were positive in that the movie roles were correctly described.

Korman (61) conducted a review of the literature on the relationship between Consideration, Structure, and various organizational criteria. His results showed a predominance of moderate correlations, but almost all of a concurrent validity nature. There was almost no evidence of predictive validity for the two dimensions, where productivity was concerned, nor on the kinds of situational moderators which might affect such validity. Earlier, Spitzer and McNamara (95) had found indications of predictive validity, as did Bass (3). These two studies were the only ones predictive of productivity and reflected the dearth of work in this area. One
of the purposes of the present study is to attempt to establish predictiveness using the Consideration and Structure dimensions.

Only one study (75) questioned the reliability and validity of the LBDQ and SBD. This study was unique in that it used a new statistical technique called the multitrait-multimethod analysis of validity. The results indicated that the variables under examination (Structure and Consideration) had relatively low reliabilities and an almost complete lack of relative construct validity. These findings did not cast any real doubt upon the numerous studies cited previously, especially in view of the fact that this research involved ratings of a very short interaction situation by ad hoc group members. The leaders of these groups were also arbitrarily assigned rather than elected or appointed by management. Design deficiencies of this kind made the credibility of the entire study doubtful.

Thus, the lengthy and painstaking research on leadership had produced a useful questionnaire which clearly delineated the two dimensions along which leader behavior varied. These two dimensions did identify two separate, reliable continua.

Teacher Evaluation and Consideration and Structure

A review of recent teacher evaluation research showed a very interesting parallel between what is typically asked for on a majority of teacher evaluation forms and the Consideration and Structure dimensions of the SBD. Coats,
Swierenga, and Wickert (8) had already identified a teacher-centered (Structure) dimension and a separate student-centered (Consideration) dimension. Similarly, McKeachie, Lin, and Mann (71) spoke of a structure and a rapport dimension which corresponded very closely in definition to Structure and Consideration, respectively.

When these two dimensions are again defined, the parallels presented become even more striking. Consideration reflected mutual trust, respect for subordinates' ideas and feelings, two-way communication, and good interpersonal relations. Structure reflected the defining and structuring of subordinate roles, active participation in directing group activities, planning, communicating information, scheduling, and trying out new ideas.

McDaniel and Ravitz (70) also offered data to support this idea and pointed out that the best liked professors were those who were well prepared for their class presentation, were able to communicate their material, exhibited a personal interest in students, and were fair in their grading. When students were asked to describe their best and worst experiences with college teaching, their responses clustered strongly about two important points: attention to or neglect of the task at hand, and sensitivity or insensitivity to student feelings. It was not difficult to see that these two points were identical to Structure and Consideration respectively, and teachers who exhibited high
involvement in these two aspects of teaching were most likely to be judged best as teachers.

Likewise, Roberson (91) discussed the self-examination of teaching profiles by teachers, and directly implicated the two dimensions. For example, he spoke of the discovery by a teacher that the class discussion planned had not taken place simply because he had lectured all period and had not permitted student responses, questions, or discussion. This was obviously indicative of too much Structure and not enough Consideration.

To demonstrate more effectively the parallels between typical teacher evaluation items and SBD items, a series of current evaluation questionnaires were analyzed in relation to item make-up, intent of the items, and the corresponding SBD dimension category.

Thomas (105) proposed an instrument for obtaining student evaluations of teachers. The fifty items that made up this instrument were quite easily classified into either the Consideration or Structure dimension, and the test was made up of approximately an equal number of each. Table I gives some typical items from this questionnaire and some parallel SBD items.
TABLE I

SAMPLE ITEMS FROM THOMAS' QUESTIONNAIRE
WITH CORRESPONDING SBD DIMENSIONS
AND PARALLEL SBD ITEMS

<table>
<thead>
<tr>
<th>Corresponding SBD Dimension</th>
<th>Items</th>
<th>Parallel SBD Item(s)*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Structure</td>
<td>&quot;The teacher makes us recall what we have just learned.&quot;</td>
<td>5, 9, 38</td>
</tr>
<tr>
<td></td>
<td>&quot;The teacher uses objects to demonstrate what he is trying to tell us.&quot;</td>
<td>3, 24</td>
</tr>
<tr>
<td></td>
<td>&quot;The teacher gets mad when we do not do our assignments correctly.&quot;</td>
<td>43, 46, 48</td>
</tr>
<tr>
<td>Consideration</td>
<td>&quot;The teacher shows disapproval when we misbehave.&quot;</td>
<td>17, 28, 31</td>
</tr>
<tr>
<td></td>
<td>&quot;The teacher has other students respond to what they think about answers that have been given to questions.&quot;</td>
<td>13</td>
</tr>
<tr>
<td></td>
<td>&quot;The teacher shows approval of good work.&quot;</td>
<td>21</td>
</tr>
<tr>
<td></td>
<td>&quot;The teacher will not allow us to discuss both sides of an issue.&quot;</td>
<td>7</td>
</tr>
</tbody>
</table>

*See Appendix A.

Hunter (53) believed evaluation to be too much of a one-way process and designed a report card for teachers to be filled out by students. This forty-item instrument was made up primarily of Consideration items and possessed limited response possibilities of "yes," "no," and "sometimes." Table II shows some typical items from this questionnaire and some parallel SBD items.
<table>
<thead>
<tr>
<th>Corresponding SBD Dimension</th>
<th>Items</th>
<th>Parallel SBD Item(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Structure</strong></td>
<td>&quot;The kids fool around a lot in this class.&quot;</td>
<td>33</td>
</tr>
<tr>
<td></td>
<td>&quot;This class understands the teacher's explanations.&quot;</td>
<td>37</td>
</tr>
<tr>
<td></td>
<td>&quot;Homework in this class is useful to me.&quot;</td>
<td>15</td>
</tr>
<tr>
<td></td>
<td>&quot;I usually get my work done on time.&quot;</td>
<td>20</td>
</tr>
<tr>
<td><strong>Consideration</strong></td>
<td>&quot;My teacher is careful not to hurt the children's feelings.&quot;</td>
<td>22, 28</td>
</tr>
<tr>
<td></td>
<td>&quot;My teacher makes me feel important in this class.&quot;</td>
<td>8, 29</td>
</tr>
<tr>
<td></td>
<td>&quot;My teacher helps kids when they make mistakes.&quot;</td>
<td>10, 17, 31</td>
</tr>
<tr>
<td></td>
<td>&quot;My teacher praises me.&quot;</td>
<td>21</td>
</tr>
<tr>
<td></td>
<td>&quot;My teacher uses ideas that are suggested by pupils.&quot;</td>
<td>7, 41</td>
</tr>
</tbody>
</table>

*See Appendix A.

Jones (55) suggested some evaluation forms for appraising teachers which contained items that were evenly divided between the two SBD dimensions. There were three forms of this rating scale: one for superior appraisal, one for self-appraisal, and one for student appraisal. All of the forms contained the same items with different wordings to
fit the situation. The item statements were to be checked either "satisfactory" or "needs improvement" which limited the discrimination of teacher attributes, and made recommendations for improvement difficult. Table III contains some typical items from the superior as evaluator form of this questionnaire and some parallel SBD items.

TABLE III
SAMPLE ITEMS FROM JONES' QUESTIONNAIRE WITH CORRESPONDING SBD DIMENSIONS AND PARALLEL SBD ITEMS

<table>
<thead>
<tr>
<th>Corresponding SBD Dimension</th>
<th>Items</th>
<th>Parallel SBD Item(s)*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Structure</td>
<td>&quot;A variety of teaching materials are utilized.&quot; 3, 24</td>
<td>30</td>
</tr>
<tr>
<td></td>
<td>&quot;All class assignments are clearly explained.&quot;</td>
<td>20, 37</td>
</tr>
<tr>
<td></td>
<td>&quot;Tests are strictly related to measuring accomplishment of course objectives.&quot;</td>
<td>26, 35</td>
</tr>
<tr>
<td></td>
<td>&quot;This teacher is prepared almost every day.&quot; 30</td>
<td></td>
</tr>
<tr>
<td>Consideration</td>
<td>&quot;Students have a major role in forming objectives.&quot; 41</td>
<td>11, 14, 22, 26, 35</td>
</tr>
<tr>
<td></td>
<td>&quot;Course objectives meet most students' individual differences.&quot;</td>
<td>22, 47</td>
</tr>
<tr>
<td></td>
<td>&quot;Teaching methods and materials provide for students' individual differences.&quot;</td>
<td></td>
</tr>
<tr>
<td></td>
<td>&quot;Class discussion is encouraged.&quot; 42, 45</td>
<td></td>
</tr>
</tbody>
</table>

*See Appendix A.
Roberson (91) offered a Teacher Self-Appraisal (TSA) system which included nine possible methods a teacher might select to use, nine levels of objectives to accomplish, and seven levels of verbal and nonverbal expressions to utilize. The items and definitions were primarily Consideration items with a few Structure items included. Table IV contains some typical terms and definitions used for the TSA system and some parallel SBD items.

**TABLE IV**

SAMPLE TERMS AND DEFINITIONS FROM ROBERSON'S QUESTIONNAIRE WITH CORRESPONDING SBD DIMENSIONS AND PARALLEL SBD ITEMS

<table>
<thead>
<tr>
<th>Corresponding SBD Dimension</th>
<th>Items</th>
<th>Parallel SBD Item(s)*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Structure</td>
<td>Demonstrate--&quot;Teacher supplements talk with visual clues or external props.&quot;</td>
<td>3, 24</td>
</tr>
<tr>
<td></td>
<td>Direction--&quot;Teacher commands or insists students comply.&quot;</td>
<td>9, 37</td>
</tr>
<tr>
<td></td>
<td>Mastery--&quot;Teacher drills or practices specifics with students.&quot;</td>
<td>38</td>
</tr>
<tr>
<td>Consideration</td>
<td>Support--&quot;Teacher praises, repeats student response, or uses student idea.&quot;</td>
<td>41</td>
</tr>
<tr>
<td></td>
<td>Receptive--&quot;Teacher indicates to a student that the lines of communication are open.&quot;</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td>Unresponsive--&quot;Teacher openly ignores student question, request, or response.&quot;</td>
<td>47</td>
</tr>
</tbody>
</table>

*See Appendix A.*
After videotaping a teaching session, each teacher was supposed to appraise himself on an IBM card as to which term given best described the method, objectives, and expressions used at random intervals identified by an electronic "beep." In this way a self-profile was established and examined for improvement.

Gromish, Bamford, Rous, Sall, and Rubin (33) condensed a variety of teacher traits to be evaluated into four major categories. Their research included the same categories under student and superior ratings of teacher effectiveness. Two of the categories fell into the Consideration dimension, and the other two fell into the Structure dimension. Table V contains the four categories used and some parallel SBD items.

**TABLE V**

CATEGORIES FROM GROMISH, ET AL. QUESTIONNAIRE WITH CORRESPONDING SBD DIMENSIONS AND PARALLEL SBD ITEMS

<table>
<thead>
<tr>
<th>Corresponding SBD Dimension</th>
<th>Items</th>
<th>Parallel SBD Item(s)*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Structure</td>
<td>&quot;Organization and Clarity&quot;</td>
<td>9, 30, 37</td>
</tr>
<tr>
<td></td>
<td>&quot;Instructional Methods&quot;</td>
<td>3, 24</td>
</tr>
<tr>
<td>Consideration</td>
<td>&quot;Teaching Attitude&quot;</td>
<td>1, 12, 40</td>
</tr>
<tr>
<td></td>
<td>&quot;Teaching Approach&quot;</td>
<td>1, 12, 40</td>
</tr>
</tbody>
</table>

*See Appendix A.*

Watman (109) proposed a list of examples of factors that teachers and supervisors should be concerned about in
evaluation. Out of five example factors put forth, only one could be classified as a Structure item, and the rest were Consideration. Table VI contains four of the proposed factors and some parallel SBD items.

<table>
<thead>
<tr>
<th>Corresponding SBD Dimension</th>
<th>Items</th>
<th>Parallel SBD Item(s)*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Structure</td>
<td>&quot;Does the teacher plan for instruction?&quot;</td>
<td>37</td>
</tr>
<tr>
<td></td>
<td>&quot;Is the teacher open-minded, flexible, introspective, and ethical?&quot;</td>
<td>11, 14, 26,</td>
</tr>
<tr>
<td></td>
<td>&quot;Is the teacher enhancing, and not threatening?&quot;</td>
<td>35</td>
</tr>
<tr>
<td></td>
<td>&quot;Does the teacher provide for a wide opportunity for student</td>
<td>12, 14, 26</td>
</tr>
<tr>
<td></td>
<td>participation?&quot;</td>
<td></td>
</tr>
</tbody>
</table>

*See Appendix A.

