

1988

Rating teacher performance to determine career ladder advancement: An analysis of bias and reliability

David W. Peterson
Iowa State University

Follow this and additional works at: <https://lib.dr.iastate.edu/rtd>

 Part of the [Educational Administration and Supervision Commons](#)

Recommended Citation

Peterson, David W., "Rating teacher performance to determine career ladder advancement: An analysis of bias and reliability" (1988). *Retrospective Theses and Dissertations*. 9715.
<https://lib.dr.iastate.edu/rtd/9715>

This Dissertation is brought to you for free and open access by the Iowa State University Capstones, Theses and Dissertations at Iowa State University Digital Repository. It has been accepted for inclusion in Retrospective Theses and Dissertations by an authorized administrator of Iowa State University Digital Repository. For more information, please contact digirep@iastate.edu.

INFORMATION TO USERS

The most advanced technology has been used to photograph and reproduce this manuscript from the microfilm master. UMI films the original text directly from the copy submitted. Thus, some dissertation copies are in typewriter face, while others may be from a computer printer.

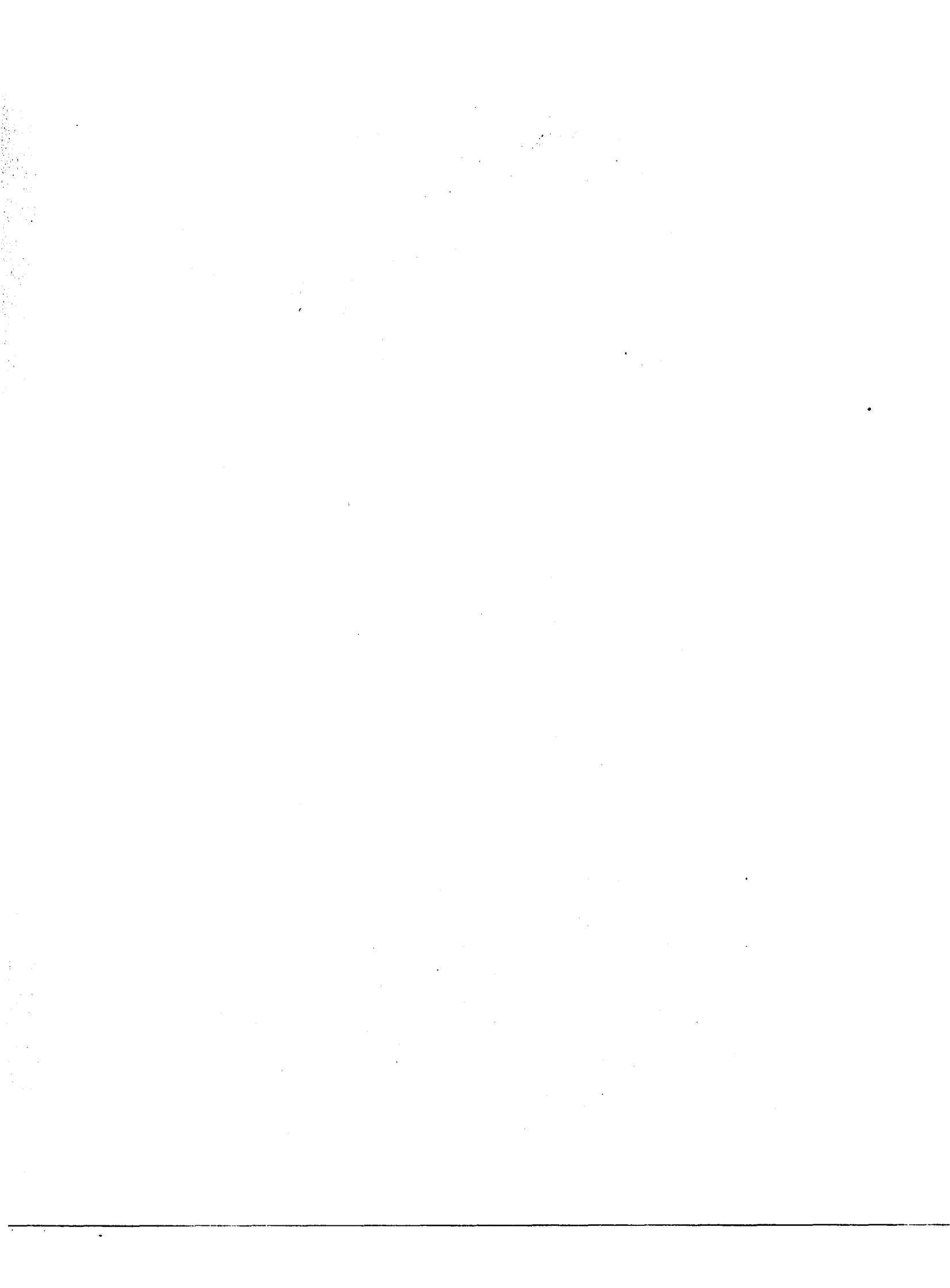
In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyrighted material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps. Each oversize page is available as one exposure on a standard 35 mm slide or as a 17" x 23" black and white photographic print for an additional charge.

Photographs included in the original manuscript have been reproduced xerographically in this copy. 35 mm slides or 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.