An informal questionnaire for measuring student reaction to instruction, for the purpose of being an input to the teacher evaluation process, had been used in certain classes at North Texas State University (83) for the past several years. The twenty-five items on the questionnaire were easily classified into either the Consideration or Structure dimensions, and reflected a preponderance of Structure items. Table VII contains some typical items.
from this informal evaluation instrument and some parallel SBD items.

### TABLE VII

**SAMPLE ITEMS FROM NORTH TEXAS STATE INFORMAL QUESTIONNAIRE WITH CORRESPONDING SBD DIMENSIONS AND PARALLEL SBD ITEMS**

<table>
<thead>
<tr>
<th>Corresponding SBD Dimension</th>
<th>Items</th>
<th>Parallel SBD Item(s)*</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Structure</strong></td>
<td>&quot;Students knew at the beginning of the course what all of the require-</td>
<td>20, 30, 37</td>
</tr>
<tr>
<td></td>
<td>ments would be.&quot;</td>
<td></td>
</tr>
<tr>
<td></td>
<td>&quot;Both the daily and long range assignments were clear in terms of</td>
<td>30, 37</td>
</tr>
<tr>
<td></td>
<td>what the instructor expected.&quot;</td>
<td></td>
</tr>
<tr>
<td></td>
<td>&quot;The instructor's lectures seemed well prepared and stimulating.&quot;</td>
<td>3, 24</td>
</tr>
<tr>
<td></td>
<td>&quot;It appeared that the instructor adhered to the announced bases for</td>
<td>37, 38</td>
</tr>
<tr>
<td></td>
<td>evaluation of students' work.&quot;</td>
<td></td>
</tr>
<tr>
<td></td>
<td>&quot;The objectives of the course were clear from the beginning of the</td>
<td>30, 37</td>
</tr>
<tr>
<td></td>
<td>course.&quot;</td>
<td></td>
</tr>
<tr>
<td><strong>Consideration</strong></td>
<td>&quot;Assignments and projects were both meaningful and reasonable in</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>terms of the course objectives.&quot;</td>
<td></td>
</tr>
<tr>
<td></td>
<td>&quot;Adequate time was allowed for students to complete the assignments</td>
<td>4, 25, 39</td>
</tr>
<tr>
<td></td>
<td>and projects.&quot;</td>
<td></td>
</tr>
<tr>
<td></td>
<td>&quot;There was opportunity for students to pursue their own special</td>
<td>8, 10</td>
</tr>
<tr>
<td></td>
<td>interests within the framework of the course.&quot;</td>
<td>19, 29</td>
</tr>
<tr>
<td></td>
<td>&quot;The instructor appeared interested in the particular needs and</td>
<td></td>
</tr>
<tr>
<td></td>
<td>interests of individual students.&quot;</td>
<td></td>
</tr>
<tr>
<td></td>
<td>&quot;There was ample opportunity for student participation in class</td>
<td>13, 41</td>
</tr>
<tr>
<td></td>
<td>discussions; questions were answered.&quot;</td>
<td></td>
</tr>
</tbody>
</table>

*See Appendix A.*
There were a large number of evaluation forms that showed identical dichotomies when their items were classified into the two SBD dimensions. The evidence is overwhelmingly in favor of the viewpoint that, whether they know it or not, educational appraisers are asking evaluators to describe a teacher's leadership style in terms of the two SBD dimensions. What had been sought is how much or how little Consideration and Structure a teacher possesses. Factor analyses of these evaluation instruments would likely reveal two factors comparable to those from the SBD which could account for a vast majority of the variance in the questionnaire. The fact that each questionnaire item could be paired easily with one or more parallel items from the SBD made this contention appear credible.

Forty out of the forty-eight total SBD items were utilized in describing parallels in the relatively small samples of items used in the preceding tables. The fact that a majority of the questionnaires dealt with Consideration items while others dealt primarily with Structure items reflected a lack of understanding of the items used.

Assuming that the goal of evaluation instrument designers is a description of a teacher's leadership style, it seems logical that such an approach can be applied more effectively if the persons involved are made aware of the real nature of the evaluations. As House had stated,
If one could develop an instrument that was easy to administer, non-threatening to the teacher, that spoke in terms meaningful in the teacher's world (unlike most psychometric instruments which measure things like "con-voluted neurasthenia"), and was cheap enough for even an administrator to purchase, the knot of evaluation would be partly unraveled (52, p. 406).

The SBD approach to teacher evaluation just might be such an instrument. It is one purpose of the present study to see if its use in this manner is as feasible as it appears to be.

**Productivity**

The research related to productivity was sketchy and inconclusive, and the relationship between style of leadership and group performance received but a small portion of the leadership research effort. Korman's review article (61) revealed that only two studies had been done using the LOQ with objective measures of productivity as criteria. Only three studies had been done using the LBDQ with the same criteria.

Katz, Maccoby, and Morse (56) found that the supervisors of high producing sections seemed to think and act differently with respect to their supervisory functions than did the supervisors of low producing sections. These results could have been the effect of productivity differences. In another study (67), the level of group performance was manipulated and was found to shape the leader's
style of management. Performance changed supervision instead of vice versa.

Lawshe and Nagle (64) established that the supervisor's behavior, as perceived by the employees, was highly related to the output of the work group. Similar findings were evident in a number of other studies as well (12, 24, 32, 59, 114).

In a recent study, Cummins (11) found support for the hypothesis that the effect of Structure on rated productivity and quality of workgroup performance, in an industrial manufacturing plant, was moderated by considerate leader behavior. The results supported the hypothesis when quality was the criterion measure, but not when productivity was the criterion measure. Thus, what research had been related to productivity was inconclusive.

Ghiselli (29) indicated that any single dimension of job performance could not be taken as synonymous with job performance in its entirety. Performance on any job was said to be best described in terms of at least several independent dimensions. Most efforts to measure performance had been limited to the use of one measuring mode which was expected to measure total job performance.

Taylor, Parker, and Ford (102) questioned the ability of the average supervisor to provide an appraisal of his subordinates adequate enough to be used as the true measure of performance. Supervisor ratings were seen as a measure
of only one aspect of job performance; many other aspects remained.

Kipnis (59) advocated the development of other measures of job performance without the expectation that the measures need correlate with each other or, more importantly, correlate with supervisors' ratings. Instead, he recommended that if technical knowledge was an important component of the criterion, job knowledge tests might be used; if willingness and diligence were important, supervisor ratings could be included; if human relations skills were to be measured, peer or subordinate ratings could be used.

One purpose of the present study is the development and use of a better measure of productivity, and to include as many job aspect measures as possible. It seems that the more job aspects measured, the better the productivity estimate. In spite of Kipnis' warnings as to the difficulty of obtaining such data, this study obtained and used as many productivity measures as could be found, and combined these measures in a way which reflected something more than just classroom effectiveness for a teacher's performance.


47. Hollander, E. P. Authoritarianism and leadership choice in a military setting. American Psychologist, 1953, 8, 368 (abstract).


51. Holmes, D. S. Effects of grades and disconfirmed grade expectancies on students' evaluations of their instructor. *Journal of Educational Psychology*, 1972, 63, 2, 130-133.


82. Neujahr, J. L. Classroom observational research. The Educational Forum, 1972, 36, 2, 221-228.


93. Rohde, K. J. The relation of authoritarianism of the aircrew member to his acceptance by the airplane commander. *American Psychologist,* 1951, 6, 323.


100. Swanson, C. & Weaver, C. N. Faculty appraisal: current practices in schools of business. *Journal of Business Education,* 1972, 47, 6, 237-238.

101. Swanson, R. A. & Sisson, D. J. The development, evaluation, and utilization of a departmental faculty appraisal system. *Journal of Industrial Teacher Education,* 1971, 9, 1, 64-79.


CHAPTER III

METHOD

Subjects
To obtain groups of men who perform as instructors and leaders in the presence of superiors, students, and peers on a daily basis, the United States Air Force training system, more specifically the pilot training program, was selected. The subjects were eighty male instructor pilots from a United States Air Force undergraduate pilot training base. They ranged in age from twenty-two to thirty years, and possessed similar educational backgrounds with all holding bachelor's degrees, and four holding master's degrees.

Instrument
Based on the previously established premise that what teacher evaluation questionnaires were after was description of a teacher's leadership style, the present study used the Supervisory Behavior Description (SBD) to measure teaching/leadership style directly. In order to provide data to ascertain which mode of teacher evaluation was most efficient in predicting a productivity criterion, four separate evaluations of each instructor's leadership style were obtained: (a) students rated the instructors; (b) superiors rated the instructors; (c) peers rated the instructors, and
(d) the instructors rated themselves. Hence, four sets of evaluations were obtained on the same groups of subjects.

The SBD was originally designed for the specific purpose of having subordinates rate superiors. The reliability of the full scale questionnaire, which the students filled out on their instructors, had been established by a number of studies. Fleishman (2) summarized all the available evidence on the SBD and found both of the dimensions to be independent and reliable. Reliability was assessed in terms of internal consistency, inter-rater agreement, and stability on repeated measures. In a similar work, Fleishman and Harris (4) found that leader behavior, as described by the SBD dimensions, tended to be quite stable over a period of several years.

Other evidence of reliability was provided by Lennerlof (10) who obtained satisfactory split-half reliability coefficients in a Swedish population. Greenwood and McNamara (5) obtained similar results, and Stogdill and Coons (16) provided a close look at the reliability of the LBDQ, which was a direct precursor of the SBD, and demonstrated increasing reliability coefficients as the questionnaire developed.

The most comprehensive reliability data on the SBD was provided by Fleishman (3) who demonstrated split-half reliabilities which, when corrected for full length by the Spearman-Brown formula (12), ranged from .89 to .98 for the Consideration dimension, and from .68 to .87 for the...
Structure dimension. Some evidence for reliability was provided by the inter-rater agreement of descriptions by several different people on the same supervisor. The coefficients used were the unbiased correlation ratio (eta), and the Horst Coefficient ($\rho$) which yielded results in the same general range. These coefficients ranged from .55 to .73 for the Consideration dimension, and from .47 to .90 for the Structure dimension. A more recent study (17) confirmed the stability of the scales over time and diverse situations. This stability held over a twenty year interval and in a country other than the United States.

The author of the questionnaire, Fleishman, developed the SBD to maximize construct validity, but urged that empirical validity studies be conducted when possible (2). Some evidence for the validity of the SBD was provided by two authors (2, 14), but as Korman (9) had pointed out, few works had provided for anything but concurrent validity. Spitzer and McNamara (13) and Bass (1) found indications of predictive validity, but these two studies were the only ones of a predictive nature and reflected a dearth of work in this area.

Fleishman (3) provided a comprehensive report of empirical validities for the SBD, but the results indicated mixed, inconclusive validity coefficients as well as diverse criterion measures. One possible derivative of the present study is to provide a basis on which to build predictive
validity for the SBD. Though not predictive in itself, the present study is seen as a needed precursor to establish a relationship from which to predict in the future.

The single study (11) which questioned the reliability and validity of the SBD had design deficiencies which cast doubt on the credibility of the study. The results indicated that the Consideration and Structure dimensions had low reliabilities and a complete lack of construct validity. In view of the fact that this study involved ratings of a very brief interaction situation by ad hoc group members, and that the leaders of these groups were arbitrarily assigned rather than selected or appointed by management, the conclusions seemed unwarranted. This was even more obvious when all of the research supportive of the SBD's reliability and validity were considered.

While the full scale SBD was used for student ratings, a slightly modified form of the SBD was used for the self, peer, and superior ratings. The original SBD consisted of forty-eight items. Twenty of these items measured the Structure dimension and twenty-eight items measured the Consideration dimension. Since this questionnaire was designed for subordinate ratings of supervisor behavior, a number of items were worded inappropriately for self, peer, and superior ratings. Ten items were eliminated leaving twenty-one Consideration items and seventeen Structure items for a total of thirty-eight items. (See Appendix A.)
The use of shortened versions of the SBD had been demonstrated in a number of studies which sought to decrease the time involved in filling out the questionnaire. Stogdill and Shartle (16) showed that reliabilities for descriptions by other persons ranged from .51 to .85 for the separate scales, and .89 for the total score. These reliabilities were as high as could be expected with each scale composed of only two to six items. Hemphill (7) and Halpin (6) used questionnaires of thirty items in length (fifteen on each dimension) with highly satisfactory results and reliability. Therefore, the elimination of the ten inappropriate items in this study was not seen as an important factor.

At this point in the study, the two SBD dimension scores, as rated by the four different evaluator groups, were the two available evaluation inputs. Since the SBD dimension scores were descriptive of highly specific behaviors, it was felt that a more inclusive, general, all-encompassing evaluation input was needed. A forced-ranking technique was decided upon in relation to each subject's overall ability as an instructor. Once again, these forced rankings were obtained from the four sets of raters and represented four separate sets of evaluations on the same group of subjects.

Procedure
The typical pilot training base (wing) consisted of two training squadrons which were made up of from six to eight training flights. Each flight consisted of between
ten and twelve instructor pilots, and each instructor had three or four student pilots assigned to him. Every flight had a flight commander who was superior to the instructor pilots and their students, and responsible for the total operation of the flight.

The training flights were close groups, and the flight commander, instructors, and students came to know one another very well in a short period of time. The daily interaction among all members of the flight was considerable, and each had many opportunities to observe the leader behavior of the flight instructors. Since each instructor and his students sat at tables in large open flight rooms, and since many syllabus missions required instructors to brief and fly with each other, each other's students, and the flight commander, every flight member had multiple occasions to observe what kind of leaders and instructors each instructor pilot was both on the ground and in-flight.

The flight commander, also a qualified instructor pilot, was in an excellent position to observe the instructional and leadership behavior of all the instructors under him since he flew with and observed each of them routinely. All instructors flew two-engine jet training aircraft which carried an instructor pilot and one student pilot. The course of training was specifically outlined in a syllabus which included four major areas of concentration: contact (visual flight), instrument flying, formation (multiple aircraft flown close
together), and navigation. The formation and navigation areas required from two to four aircraft to fly together as a "flight" with one aircraft being the flight leader. These formation missions allowed students, instructors, and flight commanders alike to observe the instructional and leadership behavior, skill, and planning of each flight instructor regularly.

Each student pilot flew primarily with his own assigned instructor. It was common, however, for every student to have flown with, been briefed by, or had other extensive contact with all of the instructors in the flight.

The student pilots were asked to fill out the long form of the SBD on their own instructor, and to make a forced-ranking of all of the flight instructors in their flight after being read these instructions:

Gentlemen:
In a couple of minutes I will give each of you a questionnaire to fill out. I want you to fill the questionnaire out on your own assigned instructor. Please do not rate the flight commander, staff personnel, or any other instructor but your own. Each of you observes the leadership behavior of your instructor daily, and that is what you are to describe with the items in this questionnaire.

Do not fill out anything on the front page. When you begin to fill the questionnaire out, just print your instructor's name near the center fold. I do not require your own name on the form--just that of your instructor. The names are necessary only to allow later correlation of the data. I guarantee that each of you will remain completely anonymous--name, squadron, and base will not be part of the later analysis and findings.

I want to caution you that there are no right or wrong answers. The items simply describe the
leadership behavior of an individual; they do not judge whether his behavior is desirable or undesirable. Differences are expected, so please do not hesitate to mark an extreme answer if that is appropriate, or right down the middle if that is appropriate. The items are simple multiple choice and require only a check mark for a response. Please make sure that your check mark lands in its intended box. Again, be sure that your instructor's name gets inside the questionnaire.

Clipped to each questionnaire you will find a small form on which you are asked to force-rank the instructors in your flight as to their overall ability as an instructor pilot. Rank them from best to worst, in your opinion, including your own instructor—no ties allowed. Please place this ranking inside the questionnaire when you finish, and put the questionnaire inside the furnished envelope and seal it.

Are there any questions?

A slightly modified form of the SBD was used for the self and peer ratings (see Appendix A). These ratings were obtained from the instructor pilots themselves after being read the following instructions:

Gentlemen:

In a couple of minutes, I will give each of you four questionnaires to fill out. One of these questionnaires is marked "self" in red in the centerfold. You are asked to enter your own name in the inside centerfold of this questionnaire, and fill the form out describing how you actually behave as a leader. Please be as objective and honest as you can. This whole research project depends upon that. So, please fill it out with how you actually do behave, not as you think you should behave.

The other three questionnaires are to be used for rating the leader behavior of three other instructors in your flight. Do not rate the flight commander or any of the staff personnel—use only your peers. Each of you observes the leadership behavior of your fellow flight members daily. I want you to select the three members of your flight that you feel most qualified to rate on their leadership behavior. This is not a popularity or personality contest, so please be as objective as possible.
in your selections. Select those people whose leader behavior you are most familiar with.

Do not fill out anything on the front page. When you begin to fill each questionnaire out, just print the name of the person you are rating near the centerfold. The names are necessary only to allow later correlation of the data. I guarantee that each of you will remain anonymous—name, squadron, and base will not be used in later analysis. I do not require your name on any questionnaire except the self-questionnaire.

I want to caution you that there are no right or wrong answers. The items simply describe the leadership behavior of an individual; they do not judge whether his behavior is desirable or undesirable. Differences are expected, so please do not hesitate to mark an extreme answer if it is appropriate, or right down the middle if that is appropriate.

When you open the questionnaires you will notice that a number of items have been crossed out. Ignore these items. Fill out only unmarked items. They are simple multiple choice and require only a check mark for a response. Please be sure your check marks land in their intended boxes. Again, be sure each questionnaire has a name in it.

Clipped to your questionnaires you will find a small form on which you are asked to force-rank the instructors in your flight as to their overall ability as an instructor pilot. Rank them from best to worst, in your opinion, including yourself—no ties allowed. Place this ranking form inside the questionnaires when you finish, and put the questionnaires inside the furnished envelope and seal it.

Are there any questions?

The superior ratings were obtained from the flight commanders who were asked to fill out a shortened form of the SBD on each of their instructors after being read the following instructions:

Sir:

In a couple of minutes, I am going to ask you to fill out a short questionnaire on each instructor in your flight. Do not fill out a questionnaire on anyone except instructors from your flight. You daily observe the leadership behavior of your instructors and that is what the items in the questionnaire will give you an opportunity to describe.
Do not fill out anything on the front page. When you begin to fill each questionnaire out, just print the name of the person you are rating near the inside centerfold. The names are necessary only to allow later correlation of the data. I do not need your name on any questionnaire. I guarantee that you will remain anonymous—name, flight, squadron, and base will not be used in later analysis.

I want to caution you that there are no right or wrong answers. The items simply describe the leadership behavior of an individual; they do not judge whether his behavior is desirable or undesirable. Differences are expected, so please do not hesitate to mark an extreme answer if it is appropriate, or right down the middle if that is appropriate.

When you open the questionnaires you will notice that some of the items are crossed out. Ignore these items. Fill out only unmarked items. They are simple multiple choice and require only a check mark for a response. Please be sure your check marks land in their intended boxes. Again, make sure that each questionnaire has a name in it.

Clipped to your questionnaires you will find a small form on which you are asked to force-rank the instructors in your flight as to their overall ability as instructor pilots. Rank them from best to worst, in your opinion—no ties allowed. Place this ranking form inside the questionnaires when you are finished, and put the questionnaires inside the furnished envelope and seal it.

Are there any questions?

A total of 114 instructor pilots were originally sampled, but due to a number of sampling restrictions this number was reduced. Only instructors who had been in their present flight for a minimum of three months, had at least two student ratings and at least two peer ratings filled out on them, were retained as viable subjects. All subjects had self and superior ratings.
Of the fourteen flights in the wing, two were unusable because the student pilots were equivalent to being the freshman class and had not been in their respective flights long enough to be able to render valid responses to the questionnaire items. A third flight was unusable because the students had just finished the course, and the instructors were being utilized by a number of other flights until the next student class arrived. Hence, these instructors were without regular students and not functioning within their own flight structure.

The remaining eleven flights each had between eight and thirteen instructor pilots in them. When the instructors who did not meet all of the restriction criteria were eliminated, there remained one flight of six, eight flights of eight, and two flights of nine viable subjects. In order to use the forced rankings previously described, it was necessary to insure that the rankings would be equitable across flights. This required that each flight have the same number of instructor pilots. To make this possible, the one flight of six was eliminated after determining that the mean questionnaire and productivity values for this flight were very close to the mean values for the other remaining flights. The two flights of nine were converted into flights of eight by randomly selecting one instructor pilot per flight to be eliminated. This resulted in ten flights of eight instructors, or a total of eighty subjects.
The eighty subjects which remained, after having met all of the restrictions, had the SBDs pertaining to them scored via the standard method of separately totalling the two dimensional scores. Each item on the questionnaire had five alternatives, which were scored 0, 1, 2, 3, 4, or the reverse whichever was appropriate (see Appendix A). Each item, and its appropriate dimension, had a mean of 2.00. Each dimension score was then a mean of all pertinent items, and this mean represented an individual’s status on the two SBD dimensions.

The productivity measure used in the present study was a composite of five separate pieces of information considered relevant to an instructor pilot’s performance. Since the literature review demonstrated a lack of agreement as to what constituted an instructor’s productivity, it was felt that the best possible solution to this problem was to include not only an instructor’s own performance, but his students’ performances as well.

Each instructor pilot was given an annual flying evaluation, in the four areas of emphasis, which encompassed the skills necessary for the training program. These major areas were contact flying, instruments, formation, and navigation. On each of these evaluation flights, the instructor pilot flew with a member of the wing standardization board which was responsible for maintaining the professionalism and proficiency of all wing instructor pilots. Though it
was possible for the same evaluator to fly more than one of these annual check flights with any individual instructor, it was most common that different evaluators were used. This procedure was followed to reduce the effects of personal biases or personality conflicts, and to maintain as much objectivity as possible.

These evaluation flights put the evaluatee in the role of an instructor pilot while the evaluator assumed the role of a student pilot. A typical mission profile was selected, briefed, and flown by the instructor while the evaluator graded both instructional ability and flying proficiency by objectively defined criteria. Instructional ability was a reflection of how the instructor briefed and prepared the student for the mission, directed and planned maneuvers, analyzed poor performance, and debriefed after the mission. Flying proficiency scores were a reflection of actual aircraft handling, demonstration of maneuvers, and patterns and landings.

The instructional scores from the most recent evaluations, in the four areas of emphasis, were averaged for each instructor and constituted the first aspect of productivity. The mean proficiency scores constituted the second. Both the instructional and proficiency scores were in the form of points possible out of 100.

The third aspect of productivity was the number of errors given during the flight evaluations. An error was
an objectively defined mistake or oversight on the part of the instructor which could have led to a compromise of flight safety, possible damage to equipment, or violation of flying procedures and regulations. Errors were assessed over and above instructional and proficiency scores and affected neither. A total of three or more errors on any single evaluation constituted failure of that mission. Since errors represented a negative aspect of performance, they had to be weighted and subtracted from the total productivity score. When the total number of errors assessed per instructor during the most recent evaluations in the four emphasis areas was determined, the weighting schema as shown in Table VIII was used.

**TABLE VIII**

**WEIGHTS ASSIGNED RELATIVE TO TOTAL ERRORS ASSESSED**

<table>
<thead>
<tr>
<th>Total Errors</th>
<th>Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 - 2</td>
<td>0.0</td>
</tr>
<tr>
<td>3 - 5</td>
<td>-1.0</td>
</tr>
<tr>
<td>6 - 7</td>
<td>-2.0</td>
</tr>
<tr>
<td>8 - 9</td>
<td>-3.0</td>
</tr>
<tr>
<td>10 - 11</td>
<td>-4.0</td>
</tr>
</tbody>
</table>

The fourth aspect of productivity was reflected in written job knowledge tests. Such tests were often given in conjunction with the flying evaluations; however, no-notice and higher headquarters tests were administered on
other occasions. These tests measured knowledge of aircraft systems, emergency procedures, operating procedures, flying regulations, and a host of other topics. The four most recent written test scores were averaged for each instructor as a measure of the fourth aspect of performance. These test scores were also in the form of points possible out of 100.