300 North Zeeb Road, Ann Arbor, MI 48106-1346 USA



Order Number 8825436

**Rating teacher performance to determine career ladder
advancement: An analysis of bias and reliability**

Peterson, David W., Ph.D.

Iowa State University, 1988

U·M·I

**300 N. Zeeb Rd.
Ann Arbor, MI 48106**



PLEASE NOTE:

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

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

U·M·I



**Rating teacher performance
to determine career ladder advancement:
An analysis of bias and reliability**

by

David W. Peterson

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

**Department: Professional Studies in Education
Major: Education (Educational Administration)**

Approved:

Signature was redacted for privacy.

In Charge of Major Work

Signature was redacted for privacy.

For the Major Department

Signature was redacted for privacy.

For the Graduate College

**Iowa State University
Ames, Iowa**

1988

TABLE OF CONTENTS

	<u>Page</u>
CHAPTER I. INTRODUCTION	1
Statement of the Problem	3
Purpose	6
Objectives	6
Research Hypotheses	7
Basic Assumptions	8
Delimitations	9
Definition of Terms	10
CHAPTER II. REVIEW OF LITERATURE	12
Background	12
Technical Aspects of Evaluation	15
Criteria	15
Instrumentation	18
Implementation strategies	20
Human Aspects of Evaluation	31
Leniency/severity/central tendency	32
Halo	33
Rater characteristics	34
Rater position	35
Personal bias	35
Summary	36
CHAPTER III. METHODS AND PROCEDURES	38
Introduction	38
Identification of Research Subjects	39
Collection of Data	40
Human Subjects Release	42

	<u>Page</u>
Methods of Statistical Treatment	42
Sampling procedures	43
Statistical analysis	45
CHAPTER IV. ANALYSIS AND RESEARCH FINDINGS	50
Descriptive Data	50
Hypothesis Testing	57
CHAPTER V. SUMMARY, CONCLUSIONS, AND RECOMMENDATIONS	80
Summary	80
Internal consistency of the appraisal instrument	80
Effect of rater characteristics on teacher appraisal scores	81
Inter-rater agreement	82
Intra-rater agreement	83
Conclusions	83
Limitations	85
Discussion	86
Impact on career ladder implementation	97
Recommendations for Practitioners	101
Recommendations for Further Research	103
BIBLIOGRAPHY	105
ACKNOWLEDGMENTS	117a
APPENDIX A. TEACHER APPRAISAL SYSTEM	118
Philosophy of Education	119
Philosophy of Instruction	122
Philosophy of Professional Employee Evaluation	124
Teacher Appraisal System: Purpose	124
Teacher Appraisal System: Procedures	125

	<u>Page</u>
APPENDIX B. INSTRUMENTS	132
DISD Test for Evaluators: Demographic Information	133
Classroom Observation Form	134
Written Record of Observation: Formative Appraisal	135
Written Record of Observation, Formative Appraisal: Criteria and Descriptors	136
Summative Evaluation Form	146
APPENDIX C. LETTER OF COMMUNICATION	149
APPENDIX D. DISD SUMMARY OF EVALUATOR TRAINING, 1985-86	151
APPENDIX E. CRITERIA FOR CAREER LADDER PLACEMENT AND ADVANCEMENT	153

LIST OF TABLES

	<u>Page</u>
Table 1. Mean scores and standard deviations by criteria for teacher appraisal instrument (N=112 appraisers and 3,460 teachers)	58
Table 2. Pearson product-moment correlation coefficients for subgroups of criteria within teacher evaluation instrument (N=3,460 teachers and 112 appraisers)	60
Table 3. Analysis of variance of mean elementary teacher appraisal scores by gender	62
Table 4. Analysis of variance of mean elementary teacher appraisal scores by race	67
Table 5. Analysis of variance of mean differences in elementary teacher appraisal scores by appraiser's highest level of education	67
Table 6. Analysis of variance of mean differences based upon total years of experience in education of appraiser	71
Table 7. Teacher appraisal score correlations between first and second appraisers	72
Table 8. Correlation between first semester mean teacher appraisal scores assigned by selected evaluators	73
Table 9. T-test analysis for significance of differences in mean teacher appraisal ratings for selected elementary evaluators	74
Table 10. Two-tailed T-test analysis for mean teacher appraisal scores assigned by first and second appraisers	75
Table 11. Teacher appraisal score correlations between an appraiser's first semester and second semester evaluations	77
Table 12. Correlation between mean teacher appraisal scores assigned by selected appraisers for first and second semesters	79

LIST OF FIGURES

	<u>Page</u>
Figure 1. Teacher performance appraisal instrument (formative), Dallas Independent School District, 1985-86	41
Figure 2. Number and percent of teacher evaluators by level, Dallas Independent School District, 1985-86	52
Figure 3. Number and percent of teachers by level, Dallas Independent School District, 1985-86	53
Figure 4. Number and percent of elementary evaluators and teachers by gender, Dallas Independent School District, 1985-86	54
Figure 5. Number and percent of elementary evaluators and teachers by race, Dallas Independent School District, 1985-86	54
Figure 6. Number and percent of elementary evaluators by highest level of education attained, Dallas Independent School District, 1985-86	55
Figure 7. Number and percent of elementary evaluators by total years of experience in education, Dallas Independent School District, 1985-86	56
Figure 8a. Mean first semester appraisal scores, by criteria, assigned to 3,460 elementary teachers by 112 evaluators, Dallas Independent School District, 1985-86	59
Figure 8b. Mean first semester appraisal scores, by groups of criteria, assigned to 3,460 elementary teachers by 112 evaluators, Dallas Independent School District, 1985-86	59
Figure 9. Research design and number of subjects used to determine differences in teacher performance appraisal ratings based upon gender	62
Figure 10. Mean first semester teacher appraisal scores by gender groups for selected elementary evaluators, Dallas Independent School District, 1985-86	63

	<u>Page</u>
Figure 11. Research design and number of subjects used to determine differences in teacher performance appraisal ratings based upon race	64
Figure 12. Mean first semester teacher appraisal scores by race groups for selected elementary evaluators, Dallas Independent School District, 1985-86	66
Figure 13. Mean first semester teacher appraisal scores by highest level of educational training of evaluator	68
Figure 14. Mean first semester teacher appraisal scores by evaluators' total years of experience in education	70

CHAPTER I. INTRODUCTION

Major studies have indicated that far-reaching school reform measures are necessary to improve the quality of learning in America's schools (Goodlad, 1984). The Carnegie Forum on Education and the Economy acknowledged the progress made in pursuing these reforms in the report, A Nation Prepared: Teachers for the 21st Century:

In the past three years, the American people made a good beginning in the search for an educational renaissance. They have pointed to educational weaknesses to be corrected; they have outlined ways to recapture a commitment to quality. They have reaffirmed the belief that the aim for greater productivity is not in conflict with the development of independent and creative minds. There is a new consensus on the urgency of making our schools once again the engines of progress, productivity and prosperity (1986, p. 2).

Shanker (1986) insists that continued progress in improving schools will occur only if teachers are given additional responsibility for the design and implementation of reform measures. Others (Keppel, 1986) cite the need for the federal government to take a stronger leadership role in reshaping the structure of education. State-promoted reform measures, however, have provided the majority of changes in the initial outpouring of school improvement measures. All states have expanded their school improvement programs and nearly all have increased graduation requirements for students (Odden, 1986).

Some states have also directed their attention to increasing standards for entry into, and continuance in, the teaching profession (Murray, 1986). Creating teacher incentive plans has also been a popular state initiative with 40 states having proposed some form of incentives

for teachers (Olson, 1987). The most common system is some form of career ladder/master teacher plan which provides differentiation in responsibility and compensation for teachers based upon the quality of their performance (Allen, 1986).

Bell (1983) described and advocated a career ladder model which makes use of supervisor and peer evaluations as a source of teacher advancement. Tennessee Governor Lamar Alexander, stating that "nothing is more indispensable to every recipe for better schools than high quality teaching" (Alexander, 1985), was successful in promoting legislation making Tennessee the first state to implement a state-wide career ladder system for teachers (State of Tennessee, 1984). Twelve states are fully implementing such programs with state funding, those being California, Florida, Massachusetts, Mississippi, Missouri, New Jersey, New York, Pennsylvania, Tennessee, Texas, Utah, and Washington (Olson, 1987). All states which currently have career ladder programs use supervisor and/or peer evaluations based upon performance-based criteria as part of the process for advancement of teachers (Allen, 1986).

The determination of which criteria to use for these evaluations has been the subject of intense discussion and debate among state planners as they developed their evaluation instruments (Astuto and Clark, 1985; Holdzkom, 1987; Smith, Peterson and Micceri, 1987). Evaluation criteria in career ladder states reflect the considerable research in the past ten years on effective teaching practices (Good and Brophy, 1984; Hunter, 1984; Manatt and Stow, 1984; McGreal, 1984).

These criteria tend to be clustered around various facets of teacher decision making. Berliner (1984) summarized the literature on effective teaching practices by citing four major areas that make a difference in producing student gains, those being (1) pre-instructional factors, (2) during-instruction factors, (3) climate factors, and (4) post-instructional factors.

While some general agreement on the definition of "good teaching" has emerged from the research, less consensus exists on the process for measuring teachers' performance. Forty-six states currently have a law or administrative regulation mandating the evaluation of teachers (Duke and Stiggins, 1986). The procedures used in most of the systems contain the use of direct observations and judgments by supervisors and some systems use the review of work samples, peer observations, and teacher interviews. The lack of consistency in collective bargaining agreements also tends to add complexity and a lack of uniformity to the process of evaluation.

Arthur Wise (1984) stated that the process of evaluation remains the crucial element in the successful implementation of performance-based pay systems and that people generally still believe that performance evaluation systems are not valid or reliable. Teacher unions in Tennessee and Texas also contend that improvements in the evaluation process are necessary in order to distinguish "good teachers from excellent teachers" with validity and reliability (Furtwengler, 1987; Olson, 1987).

Statement of the Problem

Performance appraisal can produce many positive outcomes both for the employee and for the organization (Decotiis and Petit, 1978; Eichel and

Bender, 1984). Those include (1) motivating employees by providing even-handed recognition of their efforts, (2) helping map out career paths for the employee, (3) giving guidance to needed training and development for employees, and (4) reducing the risk of legal challenges based upon equal employment opportunity regulations.