The fifth and last aspect of instructor productivity was a measure of the performance of his students. Many of the studies cited in the review of the literature called for some kind of input from student performance as an index of teacher effectiveness. Each student pilot was given evaluation flights and examinations at intervals during the training program. They were checked several times in each of the four major emphasis areas. Since students in different classes were at different stages of training at any given point in time, it was impossible to get scores on the same identical check flight or written examination for each student. Hence, the three most recent check flight scores and written test scores for all of a given instructor's students were averaged together as the fifth productivity aspect. Student check flights and written test scores were also in points possible out of 100.

The overall productivity score was a weighted composite of all five of these performance aspects added together in the following manner: (a) Each instructor pilot's own mean
instructional and flying proficiency scores from the four most recent evaluations were averaged. From this average the error weight, as determined from Table VIII, was subtracted; (b) Each instructor's four most recent written test scores were averaged; (c) Each instructor's students mean flying and written scores were averaged together. The rationale behind such a weighting scheme was that no single productivity aspect should receive an inordinate amount of weight in the final composite score. The above weighting scheme gave equal input to an instructor's flying evaluation abilities, his written test abilities, and the overall abilities of his students. Hence, the above three inputs (a, b, c) were averaged to yield the overall productivity value used as the criterion in this study. Two calculation examples are presented in Table IX which demonstrate how the inputs were combined for the highest and lowest overall productivity scores obtained in the study.

TABLE IX
SAMPLE CALCULATIONS OF HIGHEST AND LOWEST OVERALL PRODUCTIVITY SCORES

<table>
<thead>
<tr>
<th>Example</th>
<th>Mean of S Score</th>
<th>Mean of W Score</th>
<th>Mean of P Score</th>
<th>Total Errors</th>
<th>p'</th>
<th>OPS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Highest</td>
<td>84.20</td>
<td>98.00</td>
<td>91.50</td>
<td>1</td>
<td>0.00</td>
<td>91.50</td>
</tr>
<tr>
<td>Lowest</td>
<td>63.00</td>
<td>93.75</td>
<td>80.00</td>
<td>11</td>
<td>-4.00</td>
<td>76.00</td>
</tr>
</tbody>
</table>

S = Student flying and written scores
W = Instructor's own written test scores
P = Instructor's own flying proficiency and instructional scores
P' = P less error weight
OPS = Overall productivity score (mean of S, W, P')
Analysis of Data

Since the purpose of the study was to ascertain which group of evaluators was most efficient in the appraisal of teacher performance, and since some mode of predictiveness was desired for future use, a correlational approach was warranted. Multiple predictor variables justified use of the multiple linear regression model.

The SBD yields two separate, orthogonal dimension scores that had, in previous research, been shown to be positively related to performance—i.e., high on both dimensions was most productive; low on both dimensions was least productive. Since each of these two dimension scores varied independently and were orthogonal, it followed that the Consideration and Structure scores could be two independent predictor variables in each multiple regression equation. In order to provide a more global input variable to each equation, forced rankings in relation to each subject's overall ability as an instructor were obtained from each group of evaluators, and were averaged for each instructor. These average ranking scores became the third predictor variable in each regression equation.

The productivity criterion was obtained by combining the instructor's own productivity with the productivity of his students. This yielded the overall productivity measure used in this study.

With the three predictor variables defined and the composite productivity criterion established, four separate
multiple regression equations were calculated: one for superiors, one for students, one for self, and one for peers. Though the four predictor equations were based on different Consideration, Structure, and ranking scores, the same productivity criterion scores were used in all the equations.

To test Hypothesis 1, it was necessary to compare the four multiple correlation coefficients to ascertain if they were significantly different from each other. Since the same eighty subjects were rated by each of the four evaluator groups, these groups had to be considered related rather than independent. This posed an unusual problem in that the common method for the comparison of correlation coefficients, the Fisher $r$ to $z$ transformation, was for independent group first-order correlations, and was not designed for use with related multiple correlation coefficients.

Review of a number of statistical textbooks revealed no available method for the comparison of related multiple Rs. Personal communications with a number of the authors of these textbooks confirmed that no known statistical method existed for this specific problem. It became obvious that some workable method for comparing related multiple correlation coefficients had to be found.

The efficiency of any multiple correlation coefficient depends upon how accurately one can predict values on the criterion variable from knowledge of the predictor variables. This concept depends on the amount of error involved—i.e., higher correlations would have less error in prediction and
vice versa. It seemed plausible then to analyze or compare not the coefficients themselves, but rather the prediction error values.

In order to put such logic into practice, the four multiple regression equations obtained were used to generate predicted productivity scores. These predicted scores were then compared to the actual obtained productivity scores for each subject in each group, and the absolute difference values were obtained. The absolute difference values represented the magnitude of prediction error for each equation.

To determine if the prediction errors of each equation were significantly different, a one-way analysis of variance for repeated measures was conducted on the absolute difference values (18). This procedure resulted in an $F$ ratio which indicated whether significant differences between groups were present or not, but did not reveal specifically which groups were involved. Hence, a Newman-Keuls Test (18) was used to find which groups were significantly different from each other. A .05 level of significance was selected for this statistical test.

Hypothesis 2 was tested via a step-wise regression procedure to ascertain which predictor variables contributed significantly to the multiple $R$ and which did not. A .05 significance level was selected for this statistical test.


CHAPTER IV

RESULTS

The four obtained multiple regression equations with all variables included are presented in Table X. The means and standard deviations for each variable in each group are available in Appendix B.

TABLE X

OBTAINED MULTIPLE REGRESSION EQUATIONS AND MULTIPLE CORRELATION COEFFICIENTS FOR THE FOUR SETS OF PREDICTORS

<table>
<thead>
<tr>
<th>Formula Group</th>
<th>Regression Equation</th>
<th>R</th>
</tr>
</thead>
<tbody>
<tr>
<td>Students</td>
<td>$0.29000X_1 + 1.03787X_2 + 0.65868X_3 + 78.66727$</td>
<td>.2344</td>
</tr>
<tr>
<td>Peers</td>
<td>$1.10922X_1 + 1.60476X_2 + 0.39099X_3 + 75.57809$</td>
<td>.8293*</td>
</tr>
<tr>
<td>Self</td>
<td>$0.49781X_1 + 1.05463X_2 - 0.78138X_3 + 81.54816$</td>
<td>.4317*</td>
</tr>
<tr>
<td>Superiors</td>
<td>$0.52672X_1 + 0.84889X_2 - 0.00056X_3 + 80.68060$</td>
<td>.5164*</td>
</tr>
</tbody>
</table>

$X_1$ = variable one--ranking scores
$X_2$ = variable two--SBD structure scores
$X_3$ = variable three--SBD consideration scores

To determine if significant differences existed among the four evaluator groups, it was necessary to obtain error scores between actual and predicted criterion values. The results of a one-way repeated measures analysis of variance of these error residuals are presented in Table XI.
TABLE XI

SUMMARY OF REPEATED MEASURES ANALYSIS OF VARIANCE FOR THE DIFFERENCES BETWEEN PREDICTED AND OBTAINED PRODUCTIVITY SCORES

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between Subjects</td>
<td>79</td>
<td>465.7627</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Within Subjects</td>
<td>240</td>
<td>311.0395</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Treatments</td>
<td>3</td>
<td>51.7451</td>
<td>17.2484</td>
<td>15.7654</td>
<td>0.0000</td>
</tr>
<tr>
<td>Residual</td>
<td>237</td>
<td>259.2944</td>
<td>1.0941</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>319</td>
<td>776.8022</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The $F$ value obtained in Table XI indicated that significant differences existed between some or all of the treatment groups. To determine which evaluator group mean error scores were significantly different from each other, a series of Newman-Keuls studentized range statistical tests were performed. A summary of the results of these tests is presented in Table XII. See Appendix C for the calculations involved in these tests.

TABLE XII

SUMMARY OF NEWMAN-KEULS STUDENTIZED RANGE TESTS OF SIGNIFICANCE FOR THE MEAN DIFFERENCE SCORES

<table>
<thead>
<tr>
<th>Group</th>
<th>Peer</th>
<th>Superior</th>
<th>Self</th>
<th>Student</th>
</tr>
</thead>
<tbody>
<tr>
<td>Peer Superior</td>
<td>-</td>
<td>*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Self Student</td>
<td></td>
<td>*</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*p < .05.
Step-wise multiple regression procedures were used to ascertain the contribution of each variable to each regression equation. The results of this analysis are presented in Table XIII.

**TABLE XIII**

**STEP-WISE MULTIPLE REGRESSION ANALYSIS FOR THE FOUR MULTIPLE CORRELATIONS**

<table>
<thead>
<tr>
<th>Evaluator Groups</th>
<th>Students</th>
<th>Peers</th>
<th>Self</th>
<th>Superior</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Variable 3 (Rankings) Entered</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Multiple $R^2$</td>
<td>.0342</td>
<td>.6570</td>
<td>.1581</td>
<td>.2522</td>
</tr>
<tr>
<td>Increase in $R^2$</td>
<td>.0342</td>
<td>.6570</td>
<td>.1581</td>
<td>.2522</td>
</tr>
<tr>
<td>$F$</td>
<td>2.7635</td>
<td>149.4111</td>
<td>14.6471</td>
<td>26.3026</td>
</tr>
<tr>
<td>$p$</td>
<td>.1004</td>
<td>.0000*</td>
<td>.0003*</td>
<td>.0000*</td>
</tr>
<tr>
<td><strong>Variable 2 (Structure) Entered</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$R^2$</td>
<td>.0475</td>
<td>.6859</td>
<td>.1801</td>
<td>.2666</td>
</tr>
<tr>
<td>Increase in $R^2$</td>
<td>.0133</td>
<td>.0289</td>
<td>.0220</td>
<td>.0145</td>
</tr>
<tr>
<td>$F$</td>
<td>1.0723</td>
<td>7.0872</td>
<td>2.0635</td>
<td>1.5182</td>
</tr>
<tr>
<td>$p$</td>
<td>.3036</td>
<td>.0094*</td>
<td>.1549</td>
<td>.2216</td>
</tr>
<tr>
<td><strong>Variable 1 (Consideration) Entered</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$R^2$</td>
<td>.0549</td>
<td>.6877</td>
<td>.1864</td>
<td>.2666</td>
</tr>
<tr>
<td>Increase in $R^2$</td>
<td>.0075</td>
<td>.0018</td>
<td>.0063</td>
<td>.0000</td>
</tr>
<tr>
<td>$F$</td>
<td>.5998</td>
<td>.4451</td>
<td>.5897</td>
<td>.0000</td>
</tr>
<tr>
<td>$p$</td>
<td>.4411</td>
<td>.5067</td>
<td>.4449</td>
<td>.9993</td>
</tr>
</tbody>
</table>

*p < .05

**Tests for the significance of increase in $R^2$**
To aid in the interpretation of the relationships between all of the variables, including the criterion measure, the first-order correlations were calculated. A summary of all possible intercorrelations is presented in Table XIV.

### TABLE XIV

**SIMPLE CORRELATIONS**

<table>
<thead>
<tr>
<th>Evaluator Group</th>
<th>Student</th>
<th>Peer</th>
<th>Self</th>
<th>Superior</th>
<th>Criterion</th>
</tr>
</thead>
<tbody>
<tr>
<td>Variables</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>1 C</td>
<td>0.01</td>
<td>0.26</td>
<td>0.09</td>
<td>0.03</td>
<td>0.19</td>
</tr>
<tr>
<td>2 S</td>
<td>-0.17</td>
<td>0.25</td>
<td>0.12</td>
<td>0.16</td>
<td>0.32</td>
</tr>
<tr>
<td>3 R</td>
<td>0.27</td>
<td>0.36</td>
<td>0.02</td>
<td>0.09</td>
<td>0.35</td>
</tr>
<tr>
<td>4 C</td>
<td>0.36</td>
<td>0.51</td>
<td>0.01</td>
<td>0.16</td>
<td>0.45</td>
</tr>
<tr>
<td>5 S</td>
<td>0.61</td>
<td>0.03</td>
<td>0.19</td>
<td>0.25</td>
<td>0.11</td>
</tr>
<tr>
<td>6 R</td>
<td>0.32</td>
<td>0.32</td>
<td>0.16</td>
<td>0.01</td>
<td>0.16</td>
</tr>
<tr>
<td>7 C</td>
<td>0.29</td>
<td>0.25</td>
<td>0.08</td>
<td>0.28</td>
<td>0.25</td>
</tr>
<tr>
<td>8 S</td>
<td>0.50</td>
<td>0.16</td>
<td>0.40</td>
<td>0.41</td>
<td>0.40</td>
</tr>
<tr>
<td>9 R</td>
<td>0.65</td>
<td>0.43</td>
<td>0.25</td>
<td>0.42</td>
<td></td>
</tr>
<tr>
<td>10 C</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11 S</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12 R</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

C = Consideration  
S = Structure  
R = Rankings  
$r > .21$ significant at .05
CHAPTER V

DISCUSSION

In testing Hypothesis 1, four of the six Newman-Keuls comparisons achieved statistical significance at the .05 level. The two comparisons that were not significantly different in prediction error values were self-ratings paired with student and superior ratings. It was noteworthy that the peer group had significantly less prediction error than all other rating groups, and that superiors had significantly less prediction error than students.