These positive outcomes have been cited by researchers in the private sector since the early 1900s when a need arose to improve the quality of personnel decisions (Landy, Zedeck, and Cleveland, 1983). Only in the past 10 to 15 years, however, has there been an intensive focus on the effects of teacher performance appraisal (Doyle, 1983).

Other practitioners and researchers (Blumberg, 1974; Astuto and Clark, 1985; Smith, Peterson, and Micceri, 1987) have noted a number of areas of concern regarding teacher evaluation, among them: (1) the promotion of an adversarial relationship between teachers and school administrators, (2) the lack of ability of the evaluator to make valid and reliable judgments of teacher performance, (3) the lack of sufficient funding to make career ladder/merit pay systems truly effective, and (4) the lack of evidence from research to substantiate that student achievement gains are higher in schools using incentive systems such as career ladders.

Career ladder systems which make use of performance appraisal data for promotion of teachers are especially vulnerable if the system does not provide equal opportunities for all to advance (Murnane and Cohen, 1986). The problem for this study is centered around the twin themes of fairness

and equal access as they relate to the evaluation process in career ladder systems.

Teacher evaluation has been refined and improved, thus enhancing the validity and reliability of such systems, through such efforts as the School Improvement Model (SIM) (Manatt and Stow, 1984; Manatt, 1987). The SIM project assisted the Dallas, Texas Independent School District in the design of an evaluation system to implement the Texas career ladder program for teachers. Through the course of the on-going training of evaluators, the teacher evaluations for the 1985-86 school year were collected by the SIM team. Thus, these data, which consisted of seven separate evaluations (six formative appraisals and one summative evaluation) for each of the district's 7,169 teachers, were available for the present study.

Within this context, the problem for this study can be more specifically defined by the following questions:

1. Will there be agreement among the ratings for different criteria used on the appraisal instrument?
2. Will the use of teacher evaluations for promotion in a career ladder system be subject to systematic error due to certain characteristics of the rater?
3. Will teacher evaluation ratings be subject to systematic error due to an interaction effect between certain characteristics of the rater and ratee?
4. Will there be agreement between appraisal ratings assigned by two different evaluators for a common group of teachers?

5. Can evaluators make consistent ratings on repeated measures of a teacher's performance?

Purpose

The likelihood that teachers will be dealt with fairly in a career ladder/merit pay system increases if it can be demonstrated that the evaluation system is both valid (truthful; measuring what it purports to measure) and reliable (the results are consistent across time and evaluators). Thus, the intention of this study is to:

1. Determine from the literature which teacher evaluation procedures tend to produce results that are valid and reliable.
2. Determine from the literature the major sources of systematic errors in performance evaluation, in particular those related to the rater's characteristics of gender, race, education, and experience.
3. Assess the level of systematic error due to gender, race, education, and experience within a sample of teacher evaluations from a school system using evaluation data to implement a career ladder system.
4. Determine if the evaluation instrument used by the school district in the study produced results that provided equal access to career ladder promotion for teachers.

Objectives

In order to accomplish the purposes of this study, it will be necessary to:

1. Conduct a thorough review of the literature as it relates both to teacher evaluation practices and to sources of systematic error in performance evaluation.

2. Determine the degree of agreement between two different sub-groupings of criteria within the teacher evaluation instrument used by the school district in this study.

3. Determine the degree to which teacher evaluation ratings vary based upon the evaluator's gender, race, level of educational training, and experience in education.

4. Determine if the characteristics of the evaluator's race and gender interact with the teacher's race and gender to produce differences in evaluation ratings of the teacher.

5. Determine the degree to which two different evaluators are able to identify the same level of performance when conducting independent appraisals of the same teacher.

6. Determine the degree to which an evaluator's ratings of teachers remain consistent over time.

Research Hypotheses

In order to fulfill the purposes of this study, the following hypotheses were developed and tested:

1. There will be no significant positive correlation between an evaluator's ratings for two different subsets of performance criteria on the appraisal instrument.

2. There will be no significant difference in mean teacher evaluation scores based upon the rater characteristics of gender, race, level of training, or years of experience in education.

3. There will be no significant difference in mean teacher evaluation ratings due to an interaction effect between the race and gender of the evaluator and the race and gender of the teacher.

4. There will be no significant degree of agreement between mean teacher evaluation ratings assigned by the first appraiser and those assigned by the second appraiser.

5. There will be no significant positive correlation between an evaluator's first and second semester teacher appraisal ratings.

Basic Assumptions

This study was based upon the assumptions that:

1. Research studies and current literature have identified the sources of systematic rating errors caused by rater bias.

2. An evaluator's rating represents a valid measure of a teacher's performance at that point in time.

3. Mean scores derived from the teacher appraisal instrument are normally distributed and variances between comparison groups are equal.

4. Each appraiser followed procedures prescribed by the school district in the study as they relate to observing the teacher for at least 30 uninterrupted minutes, completing the evaluation instrument independently from other evaluators and conferring with the teacher following the observation.

5. The presence of a career ladder system, and the use of appraisal scores to determine teacher advancement, had an equal impact among all teachers and evaluators in the school district used in this study.

6. The instrument used to evaluate teachers contained performance criteria that are supported by research as those that promote student achievement.

7. The presence of a career ladder system, and its use of performance appraisal data as a means for promotion, created a greater disposition for all evaluators to make more lenient judgments of performance than if the data were gathered only for research purposes.

Delimitations

This study was intended to generate knowledge about the effects that rater characteristics had on the actual teacher performance appraisals that were collected within the context of a career ladder system for teachers. Performance appraisal ratings from the Dallas, Texas Independent School District, with 475 evaluators and 7,169 teachers during the 1985-86 school year, were selected for analysis in this study.

It is presumed that the presence of a career ladder system, as well as the presence of an on-going training program to help evaluators implement the system, had an effect on the ratings. While this effect is mentioned in the study, no attempt was made to examine it in depth. Further study of the comparison of performance appraisal data between career ladder and non-career ladder systems would help show the effect of purpose on evaluation ratings, but this was not the intent of this study.

Also, while it is acknowledged that many rater characteristics exist that may have an effect on performance appraisal ratings, only the frequently studied characteristics of gender, race (black, white Hispanic), experience in education, educational training, and the rater's relationship to the ratee (first or second appraiser) were selected for analysis.

Definition of Terms

Career ladder - A performance incentive plan which provides recognition for teachers with differential pay featuring several career steps with additional responsibilities.

Criterion - A research-based behavior used in making judgments about a teacher's performance that is uniformly applied.

Evaluation system - Procedures which provide fair, objective, and consistent analysis of teaching performance.

Evaluator - A person assigned the task of making periodic judgments about the work performance of another. In this study, the terms "evaluator," "rater," and "appraiser" are used synonymously.

First appraiser - The primary evaluator who is responsible for submitting the teacher's final evaluation rating each year.

Formative appraisal - The gathering of data and assigning of evaluation ratings for the purpose of making preliminary judgments and suggestions for improvement during the school year.

Rater bias - Systematic error in the rating of performance which is traced not to actual performance but rather to characteristics of the rater or of the situation in which the rating occurs.

Reliability - The extent to which measurements (teacher evaluation ratings, in this study) are consistent across time and evaluators.

Second appraiser - An evaluator other than the teacher's direct supervisor. In this investigation all second appraisers were elementary principals in a school different from the teacher.

Stability - The consistency of performance measures over time; otherwise known as test-retest reliability.

Summative evaluation - The end-of-the-year summary rating of the teacher's performance.

Systematic error - Error in rating scores which is consistent within an individual or group of persons, as opposed to random error which is not consistent.

Validity - The degree to which an instrument is truthful in measuring what it purports to measure.

CHAPTER II. REVIEW OF LITERATURE

The ability to rate accurately is a prerequisite for success of any human performance evaluation system regardless of the purpose of the system. The review of literature assumes that a body of information exists which addresses this concept of accuracy. The sources for the search consisted of two major areas, one of those being studies from performance appraisal in business and industry. The other major source came from studies, most of them occurring in the past 10 to 15 years, of the evaluation of educational personnel.

The review of literature within these two broad areas concentrated on an attempt to: (1) provide a brief background on the state of the art in teacher evaluation, (2) identify and describe the major technical aspects that affect the accuracy and usefulness of evaluations, and (3) identify the human factors that have an effect on the validity and reliability of performance evaluations.

Background

People have been making informal judgments about each other's performance for as long as the human race has engaged in group activities (Fletcher and Williams, 1985). Most authors, however, trace the beginning of the formal practice of performance evaluation to the beginning of the twentieth century (Doyle, 1983; Landy, Zedek, and Cleveland, 1983).

Throughout the first half of the century, studies focused on issues such as methodology, statistical techniques, and psychometric properties of ratings. Few activities of significance in the area of personnel

appraisal occurred until the 1960s, and events since that time have helped this interest continue.

The passage of the 1964 Civil Rights Act generated interest in human resource planning, selection validation and performance appraisal. Economic decline, the growth of Reaganomics, and the loss of competitiveness in international markets have also focused attention on the contributions that personnel/human resources can make to organizations (Bernardin and Beatty, 1984, p. 3).

Events within the field of education in the past 10 to 15 years have provided motivation for intensive studies of performance appraisal, in particular, teacher evaluation. Concerns were expressed throughout the 1970s about eroding levels of achievement by students in America's schools. This led several governmental agencies and educational organizations to commission studies to determine the magnitude of the problems and to suggest solutions. Some of these reports included A Nation at Risk (National Commission on Excellence in Education, 1983), A Place Called School (Goodlad, 1984), High School (Boyer, 1983), and Teachers for the 21st Century (Carnegie Forum on Education and the Economy, 1986).

Many of these reports presented evidence that personnel evaluation practices were lacking in schools and that sweeping reforms were necessary. Responses at the state and local level have come under such labels as career ladders, merit pay, peer review, master teachers, mentor teachers, clinical supervision, and assessment centers (Bell, 1983; Astuto and Clark, 1985; Allen, 1986).

These systems are a response to a dissatisfaction with evaluation practices that is shared widely among professionals and lay people alike (McNeil and Popham, 1973). Medley, Coker, and Soar (1984) noted teachers' resistance to evaluation on the basis that performance appraisal systems lack objectivity, are open to bias and are not based upon relevant criteria. Arthur Wise (1984) contended that the lack of sophistication in evaluation has led to most systems being both unproductive and unfair.