Peer evaluations were most accurate in predicting productivity. Superior evaluations were next best, but were significantly less accurate than peer ratings, and significantly more accurate than student ratings. Self-ratings were significantly different in predictiveness only from peer ratings. Student ratings were the least accurate—being significantly different from peer and superior ratings, but not self-ratings.

Another vantage point, from which to compare the four evaluation groups, was direct comparison of the multiple correlation coefficients themselves. The peer multiple $R$ (with all three variables included) was $0.8293 \ (p < .0000)$. When this value was squared, an indication of total variance accounted for by the relationships of the variables used was
obtained. Hence, the peer $R$ accounted for nearly 69% of the total variance. These values could then be compared to the superior $R$ of .5164 ($p < .0000$) which accounted for nearly 27% of the total variance; the self $R$ of .4317 ($p = .0013$) which accounted for almost 19% of the total variance; and the student $R$ of .2387 ($p = .2287$) which accounted for about 5% of the total variance.

The finding that peers were the best evaluator group was contradictory to the already established fact that peer ratings were little used in teacher evaluation. However, this particular finding was not seen as an isolated or unusual one since a large number of previous studies had also found peer ratings valuable in assessment procedures.

Wherry and Fryer (48) investigated the criticism that buddy ratings were not criteria of leadership but merely popularity contests. They found stability in peer ratings in that ratings made during the first month measured the same factors three months later. In addition, it was not until the fourth month that superiors' ratings reflected the leadership factors which fellow students identified in their first month ratings. Similar results were obtained by a number of other investigators where the stability, validity, and reliability of peer ratings were concerned (4, 11, 15, 20, 24, 25, 41).

Hollander (23) found that the validity of peer nominations was not adversely affected by considerations of
friendship, while Lewin, Dubno, and Akula (30) came to the conclusion that face-to-face interaction was not a critical variable in the peer rating process. Sells, Trites, and Parish (39) found an interesting correlate of manifest anxiety and peer ratings, where student pilots with high anxiety tended to receive lower buddy ratings from classmates and vice versa. Anxiety later correlated strongly with class standing at graduation with high anxiety students graduating in the lower half of the class.

A few studies (3, 6, 47) examined the relationship between peer ratings and self-ratings. Peer ratings were found to be the more stable and reliable measure. Another study (43) found evidence that when teachers were trained to examine and weight teacher and student behaviors in each other's classrooms, their own teaching performance improved noticeably. In addition, peer ratings had been reviewed and found useful in a wide variety of situations (12, 17, 22, 27).

By far, the largest area of research on peer ratings was concerned with predictive attributes. Peer ratings had been successfully used to predict success in military flight training (21, 44, 50), success as a supervisor or leader (1, 29, 36, 37, 46), failure to complete military flight programs (7, 8), success or failure as a member of the Peace Corps (16), first year grade point average in graduate school (49), and success as a salesman (45). Rokeach (38) found that college peers were significantly better able to predict
which people would be highly dogmatic (closed-minded) or minimally dogmatic (open-minded) than their college professor superiors were. Peer ratings, then, are reliable, stable, and valid, and have had sufficient use to be considered sound for assessment purposes.

The question to be asked after all the prior and present evidence in favor of the usefulness of peer ratings is why they are not used more often. While there is no simple answer that can be given, backed by empirical data, there is anecdotal logic available through three important corollaries. First, many responsible people are simply not aware of the value of peer ratings. Second, many administrators and supervisors see the loss of their ability to rate or evaluate those below them as a loss of power and authority. Third, informal questions and discussions concerning the use of peer ratings, for purposes such as pay raises, promotion and tenure, seems to cause discomfort and anxiety among those who would rate each other. It would appear that peers are afraid of being rated by each other, and feel more comfortable in having those below them (students) or above them (superiors) do the evaluating.

Why the great fear of peer ratings exists is not understood at this time, especially in view of the fact that fellow teachers and professors probably know as much or more about their peers' productivity as either students
or superiors. Students are regularly in touch with their instructors, but only for particular classes during specific semesters or school years. Beyond that, students really know very little about a teacher's research efforts, publication successes, professional and social endeavors, and out-of-class dedication, competence, and attitude. Many of the same shortcomings are true of superiors, especially when they have a fairly large number of teachers under their jurisdiction. The superior also rarely has the time to observe and rate the classroom proficiency of the teachers. In the present study, superiors (flight commanders) had knowledge of the performance scores of their subordinates, but were still significantly less able to discern differences in performance than peers were. Logically then, peers appear to be as good an evaluator group as can be found, and the reluctance to use them in education is puzzling.

It is clear, however, that whether academia likes it or not, the evidence is in and should no longer be ignored. The findings of the present study reiterate this point, especially where teacher evaluation is concerned. This issue has become critical in modern education, and as recently noted in nationwide publicity (34), teachers are actually going on strike because they are dissatisfied with currently used evaluation methods. Perhaps peer evaluations could help to solve this dilemma if those involved were made
aware of the reliability, validity, and general usefulness of peer ratings. Though peer ratings are not perfect, the answer to the present problem is to use whatever evaluation method works best. The present study strongly indicates that peer evaluations are best out of the four possible evaluation groups.

Hypothesis 2 was tested via step-wise regression analysis of the three input variables. In each regression equation, the overall ranking variable accounted for the most variance and therefore, was the first variable entered in the step-wise analysis. The variable that accounted for the second most variance, in all four regression equations, was the SBD Structure variable. The SBD Consideration variable accounted for the least variance in all of the equations.

Examination of the step-wise regression analysis revealed that the entry of the overall ranking variable resulted in significant multiple Rs for all evaluator groups except students; i.e., the student group was the only non-significant multiple R and, therefore, remained so for the subsequent steps.

The addition of the SBD Structure variable resulted in a significant increase in $R^2$ for only the peer group. The addition of the SBD Consideration variable resulted in no significant increase in $R^2$ for any of the evaluator groups. Hence, of the three input variables used, only rankings and
Structure contributed significantly to the predictiveness of the evaluator group (peers); only rankings contributed significantly to two other evaluator groups (self and superiors); and none of the variables contributed significantly to the last evaluator group (students). These results do not support Hypothesis 2 when all of the evaluator groups are considered, but when peers alone are examined, a different conclusion emerges.

Since the major purpose of the present study was to ascertain which evaluator group was most efficient in predicting productivity, and since the peer group was significantly more accurate in this respect, the subsequent discussion will involve only the peer group evaluators. The peer group also had two significant variables whereas the superior and self groups had only one.

It is worthy of note that the simple, first-order correlations reveal at least a partial explanation as to why all three variables did not contribute significantly to the peer multiple R. Both SBD variables correlated moderately with overall rankings (.61 for Structure and .51 for Consideration). In addition, both Structure and Consideration correlated fairly well with the criterion (.63 and .46 respectively) while the correlation between these two variables was only .36. It would appear that while overall rankings accounted for something unique in themselves, they also included much of both Structure and
Consideration. Therefore, in the step-wise regression analysis, when rankings accounted for the greatest amount of total variance, and Structure added a significant contribution, there was little left for Consideration to add that had not been explained already by the other two variables. Further research using only the two SBD dimensions might be fruitful in this respect. Support for such a future study was provided via a variance reduction technique (9) where a multiple $R$ of 0.68 was obtained utilizing only Consideration and Structure as predictors while removing the effect of rankings.

At this point, McNally's (32) division of the purpose of teacher evaluation into two categories becomes interesting. He saw the purpose of accountability in terms of administrative and instructional aspects. The administrative aspect entailed deciding whether or not a teacher should be employed, transferred, tenured, promoted, or given a raise. The instructional aspect was to aid a teacher's professional improvement or help to rectify weak areas. The results of the present study indicate that peer rankings would be the simplest method for getting at the administrative aspect of teacher evaluation. The SBD would be a simple instrument to use for the instructional aspect of teaching.

Though McNeil (33) had predicted that when using peer ratings, ranks would probably yield better results than other measures, he did not discuss a severe limitation
characteristic of rankings. As the results of the present study have indicated, rankings may include a number of factors within them, but they do not yield information which could be used to counsel or improve teacher effectiveness. Evaluation is one thing; improvement is another. That is one reason the SBD was incorporated into this study. It should also be remembered that although rankings were the best single predictor variable, the SBD Structure dimension was responsible for a significant increase in the peer multiple R. While rankings yield little instructional information, the SBD dimensions have the potential for revealing much to aid a teacher's professional improvement or to help rectify weak areas.

It was established earlier in this study that what current teacher evaluation forms described, albeit unknown to their authors, was leadership style in the form of the two SBD dimensions. If these kinds of behaviors are what is valued in the teacher evaluation process, then a direct and empirical use of these factors should be promulgated. To reiterate what House has stated,

If one could develop an instrument that was easy to administer, non-threatening to the teacher, that spoke in terms meaningful in the teacher's world . . . , and was cheap enough for even an administrator to purchase, the knot of evaluation would be partly unraveled (28, p. 406).

The procedure described in the present study is seen as just such a method. It is simple, inexpensive, and
available now. Though much research remains to be done in this respect, the presently described technique, though not perfect, is empirically based and better justified than any other evaluation technique encountered in this study. In addition, while peer rankings can be used to derive the administrative aspect of evaluation, the two SBD dimensions have the capability to supply a usable instructional aspect. The specific definitions of the Consideration and Structure dimensions lend themselves to supplying information to teachers on how to change their approach to leadership, if necessary, to become more effective.

Swanson (42) established that scoring high on both SBD dimensions was indicative of high productivity, while scoring low on both dimensions was indicative of low productivity. However, at the present time, little is known about what styles of leadership are best for any given educational level. To provide a hypothetical example, in grade school, for example, teachers may need a lot of Structuring behavior and relatively little Consideration. As children grow and develop, and Consideration becomes important, teachers high on both dimensions may be best. As students progress from high school and into college, it may be most effective for teachers to decrease both Consideration and Structuring behaviors to allow the pupils to do more on their own. This changing cycle of leadership style was modeled after Hersey and Blanchard's Life Cycle Theory of Leadership (19) which
hypothesized a curvilinear cyclical progression through all
the various combinations of the SBD dimensions. It might
also be possible that such cycling could occur over the
course of any given semester or school year with high Struc-
ture and low Consideration to begin with, high on both dimen-
sions later, low Structure and high Consideration as the time
period approaches an end, and low on both dimensions to con-
clude with when pupils require the least amount of teacher
input.

There are two possible approaches to improving teacher
effectiveness when knowledge of the SBD dimensions is pro-
vided. First, as implied from the cyclical theory above, a
teacher could modify his or her style of leadership to best fit the situation. This approach assumes that teachers, as
leaders, are capable of modifying their leadership behavior
to fit the occasion. There is some evidence available (2, 31, 35) that individuals can modify their leader behavior
at least to some extent. Second, knowledge of the SBD
dimensions could provide data for the placement of the
proper teacher into specific situations. This line of
thought is similar to that of Fiedler (10), who advocates
placement of certain leadership styles into specific sit-
uations where they appear to be most effective.

The application of the SBD dimensions could be accom-
plished in either of the two described manners while teachers
are being trained in college. At some point of optimum
familiarity, such as during team teaching, peers could evaluate each other's leadership style via the SBD. With knowledge of the two SBD dimensions, new teachers could either be counselled as to what directions they should take in the development of their leader behavior, or use the dimensional information for the most advantageous placement of the teachers. A similar program could be adopted to the non-training environment of the academic world in grade schools, junior colleges, and universities. All of this would require much additional research, especially to find out which leadership styles are most efficient at given grade levels in school or college.