On the other hand, there are authors who see reasons for optimism within the otherwise unsettled and controversial field of teacher evaluation. Peer review is seen as a positive step both toward increasing the reliability of performance evaluations and gaining teacher involvement in and acceptance of evaluation systems (Thompson, 1979; Bell, 1983; Hopfengardner and Walker, 1984; Lempeis, 1984; Cummings, 1985; Spring Hill Center, 1986). Other authors see hope in the emerging research on effective teaching practices and believe that evaluation systems now can be created based upon teacher behaviors known to have a positive effect on student achievement (Manatt and Stow, 1984; McGreal, 1984; Stallings, 1986; Zahorik, 1987).

It remains to be seen whether the current interest in teacher evaluation is part of another cycle and can be expected to diminish or whether sufficient momentum has developed to institutionalize evaluation systems in the schools (Doyle, 1983). The evolution may well depend on the ability of researchers and practitioners to deal successfully with the technical and human barriers that have prevented the achievement of performance appraisal systems in the past (Henderson, 1984).

Technical Aspects of Evaluation

Through the correct design and proper implementation of the technical aspects of a teacher evaluation system, schools can increase the likelihood of deriving benefits for the organization over a long period of time. These technical factors include (1) using criteria that are relevant to the job of the teacher, (2) developing instruments which promote accurate measurement while minimizing time-consuming paperwork, and (3) using sound strategies for implementing the system.

Criteria

Performance criteria that make sense to administrators and teachers are essential for the success of an evaluation system (Manatt, 1987). Perhaps the most sensible criterion for judging a teacher's competence is a modification of the learner (McNeil and Popham, 1973). However, the difficulty associated with assessing such results has led most researchers to use more readily available criteria, those being teacher behaviors that are normally observable in the classroom. Simon and Boyer's anthology Mirrors for the Classroom (1970) identified 79 observation systems for labeling and classifying data related to the dynamics of instruction. Researchers in the 1980s have continued to test the effects of teacher behavior on student achievement, and to develop categories for these behaviors.

Berliner (1984) summarized the literature on effective teaching strategies by using four categories of behaviors, those being (1) pre-instructional factors, (2) during-instruction factors, (3) climate factors, and (4) post-instructional factors. Through their 5-year

experience with the School Improvement Model Project, Manatt and Stow (1984, 1986) developed 24 criteria that were found to be valid in linking teaching behaviors with student achievement. These criteria are grouped in four broad areas, those being (1) productive teaching techniques, (2) organized class management, (3) positive interpersonal relationships, and (4) professional responsibilities. Allen (1986) synthesized the literature on effective teaching strategies and also noted four general categories, viz., (1) planning, (2) management, (3) climate, and (4) instruction. Hunter (1984) clustered criteria around teacher decisions, those relating to (1) content, (2) learner behaviors, and (3) teacher behaviors.

Effective teaching behaviors were further categorized into two broad areas, those being (1) management and instructional techniques, and (2) personal characteristics, by the American Association of School Administrators in its research summary Effective Teaching: Observations from Research (1986). The report indicated that for the most part effective teachers:

- tend to be good managers
- use systematic instruction techniques
- have high expectations of their students and themselves
- believe in their own efficacy
- vary teaching strategies
- handle discipline through prevention
- are usually warm and caring
- are democratic in their approach
- are task-oriented
- are concerned with perceptual meanings rather than facts and events
- are comfortable interacting with students
- have a strong grasp of the subject matter
- are readily accessible to students outside of class

--tailor their teaching to student needs
--are highly flexible, enthusiastic, and imaginative
(p. 4).

There is evidence that the attitudes and beliefs held by teachers correlate positively with their ability to provide classroom instruction that meets these research-based criteria. Good and Brophy (1984) noted that teachers' expectations of students are often matched by unequal distribution of interactions between the teacher and students perceived as being either high or low achievers. Duke and Stiggins (1986) indicated that teachers who demand a lot of themselves and are flexible are likely to react favorably to making positive changes in their teaching behaviors. Noriega (1987) found that "high gain" teachers are likely to have a strong belief that they, rather than other environmental factors, have the main influence over a student's success or failure.

Several state level and local school districts have designed and adopted evaluation instruments which reflect this research on effective teaching and on teacher beliefs and characteristics. Florida clustered teacher behaviors into four categories, those being (1) instructional organization and development, (2) presentation of subject matter, (3) communication: verbal and non-verbal, and (4) management of student conduct (Smith; Peterson, and Micceri, 1987). North Carolina developed a state-wide instrument using eight functions (Holdzkom, 1987):

1. management of instructional time
2. management of student behavior
3. instructional presentations
4. instructional monitoring of student performance
5. instructional feedback
6. facilitating instruction

7. communicating within the educational environment
8. performing non-instructional duties (p. 42).

The Dallas, Texas Independent School District (1985) used ten performance criteria on its summative evaluation instrument:

1. demonstrates effective planning skills
2. implements the lesson
3. communicates effectively with students
4. uses evaluation activities appropriately
5. displays a thorough knowledge of curriculum and subject matter
6. insures student time on task
7. implements discipline management procedures
8. demonstrates sensitivity in relating to students
9. demonstrates effective interpersonal relationships with adults
10. fulfills employee responsibilities (p. 26).

Instrumentation

After defining the broad areas of teacher effectiveness, researchers have sought ways to incorporate them into evaluation instruments. Early models made extensive use of numeric rating scales for assessing each criterion, but there is a preponderance of research suggesting that using a graphic response mode to rate specific behavioral descriptors produces ratings that have greater validity and reliability (McNeil and Popham, 1973; Borman, 1977; Saal, Downey and Lahey, 1980; Wexley and Yukl, 1984). A study by Hoffman (1986) found that raters made more valid assessments of teacher competence if they first were required to rate each of the indicators for a criterion prior to rating the performance area as a whole.

Behaviorally Anchored Rating Scales (BARS) are systems developed in the early 1960s by researchers and practitioners in business and industry to provide evaluators with a low-inference tool for observing and

assessing the performance of a worker (Landy and Farr, 1983). Evaluation systems in education have frequently adopted the use of BARS in designing evaluation instruments, among them the University of Washington Teacher Assessment Center (Beal, Foster, and Olstad, 1985). For example, under the general criterion, "uses instructional time efficiently," the following descriptors are used to provide the evaluator with some "anchors" for making a judgment relating to quality level of a person's work:

1. instructional activities begin promptly
2. lesson transitions are made smoothly
3. there are no meaningless digressions
4. instruction continues until the end of the period (p. 2).

As helpful as these descriptors are, however, they still lack specificity in giving the observer/evaluator specific behaviors attached to specific response modes based on quality. The Dallas Independent School District Teacher Appraisal Handbook (1985) took the BARS approach even a step further by using specific descriptors under five levels of quality as illustrated by this example from the criterion, "implements the lesson":

1. Unsatisfactory--does not involve all students in class activities.
2. Below expectations--involves only high achieving students in class activities.
3. Satisfactory--involves all students in class activities.
4. Exceeds expectations--involves all students by using techniques which check for understanding.
5. Clearly outstanding--involves all students within a class period by using a variety of activities (p. 28).

Holdzkom (1987) described the State of North Carolina's teacher evaluation instrument and its approach to proving key anchor words in the

rating scale. In this scale there are six levels of quality, those being (1) unsatisfactory, (2) below standard, (3) at standard, (4) above standard, (5) well-above standard, and (6) superior. For the rating of "well-above standard," the teacher behavior should meet this language:

Performance within this function area is frequently outstanding. Some teaching practices are at the highest level, while others are at a consistently high level. Teacher frequently seeks to expand scope of competencies and often undertakes additional, appropriate responsibilities (Holdzkom, 1987, p. 43).

In addition to studies of the descriptive language for work performance, a number of studies have focused on the effect of the number of rating categories. Landy and Farr (1980) undertook an extensive review of the literature on this subject and cited evidence that an excessive number of categories can have a negative effect on the reliability of the ratings. They summarized their review by indicating that Miller's often-cited "seven, plus or minus two" dictum (Miller, 1956) continued to be the best guideline for selecting the number of response categories for rating the performance of workers.

Implementation strategies

In addition to having criteria that accurately reflect the research on effective teaching, and to having a response mode with anchors for specific behaviors, schools interested in pursuing performance-based evaluation systems must consider carefully a number of other strategies for implementing the system. There is evidence to suggest that the design of the system, and the manner in which it is implemented, are as important as any statistical measure in determining the "validity" of the system

(National Study of School Evaluation, 1984). Savage (1982) maintained that the development of a wholesome climate for professional growth through teacher evaluation is more important in the long run than the technical excellence of any form or procedure used. Key questions related to developing this positive climate through program implementation include (1) who should do the performance appraisals of teachers? (2) should evaluators have any special type of training? (3) what types of data should be collected in the evaluation process? and (4) how should the results of appraisals be fed back to teachers?

Much has been written on the issue of who evaluates. Traditionally, performance appraisals, regardless of their intended use, have been made only by an employee's direct supervisor. Devries et al. (1981), in an exhaustive review of the literature, showed that in 93 percent of the systems studied in business and industry, the employee's immediate supervisor took the sole responsibility for doing the performance appraisal. Similar practices have been noted in teacher evaluation, with a teacher's building principal usually being the sole person responsible for rating a teacher's performance (Grossnickle and Cutter, 1984; Duke and Stiggins, 1986). Duckett (1985), however, noted that there are numerous people who evaluate, or contribute to evaluation, of teachers, those being students, parents, peers, building level administrators, central administrative staff, and community members.

Collection of student input is increasingly regarded as a valuable source of data in the implementation of successful teacher evaluation systems. Student ratings have been used most frequently at institutions

of higher learning, and in that context have been studied by a number of researchers. Doyle (1983) cited studies indicating that student evaluations of their instructor were highly reliable with coefficients in the .80s and .90s, and were consistent across items used on the evaluation instrument. The use of student ratings in elementary and secondary school settings has not, however, been implemented as a major source of evaluation data, and therefore generalizations cannot be made at those levels.

The major reasons for the lack of use of student evaluations were summarized in Successful Teacher Evaluation:

While attitudes regarding the value of student ratings vary, the average elementary and secondary teacher is uncomfortable with the concept. Teachers generally lack faith in the student's ability to accurately rate their performance. In many respects their fears are justified. There is not a great deal of support for the accuracy of student ratings, and the support that does exist is not strong enough to justify using student ratings in any summative evaluation sense (McGreal, 1983, p. 134).