Though Hypothesis 2 was not supported in three of the four evaluator groups, it did receive a substantial amount of support from the peer group. The peer group clearly established both rankings and the SBD as valuable predictors when productivity was the criterion.

A number of research efforts remain to be made in order to check and expand upon the present study. For example, it would be desirable to conduct a study in a similar military population using the regression equations from the present study to predict the productivity criterion. If such predictive validity could be established, it would represent a huge step forward not only for teacher evaluation, but also for the usefulness of the SBD. Another desirable study would be an attempt to combine the use of rankings
and the SBD dimensions to predict performance in a civilian academic environment. The problem involved in this kind of study would again be the dearth of recorded and objective performance information.

Since the problem of teacher evaluation is predominantly a civilian one, the question of the generalizability of the present study can be raised. In order to examine this issue adequately, consideration of a number of separate factors is needed.

First, the similarity or dissimilarity of the educational situation must be considered. The pilot training situation is not solely a military effort; airlines and private schools train pilots also. Though military protocol and command structures are observed, rank and authority in the instructor pilot plays no bigger role in the military environment than it does in the civilian situation. Holloman (26) found the role of the military supervisor to be no more clearly defined or supported by social custom than the role of civilian supervisors. Additionally, the claim that the military uses more structured approaches to education can be countered by the fact that many civilian training and educational programs have highly organized and structured aspects as well. Though military education programs all have syllabi of desired instructional outcomes, the same is true of practically any grade school, high school, or university even though these goals may not be in writing. Moreover, the syllabus is used
as a guide to student progress and is not a mandate for the approach to instruction. Individual teaching styles, therefore, are common to both military and civilian education programs.

Second, the present study might be criticized because it involved what some may consider the training of primarily a motor skill which cannot be compared to purely academic settings. Flight training does involve a large amount of motor skill, but no more so than many other areas of endeavor--i.e., art, engineering, draftsmanship, architecture, biology, medicine, physics, chemistry, and industrial arts to name just a few. It must also be remembered that flight training involves four or five hours of purely academic work for every hour spent airborne.

Third, the typical pilot training instructor and student may not be equivalent to the typical grade school, high school, or college instructor and student. All military pilot trainees possess college degrees as do the instructors. This would not make pilot training personnel much different than college personnel, but it may be a legitimate criticism where grade and high school students are concerned. Pilot trainees are also a highly select group which may interfere somewhat with the generalizability of the study. However, similar selectivity occurs not only in graduate schools, but in many undergraduate programs as well.
Fourth, another possible criticism of the present study concerned teacher-student ratio differences. The military pilot training situation has a very low teacher-student ratio (about 1 to 4) compared to most civilian settings where the ratio can run as high as one to several hundred. Differences of this order might be thought to contribute to possible disparities in generalizability, but evidence does exist (5, 13, 14, 18, 40) which indicates that teacher-student ratio is not an important influence on student ratings of teachers.

In conclusion, the limitations of the study were not seen as being seriously detrimental to the generalizability of the results. The present work was a serious effort to provide empirical data to an area previously ruled by conjecture and opinion. The results indicate that peers are the group best qualified to evaluate teacher effectiveness when rankings and the two SBD dimensions are used to predict objective performance measures. Leadership style via the SBD dimensions is valuable for not only teacher evaluation, but possibly also for teacher improvement or placement. Though just a beginning, the present study has attacked a number of previously difficult evaluation problems, and it has provided a basis upon which these problems may be solved in the future.
CHAPTER BIBLIOGRAPHY


CHAPTER VI

SUMMARY AND CONCLUSIONS

This study dealt with the chronic educational problem of teacher evaluation. Efficiency-minded legislators and the general public had demanded accountability and were getting it, but there were no indications that the modes of evaluation being used were valid or that the evaluators being used were competent or accurate in their ratings. The most serious flaw in contemporary attempts at teacher evaluation was the unavailability of objective productivity measures to be used as criteria. In addition, it appeared that modern education was de-emphasizing the leadership role of the teacher.

The purpose of this study was to identify which mode of teacher evaluation (ratings by superiors, self, students, or peers) was most efficient when a fairly objective productivity criterion was obtained, and to establish a basis for examining the teacher's teaching style as style of leadership. The teacher's style of leadership was described by the Supervisory Behavior Description (SBD).

The review of the literature for this study revealed massive differences of opinion and sheer confusion where accountability was concerned. Few authors agreed on who was to rate whom, for what purpose, and in relation to
Accountability had as many definitions as it had definers. A majority of the recommended methods of evaluation were unusable due to excessive costs, great complexity, and the fact that many were based on pure opinion and conjecture rather than empirical evidence. Nevertheless, it appears that evaluation and accountability are here to stay even though there is little agreement on how it should be done. It appeared that the lack of objective criteria of performance was the key factor in the existing accountability confusion.

A review of the specific evaluation criteria used in modern education indicated that administrative (superior) ratings were the most widely used measure of instructional competence. However, superior evaluations had little substantial validity data to support them, and were subject to considerable halo effect. Student ratings appeared to be a rapidly growing mode of evaluation, but these too were subject to considerable halo effect, and tended to reflect primarily the personal charisma and social qualities of instructors. The cited research results were highly mixed and inconclusive. Self-ratings were seen as being notoriously biased and showed negligible relationships with superior ratings, student ratings, and measures of student gain. Peer ratings, though widely used in non-educational settings, were little used in educational situations.
It was evident from the literature review that little was said about the leadership role of teachers. Schools of education tended to train teachers as purveyors of knowledge rather than leaders, but a series of studies were examined which made a strong case for teachers as leaders. These studies arrived at two basic factors that were important in education; teacher-centered behaviors, and student-centered behaviors. Since these two factors corresponded to the two leadership dimensions described by the SBD (Structure and Consideration), the rationale, development, and use of this questionnaire was in order.

The SBD was the result of over twenty years of research in industry and the military. Its predecessors, the Leader Behavior Description Questionnaire (LBDQ) and the Leader Opinion Questionnaire (LOQ), as well as the SBD, arrived at two orthogonal dimensions along which a leader's behavior varied. The Consideration dimension reflected the extent to which an individual was likely to have job relationships characterized by mutual trust, respect for subordinates' ideas, and consideration of their feelings. The Structure dimension reflected the extent to which an individual was likely to define and structure his role and those of his subordinates toward goal attainment. Extensive research on the SBD determined that it had adequate reliability, but only concurrent validity. Nevertheless, the SBD was found to be psychometrically sound. One purpose of this
A review of a number of current teacher evaluation questionnaires revealed that every evaluation item on the questionnaires was parallel to one or more of the SBD items. What evaluators had been trying to evaluate and describe, although perhaps unknown to them, was how much or how little Consideration and Structure a teacher possessed. Since what evaluators appeared to be seeking was description of a teacher's leadership style, it was evident that a more efficient approach to accountability would be a direct use of empirical leadership behavior data.

Review of the literature pertaining to productivity showed the research results to be sketchy and inconclusive. Most of the studies conducted were using subjective judgments of performance and single job aspects as criteria. A number of authors recommended a multiple-aspect approach to the productivity criterion, but emphasized how difficult these were to obtain. One purpose of this study was to derive a better measure of productivity by combining as many job aspects as possible into a multiple-input criterion.

Since the purpose of the study was to ascertain which group of evaluators (superiors, self, students, or peers) were most efficient in the appraisal of teacher performance, and since some mode of prediction was desired for future use, a correlational approach was warranted. Multiple predictor
variables justified use of the multiple linear regression model.

Using the two SBD leadership dimensions and an overall ranking as predictors, and a composite teacher/student productivity measure as the criterion, two hypotheses were put forth to be tested. It was hypothesized that (1) there would be significant differences between administrator, student, self, and peer multiple correlation coefficients used to predict productivity when leadership style was an input to the prediction equations, and (2) each input variable to the regression equations would contribute significantly to their respective multiple correlation coefficients.

The subjects used in this study were eighty male instructor pilots from a United States Air Force pilot training base. Four separate evaluations of each instructor were obtained: (a) students rated the instructors; (b) superiors rated the instructors; (c) peers rated the instructors; and (d) the instructors rated themselves. The SBD was used to provide its dimension scores as two input variables to the multiple regression equations. These dimension scores were descriptive of highly specific behaviors. Therefore, it was felt that a more inclusive, general evaluation input was needed. A forced ranking technique was decided upon in relation to each subject's overall ability as an instructor. This ranking input served as the third predictor variable in the study.
The productivity criterion measure consisted of five separate pieces of information considered relevant to an instructor pilot's performance. These five inputs were weighted and combined in such a way that equal weight was given to an instructor's own flying and instructional ability, his written test knowledge, and his students' performance. These final productivity values were used as the criterion in each of the four multiple regression equations.

An unusual problem presented itself when Hypothesis 1 was to be tested in that no available technique could be found for the comparison of related multiple Rs. Normal ways of testing the difference between two correlation coefficients did not apply in the case of related multiple Rs, and queries to a number of authors eminent in the field of statistics provided no help. Hence, an original statistical test was devised which tested the difference in prediction error rather than the differences in multiple Rs themselves. The four separate regression equations were used to generate predicted criterion scores. The absolute differences between the predicted and actual criterion scores were obtained, and a one-way repeated measures analysis of variance was conducted on these residual values. When the $F$ ratio indicated that significant differences in prediction residuals existed, a series of Newman-Keuls studentized range statistical tests were conducted to find out specifically where the differences were.
Hypothesis 2 was tested via a step-wise regression procedure. A .05 level of significance was selected for all statistical tests in this study.

In testing Hypothesis 1, the results indicated that the error residuals differed significantly ($F = 15.7654$, $p < .0000$). The Newman-Keuls tests showed that the peer group had significantly less prediction error than all other groups, and that the superior group had significantly less prediction error than the student group. The multiple $R^2$s for the groups were .6570 for peers, .2522 for superiors, .1581 for self, and .0342 for students. The peer, superior, and self multiple $R$s were significant when all three variables were included. The student multiple $R$ was not significant.

In testing Hypothesis 2, it was found that the overall ranking variable accounted for the most variance in all regression equations, and contributed to significance in all except the student equation. The SBD Structure variable accounted for the next highest amount of variance in all equations, but added a significant increase only to the peer multiple $R^2$. The Consideration variable accounted for the least amount of variance in all equations, and did not add significantly to any of them.

It was concluded that peer evaluations were most accurate in predicting productivity. Superior evaluations were next best followed by self and finally student...
evaluations. The finding that peers were the best predictor group was in contradiction to the previously established fact that peer ratings were little used in teacher evaluation, but was not seen as an isolated or unusual result. A large number of studies were reviewed which demonstrated the reliability and validity of peer ratings in a wide variety of settings.

In view of this finding, the question of why peer ratings were not used more often was considered. There were three possible answers. First, responsible people were just not aware of the value of peer ratings. Second, administrators and supervisors viewed the use of peer ratings as a usurping of their power and authority. Third, those who would rate each other as peers were afraid of such an approach, and seemed to be more comfortable with superior or student ratings. The reasons for this fear of peer ratings were not known, but an examination of personal knowledge possessed by each group involved gave strong support to the fact that peers were as good a group of evaluators as could be found. The reluctance to use them in education was puzzling.

Regardless of the preferences demonstrated, the present study strongly indicated that peer evaluations were best out of the four possible evaluation groups. Therefore, Hypothesis 1 was fully supported in the case of peers, but only partially supported where the other groups were concerned.
Hypothesis 2 was partially supported in that overall rankings contributed significantly to three out of the four multiple Rs. None of the other variables contributed significantly except the SBD Structure variable in the peer multiple R. Hence, for peers, two out of the three predictor variables contributed significantly to the multiple R. The question as to why all three variables did not contribute significantly to the peer multiple R was answered by the fact that overall rankings included both Consideration and Structure in them. When rankings accounted for the greatest amount of variance, and Structure added another significant contribution, there was little left for Consideration to add to the peer multiple R.

Future research possibilities are discussed, and a number of recommendations for rapid application of the results of this study are offered. While rankings account for the most variance, they are not seen as a valuable tool for the professional improvement of teachers. The two SBD dimensions are offered as sources of information for either the improvement or placement of teachers into appropriate educational situations. The procedure described in this study is seen as a simple, inexpensive, and currently available method for teacher evaluation. Though not perfect, the technique described is empirically based and better justified than any other evaluation technique.
encountered in the study. Possible applications of the SBD
dimension scores are also discussed.

The limitations of the study are examined, and a num-
ber of possible criticisms are answered. The limitations
and criticisms are not seen as detrimental to the general-
izability of the results. The multiple correlational
approach to teacher evaluation, via peer rankings and
ratings of leadership style, is an effort to provide
empirical data to an area previously limited to conjecture and opinion.
INSTRUCTIONS:

You have observed your own supervisor and probably you know pretty well how he operates. In this questionnaire, you are simply to describe some of the things your own supervisor does with your group.

For each item, choose the alternative which best describes how often your supervisor does what that item says. Remember...there are no right or wrong answers to these questions. The items simply describe the behavior of the supervisor over you; they do not judge whether his behavior is desirable or undesirable. Everyone's supervisor is different and so is every work group, so we expect differences in what different supervisors do.

Answer the items by marking an "X" in the box (a, b, c, d or e) next to each item to indicate your choice.

Copyright 1970, Edwin A. Fleishman, Ph.D.  
Printed in U.S.A. All rights reserved.
1. HE IS EASY TO UNDERSTAND.
   a. always   b. often   c. occasionally   d. seldom   e. never

2. HE ENCOURAGES OVERTIME WORK.
   a. a great deal   b. fairly much   c. to some degree   d. comparatively little   e. not at all

3. HE TRIES OUT HIS NEW IDEAS.
   a. often   b. fairly much   c. occasionally   d. once in a while   e. very seldom

4. HE BACKS UP WHAT PEOPLE IN HIS WORK GROUP DO.
   a. always   b. often   c. occasionally   d. seldom   e. never

5. HE CRITICIZES POOR WORK.
   a. always   b. often   c. occasionally   d. seldom   e. never

6. HE DEMANDS MORE THAN WE CAN DO.
   a. often   b. fairly often   c. occasionally   d. once in a while   e. very seldom

7. HE REFUSES TO GIVE IN WHEN PEOPLE IN THE WORK GROUP DISAGREE WITH HIM.
   a. always   b. often   c. occasionally   d. seldom   e. never

8. HE EXPRESSES APPRECIATION WHEN ONE OF US DOES A GOOD JOB.
   a. always   b. often   c. occasionally   d. seldom   e. never

9. HE INSISTS THAT PEOPLE UNDER HIM FOLLOW STANDARD WAYS OF DOING THINGS IN EVERY DETAIL.
   a. always   b. often   c. occasionally   d. seldom   e. never

10. HE HELPS PEOPLE IN THE WORK GROUP WITH THEIR PERSONAL PROBLEMS.
    a. often   b. fairly often   c. occasionally   d. once in a while   e. very seldom

11. HE IS SLOW TO ACCEPT NEW IDEAS.
    a. always   b. often   c. occasionally   d. seldom   e. never

12. HE IS FRIENDLY AND CAN BE EASILY APPROACHED.
    a. always   b. often   c. occasionally   d. seldom   e. never

13. HE GETS THE APPROVAL OF THE WORK GROUP ON IMPORTANT MATTERS BEFORE GOING AHEAD.
    a. always   b. often   c. occasionally   d. seldom   e. never

14. HE RESISTS CHANGES IN WAYS OF DOING THINGS.
    a. a great deal   b. fairly much   c. to some degree   d. comparatively little   e. not at all

15. HE ASSIGNS PEOPLE UNDER HIM TO PARTICULAR TASKS.
    a. always   b. often   c. occasionally   d. seldom   e. never

16. HE STRESSES BEING AHEAD OF COMPETING WORK GROUPS.
    a. a great deal   b. fairly much   c. to some degree   d. comparatively little   e. not at all

17. HE CRITICIZES A SPECIFIC ACT RATHER THAN A PARTICULAR INDIVIDUAL.
    a. always   b. often   c. occasionally   d. seldom   e. never
35. HE REJECTS SUGGESTIONS FOR CHANGES.
   a. always  b. often  c. occasionally  d. seldom  e. never

36. HE CHANGES THE DUTIES OF PEOPLE UNDER HIM WITHOUT FIRST TALKING IT OVER WITH THEM.
   a. often  b. fairly often  c. occasionally  d. once in a while  e. very seldom

37. HE DECIDES IN DETAIL WHAT SHALL BE DONE AND HOW IT SHALL BE DONE.
   a. always  b. often  c. occasionally  d. seldom  e. never

38. HE SEES TO IT THAT PEOPLE UNDER HIM ARE WORKING UP TO THEIR LIMITS.
   a. always  b. often  c. occasionally  d. seldom  e. never

39. HE STANDS UP FOR PEOPLE UNDER HIM EVEN THOUGH IT MAKES HIM UNPOPULAR.
   a. always  b. often  c. occasionally  d. seldom  e. never

40. HE MAKES THOSE UNDER HIM FEEL AT EASE WHEN TALKING WITH HIM.
   a. always  b. often  c. occasionally  d. seldom  e. never

41. HE PUTS SUGGESTIONS THAT ARE MADE BY THE PEOPLE UNDER HIM INTO OPERATION.
   a. always  b. often  c. occasionally  d. seldom  e. never

42. HE REFUSES TO EXPLAIN HIS ACTIONS.
   a. often  b. fairly often  c. occasionally  d. once in a while  e. very seldom

43. HE EMPHASIZES THE QUANTITY OF WORK.
   a. a great deal  b. fairly much  c. to some degree  d. comparatively little  e. not at all

44. HE ASKS FOR SACRIFICES FROM HIS PEOPLE FOR THE GOOD OF THE ENTIRE DEPARTMENT.
   a. often  b. fairly often  c. occasionally  d. once in a while  e. very seldom

45. HE ACTS WITHOUT CONSULTING THE PEOPLE UNDER HIM FIRST.
   a. often  b. fairly often  c. occasionally  d. once in a while  e. very seldom

46. HE "NEEDLES" PEOPLE UNDER HIM FOR GREATER EFFORT.
   a. a great deal  b. fairly much  c. to some degree  d. comparatively little  e. not at all

47. HE INSISTS THAT EVERYTHING BE DONE HIS WAY.
   a. always  b. often  c. occasionally  d. seldom  e. never

48. HE ENCOURAGES SLOW-WORKING PEOPLE TO GREATER EFFORT.
   a. often  b. fairly often  c. occasionally  d. once in a while  e. very seldom
18. HE LETS OTHERS DO THEIR WORK THE WAY THEY THINK BEST.
   a. always  b. often  c. occasionally  d. seldom  e. never

19. HE DOES PERSONAL FAVORS FOR THE PEOPLE UNDER HIM.
   a. often  b. fairly often  c. occasionally  d. once in a while  e. very seldom

20. HE EMPHASIZES MEETING OF DEADLINES.
   a. a great deal  b. fairly much  c. to some degree  d. comparatively little  e. not at all

21. HE SEES THAT A WORKER IS REWARDED FOR A JOB WELL DONE.
   a. always  b. often  c. occasionally  d. once in a while  e. very seldom

22. HE TREATS PEOPLE UNDER HIM WITHOUT CONSIDERING THEIR FEELINGS.
   a. always  b. often  c. occasionally  d. once in a while  e. very seldom

23. HE INSISTS THAT HE BE INFORMED ON DECISIONS MADE BY THE PEOPLE UNDER HIM.
   a. always  b. often  c. occasionally  d. seldom  e. never

24. HE OFFERS NEW APPROACHES TO PROBLEMS.
   a. often  b. fairly often  c. occasionally  d. once in a while  e. very seldom

25. HE TREATS ALL WORKERS UNDER HIM AS HIS EQUALS.
   a. always  b. often  c. occasionally  d. seldom  e. never

26. HE IS WILLING TO MAKE CHANGES.
   a. always  b. often  c. occasionally  d. seldom  e. never

27. HE ASKS SLOWER PEOPLE TO GET MORE DONE.
   a. often  b. fairly often  c. occasionally  d. once in a while  e. very seldom

28. HE CRITICIZES PEOPLE UNDER HIM IN FRONT OF OTHERS.
   a. often  b. fairly often  c. occasionally  d. once in a while  e. very seldom

29. HE STRESSES THE IMPORTANCE OF HIGH MORALE AMONG THOSE UNDER HIM.
   a. a great deal  b. fairly much  c. to some degree  d. comparatively little  e. not at all

30. HE TALKS ABOUT HOW MUCH SHOULD BE DONE.
   a. a great deal  b. fairly much  c. to some degree  d. comparatively little  e. not at all

31. HE "RIDES" THE PERSON WHO MAKES A MISTAKE.
   a. always  b. often  c. occasionally  d. seldom  e. never

32. HE WAITS FOR PEOPLE UNDER HIM TO PUSH NEW IDEAS BEFORE HE DOES.
   a. always  b. often  c. occasionally  d. seldom  e. never

33. HE RULES WITH AN IRON HAND.
   a. always  b. often  c. occasionally  d. seldom  e. never

34. HE TRIES TO KEEP THE PEOPLE UNDER HIM IN GOOD STANDING WITH THOSE IN HIGHER AUTHORITY.
   a. always  b. often  c. occasionally  d. seldom  e. never
1. HE IS EASY TO UNDERSTAND.
   a. always  b. often  c. occasionally  d. seldom  e. never

2. HE ENCOURAGES OVERTIME WORK.
   a. a great deal  b. fairly much  c. to some degree  d. comparatively little  e. not at all

3. HE TRIES OUT HIS NEW IDEAS.
   a. often  b. fairly much  c. occasionally  d. once in a while  e. very seldom

4. HE BACKS UP WHAT PEOPLE IN HIS WORK GROUP DO.
   a. always  b. often  c. occasionally  d. seldom  e. never

5. HE CRITICIZES POOR WORK.
   a. always  b. often  c. occasionally  d. seldom  e. never

6. HE DEMANDS MORE THAN WE CAN DO.
   a. often  b. fairly often  c. occasionally  d. once in a while  e. very seldom

7. HE REFUSES TO GIVE IN WHEN PEOPLE IN THE WORK GROUP DISAGREE WITH HIM.
   a. always  b. often  c. occasionally  d. seldom  e. never

8. HE EXPRESSES APPRECIATION WHEN ONE OF US DOES A GOOD JOB.
   a. always  b. often  c. occasionally  d. seldom  e. never

9. HE INSISTS THAT PEOPLE UNDER HIM FOLLOW STANDARD WAYS OF DOING THINGS IN EVERY DETAIL.
   a. always  b. often  c. occasionally  d. seldom  e. never

10. HE HELPS PEOPLE IN THE WORK GROUP WITH THEIR PERSONAL PROBLEMS.
    a. often  b. fairly often  c. occasionally  d. once in a while  e. very seldom

11. HE IS SLOW TO ACCEPT NEW IDEAS.
    a. always  b. often  c. occasionally  d. seldom  e. never

12. HE IS FRIENDLY AND CAN BE EASILY APPROACHED.
    a. always  b. often  c. occasionally  d. seldom  e. never

13. HE GETS THE APPROVAL OF THE WORK GROUP ON IMPORTANT MATTERS BEFORE GOING AHEAD.
    a. always  b. often  c. occasionally  d. seldom  e. never

14. HE RESISTS CHANGES IN WAYS OF DOING THINGS.
    a. a great deal  b. fairly much  c. to some degree  d. comparatively little  e. not at all

15. HE ASSIGN PEOPLE UNDER HIM TO PARTICULAR TASKS.
    a. always  b. often  c. occasionally  d. seldom  e. never

16. HE STRESSES BEING AHEAD OF COMPETING WORK GROUPS.
    a. a great deal  b. fairly much  c. to some degree  d. comparatively little  e. not at all

17. HE CRITICIZES A SPECIFIC ACT RATHER THAN A PARTICULAR INDIVIDUAL.
    a. always  b. often  c. occasionally  d. seldom  e. never
35. He rejects suggestions for changes.  
a. always  
b. often  
c. occasionally  
d. seldom  
e. never  

36. He changes the duties of people under him without first talking it over with them.  
a. always  
b. fairly often  
c. occasionally  
d. once in a while  
e. very seldom  

37. He decides in detail what shall be done and how it shall be done.  
a. always  
b. often  
c. occasionally  
d. seldom  
e. never  

38. He sees to it that people under him are working up to their limits.  
a. always  
b. often  
c. occasionally  
d. seldom  
e. never  

39. He stands up for people under him even though it makes him unpopular.  
a. always  
b. often  
c. occasionally  
d. seldom  
e. never  

40. He makes those under him feel at ease when talking with him.  
a. always  
b. often  
c. occasionally  
d. seldom  
e. never  

41. He puts suggestions that are made by the people under him into operation.  
a. always  
b. often  
c. occasionally  
d. seldom  
e. never  

42. He refuses to explain his actions.  
a. often  
b. fairly often  
c. occasionally  
d. once in a while  
e. very seldom  

43. He emphasizes the quantity of work.  
a. a great deal  
b. fairly much  
c. to some degree  
d. comparatively little  
e. not at all  

44. He asks for sacrifices from his people for the good of the entire department.  
a. often  
b. fairly often  
c. occasionally  
d. once in a while  
e. very seldom  

45. He acts without consulting the people under him first.  
a. always  
b. fairly often  
c. occasionally  
d. once in a while  
e. very seldom  

46. He "needles" people under him for greater effort.  
a. a great deal  
b. fairly much  
c. to some degree  
d. comparatively little  
e. not at all  

47. He insists that everything be done his way.  
a. always  
b. often  
c. occasionally  
d. seldom  
e. never  

48. He encourages slow-working people to greater effort.  
a. often  
b. fairly often  
c. occasionally  
d. once in a while  
e. very seldom
INSTRUCTIONS:

You have observed your own supervisor and probably you know pretty well how he operates. In this questionnaire, you are simply to describe some of the things your own supervisor does with your group.

For each item, choose the alternative which best describes how often your supervisor does what that item says. Remember...there are no right or wrong answers to these questions. The items simply describe the behavior of the supervisor over you; they do not judge whether his behavior is desirable or undesirable. Everyone's supervisor is different and so is every work group, so we expect differences in what different supervisors do.

Answer the items by marking an “X” in the box (a, b, c, d or e) next to each item to indicate your choice.
SUPERVISORY BEHAVIOR DESCRIPTION SCORING INSTRUCTIONS

The Supervisory Behavior Description yields two scores, one for Consideration and one for Structure. The 28 items that are keyed for Consideration should be scored first. The answer to each item receives a score of 0, 1, 2, 3, or 4 as indicated on the key (see reverse). Thus for Item 1, choice "a" would be given a score of 4, choice "b" would receive a score of 3, etc. For Item 6, choice "a" would be given a score of 0, choice "c" would be scored 2, and so on. As you go through the items that are keyed to Consideration, mark the score for each item directly on the questionnaire booklet in the margin to the right of the item answer column. After the 28 items have been marked in this way, add the individual item scores to obtain the total Consideration score. (Note: If no answer is provided for an item in this key, determine the average score of Consideration items completed, enter this average score beside the blank item(s) and include in calculating the total.) Record this total in the Consideration box (C) on the cover of the questionnaire booklet.

Then, using a different color pencil, score the 20 items in the Structure key. Before recording the total, check to be sure all 20 Structure questions were answered. If not, find the average Structure score, multiply by the number of items omitted and add to the total. Record the total Structure score in the Structure box (S) on the front of the questionnaire booklet.

Conversion to Percentiles. Consult the Norms table in the Examiner's Manual for comparing each score obtained with the appropriate supervisory or management group. Record the appropriate percentile-rank equivalent in the box on the SBD booklet cover.

(Over)
<table>
<thead>
<tr>
<th>Question</th>
<th>Options</th>
</tr>
</thead>
<tbody>
<tr>
<td>18. HE LETS OTHERS DO THEIR WORK THE WAY THEY THINK BEST.</td>
<td>a. always  b. often  c. occasionally  d. seldom  e. never</td>
</tr>
<tr>
<td>19. HE DOES PERSONAL FAVORS FOR THE PEOPLE UNDER HIM.</td>
<td>a. often  b. fairly often  c. occasionally  d. once in a while  e. very seldom</td>
</tr>
<tr>
<td>20. HE EMPHASIZES MEETING OF DEADLINES.</td>
<td>a. a great deal  b. fairly much  c. to some degree  d. comparatively little  e. not at all</td>
</tr>
<tr>
<td>21. HE SEES THAT A WORKER IS REWARDED FOR A JOB WELL DONE.</td>
<td>a. always  b. often  c. occasionally  d. seldom  e. never</td>
</tr>
<tr>
<td>22. HE TREATS PEOPLE UNDER HIM WITHOUT CONSIDERING THEIR FEELINGS.</td>
<td>a. always  b. often  c. occasionally  d. once in a while  e. very seldom</td>
</tr>
<tr>
<td>23. HE INSISTS THAT HE BE INFORMED ON DECISIONS MADE BY THE PEOPLE UNDER HIM.</td>
<td>a. always  b. often  c. occasionally  d. seldom  e. never</td>
</tr>
<tr>
<td>24. HE OFFERS NEW APPROACHES TO PROBLEMS.</td>
<td>a. often  b. fairly often  c. occasionally  d. once in a while  e. very seldom</td>
</tr>
<tr>
<td>25. HE TREATS ALL WORKERS UNDER HIM AS HIS EQUALS.</td>
<td>a. always  b. often  c. occasionally  d. seldom  e. never</td>
</tr>
<tr>
<td>26. HE IS WILLING TO MAKE CHANGES.</td>
<td>a. always  b. often  c. occasionally  d. seldom  e. never</td>
</tr>
<tr>
<td>27. HE ASKS SLOWER PEOPLE TO GET MORE DONE.</td>
<td>a. often  b. fairly often  c. occasionally  d. once in a while  e. very seldom</td>
</tr>
<tr>
<td>28. HE CRITICIZES PEOPLE UNDER HIM IN FRONT OF OTHERS.</td>
<td>a. a great deal  b. fairly much  c. to some degree  d. comparatively little  e. not at all</td>
</tr>
<tr>
<td>29. HE TALKS ABOUT HOW MUCH SHOULD BE DONE.</td>
<td>a. a great deal  b. fairly much  c. to some degree  d. comparatively little  e. not at all</td>
</tr>
<tr>
<td>30. HE &quot;RIDES&quot; THE PERSON WHO MAKES A MISTAKE.</td>
<td>a. always  b. often  c. occasionally  d. seldom  e. never</td>
</tr>
<tr>
<td>31. HE RULES WITH AN IRON HAND.</td>
<td>a. always  b. often  c. occasionally  d. seldom  e. never</td>
</tr>
<tr>
<td>32. HE TRIES TO KEEP THE PEOPLE UNDER HIM IN GOOD STANDING WITH THOSE IN HIGHER AUTHORITY.</td>
<td>a. always  b. often  c. occasionally  d. seldom  e. never</td>
</tr>
</tbody>
</table>
## SUPERVISORY BEHAVIOR DESCRIPTION SCORING KEY

<table>
<thead>
<tr>
<th>Item</th>
<th>a</th>
<th>b</th>
<th>c</th>
<th>d</th>
<th>e</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>7</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>8</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>10</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>11</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>12</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>13</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>14</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>17</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>19</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>21</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>22</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>25</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>26</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>28</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>29</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>31</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>34</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>35</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>36</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>39</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>40</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>41</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>42</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>45</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>47</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Item</th>
<th>a</th>
<th>b</th>
<th>c</th>
<th>d</th>
<th>e</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>9</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>15</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>16</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>18</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>20</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>23</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>24</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>27</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>30</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>32</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>33</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>37</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>38</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>43</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>44</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>46</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>48</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

(Over)
APPENDIX B

MEANS AND STANDARD DEVIATIONS FOR THE THREE VARIABLES FOR EACH EVALUATOR GROUP
### MEANS AND STANDARD DEVIATIONS FOR THE THREE VARIABLES FOR EACH EVALUATOR GROUP

<table>
<thead>
<tr>
<th>Group</th>
<th>Variable</th>
<th>Mean</th>
<th>Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Student</strong></td>
<td>Consideration</td>
<td>2.8418</td>
<td>0.4137</td>
</tr>
<tr>
<td></td>
<td>Structure</td>
<td>2.4155</td>
<td>0.3468</td>
</tr>
<tr>
<td></td>
<td>Rankings</td>
<td>6.8108</td>
<td>1.5462</td>
</tr>
<tr>
<td><strong>Peer</strong></td>
<td>Consideration</td>
<td>2.4115</td>
<td>0.3869</td>
</tr>
<tr>
<td></td>
<td>Structure</td>
<td>2.2138</td>
<td>0.3989</td>
</tr>
<tr>
<td></td>
<td>Rankings</td>
<td>4.4624</td>
<td>1.8011</td>
</tr>
<tr>
<td><strong>Self</strong></td>
<td>Consideration</td>
<td>2.7413</td>
<td>0.3275</td>
</tr>
<tr>
<td></td>
<td>Structure</td>
<td>2.5294</td>
<td>0.5208</td>
</tr>
<tr>
<td></td>
<td>Rankings</td>
<td>5.9250</td>
<td>2.1861</td>
</tr>
<tr>
<td><strong>Superior</strong></td>
<td>Consideration</td>
<td>2.4650</td>
<td>0.5332</td>
</tr>
<tr>
<td></td>
<td>Structure</td>
<td>2.3251</td>
<td>0.5677</td>
</tr>
<tr>
<td></td>
<td>Rankings</td>
<td>4.5000</td>
<td>2.3057</td>
</tr>
<tr>
<td><strong>Criterion</strong></td>
<td></td>
<td>85.0233</td>
<td>3.0426</td>
</tr>
</tbody>
</table>
APPENDIX C

CALCULATION DATA FOR THE NEWMAN-KEULS STUDENTIZED RANGE TESTS OF SIGNIFICANCE
### CALCULATION DATA FOR THE NEWMAN-KEULS STUDENTIZED RANGE TESTS OF SIGNIFICANCE

<table>
<thead>
<tr>
<th>Group</th>
<th>Peer</th>
<th>Superior</th>
<th>Self</th>
<th>Student</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean Error</td>
<td>1.3439</td>
<td>2.0195</td>
<td>2.1971</td>
<td>2.4209</td>
</tr>
<tr>
<td>Value</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Peer</td>
<td>—</td>
<td>0.6756</td>
<td>0.8532</td>
<td>1.0770</td>
</tr>
<tr>
<td>Superior</td>
<td>—</td>
<td>0.1776</td>
<td>0.4014</td>
<td></td>
</tr>
<tr>
<td>Self</td>
<td>—</td>
<td>0.2238</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Student</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\[ r = \frac{0.95 \sqrt{MS_{error}}}{n} \]

<table>
<thead>
<tr>
<th>Truncated Range ( r )</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>[ 0.95 (r,237) ]</td>
<td>2.77</td>
<td>3.31</td>
<td>3.63</td>
</tr>
<tr>
<td>[ \frac{MS_{error}}{n} ]</td>
<td>.3241</td>
<td>.3873</td>
<td>.4247</td>
</tr>
</tbody>
</table>
BIBLIOGRAPHY

Books


Stogdill, R. M. *Manager, employees, organizations.* Columbus: Ohio State Bureau of Business Research, Ohio State University, 1966.


Articles


Christner, C. A. & Humphill, J. K. Leader behavior of B-29 commanders and changes in crew members' attitudes toward the crew. *Sociometry*, 1955, 18, 82-87.


Cooper, R. Leader's task relevance and subordinate behavior in industrial work groups. *Human Relations*, 1966, 19, 57-84.


Fleishman, E. A. The measurement of leadership attitudes in industry. *Journal of Applied Psychology*, 1953, 37, 153-158.


Hollander, E. P. Authoritarianism and leadership choice in a military setting. American Psychologist, 1953, 8, 368 (abstract).


Holmes, D. S. Effects of grades and disconfirmed grade expectancies on students' evaluations of their instructor. Journal of Educational Psychology, 1972, 63, 2, 130-133.

House, E. R. The conscience of educational evaluation. Teachers College Record, 1972, 73, 3, 405-414.


Neujahr, J. L. Classroom observational research. The Educational Forum, 1972, 36, 2, 221-228.


Rohde, K. J. The relation of authoritarianism of the aircrew member to his acceptance by the airplane commander. *American Psychologist*, 1951, 6, 323.


Tscheulin, D. Leader behavior measurement in German industry. Journal of Applied Psychology, 1972, 56, in press.


Reports

Doll, R. E. Peer rating validity as a function of rater intelligence and rating score received. U.S.N. School of Aviation Medical Research Report, 1963, 24, 5 p. (abstract).


Unpublished Material

North Texas State University. Informal unpublished student reaction to instruction questionnaire, 1972.


Questionnaires