There is support, however, for allowing the student to give the teacher feedback on his or her perception of life in the classroom. A student's degree of agreement or disagreement with the statement, "I feel my ideas are important in this class," can be rated more accurately by the student, and be accepted more readily by the teacher, than a response to the statement, "the teacher knows the subject matter" (McGreal, 1983). In this respect then, Savage (1982) believes that student perceptions can be an important "artifact of teaching."

The lack of adequate instruments to gather valid and reliable information from students has been a major roadblock preventing widespread

use of student evaluations. However, the work of Judkins (1987) is significant for its creation and validation of student evaluation instruments based upon the reading level of the students. Over 3,500 students participated in the study that resulted in separate instruments being validated for use at the K-2, 3-6, 7-8, and 9-12 grade levels.

Self-ratings are frequently used in systems whose sole purpose is employee goal setting and improvement, but are seldom used as a source for arriving at an employee's summative evaluation, especially in career ladder/merit pay systems. Reasons for this omission include the practical and legal limitations of having one's own judgments used as a source for assigning pay differential, as well as the questionable nature of these data relating to statistical error.

Fletcher and Williams (1985) cited the well-known tendency of self-appraisals to suffer from leniency. The authors noted a General Electric Company study in which, when asked to compare their performance with that of others in the company, each individual felt he or she was performing better than three-quarters of his or her peers. Similar studies within the field of education are few, but one of those studies (Noriega, 1987) found in a study of the characteristics of "high gain" teachers that these teachers on the average rated themselves higher than their supervisors (principals) rated them on 18 of 25 effective teaching criteria.

While the use and acceptance of student and self-ratings is very questionable in the literature relating to implementing evaluation systems, studies of the role of peers in the evaluation process are more

numerous. A number of authors have recommended that school districts use multiple appraisers in the implementation of evaluation systems, regardless of whether the system is geared for career ladder/merit pay purposes, or is used solely for the purpose of feedback and goal setting (Dornbusch, 1976; Brophy, 1979; Cruickshank and Applegate, 1981; Bell, 1983; Hopfengardner and Walker, 1984). Ellis (1979) reported studies showing that teachers likened observations and evaluations by supervisors to "fire drills," whereas teachers were more likely to perceive feedback from peers as being genuine and meriting serious consideration.

In addition to being accepted with greater credibility, peer ratings also get high marks from researchers for both validity and reliability. Latham and Wexley (1981) undertook an extensive review of the literature on the topic and noted numerous studies in which the validity and reliability of peer ratings exceeded that of either subordinate or supervisor ratings. The ability of the peer to see an employee's total job performance, and not just a portion of it (as occurs with most supervisory observations) was noted by the authors as a contributing factor to the validity of peer ratings. Bernardin and Beatty (1984) argued that ratings from any single evaluator are less preferable to averaging the ratings of evaluators from different levels in the organization, including peers.

Recent studies on the actual impact and acceptance of peer evaluation in education are few despite the current popularity of recommending it as part of teacher performance appraisal systems. Those studies that do exist are mixed relating to teacher acceptance of peer review. Some

school districts have made a concerted effort to support the peer review concept, with Detroit and Salt Lake City being two examples of large school systems in which the approach was reported to be received favorably by teachers (Benzley, Kauchak, and Peterson, 1985; Sofer, 1985). The Northfield, Minnesota school district is a smaller system which successfully implemented and refined peer review in its evaluation system under the guidance of the School Improvement Model at Iowa State University between 1980 and 1983 (Northfield Public Schools, 1983). Lampesis (1984) documented similar success with a peer review model at Richland Northeast High School in Columbia, South Carolina.

Implementation of some form of peer evaluation as part of a teacher's overall rating for career ladder/merit pay advancement has also occurred in the state-wide plans adopted by Texas (Dallas Independent School District, 1985) and Tennessee (Furtwengler, 1987).

Despite these seemingly promising prospects for peer review, problems in implementation do exist. First, there is a well-known tendency for peers to rate each other higher on the average than supervisors would rate an employee (Doyle, 1983). Lieberman (1985) also attacked the peer review model, citing its failure within institutions of higher learning and the susceptibility of peer ratings to biases held by the rater.

There are also political obstacles preventing universal implementation of peer review. Unions have traditionally opposed performance appraisal systems even without peer review (Wexley and Yukl, 1984), and peer review would seriously conflict with the union's basic tenet of promoting the good of all workers and not pitting members against

each other (Lieberman, 1985). McFaul and Cooper (1984) argued that the peer review model is not viable due to the context in which it often occurs, that being in large urban schools:

...the needs of the peer supervision model for collegiality and trust are incongruent with the prevailing isolation, fragmentation and hierarchial power structure found in urban schools (p. 7).

Goldsberry (1984) and Krajewski (1984) both disagreed with this assessment by indicating that too few studies of the effectiveness of peers existed to make such a generalization. McGreal (1983), however, cited several studies of peer evaluation programs that have been opposed by teachers who see this process as essentially a "popularity contest," thus producing unreliable and invalid results.

While disagreement exists over the use of peers as evaluators, there is no similar controversy in the literature regarding the value of training evaluators. There is much evidence to suggest that teacher evaluators do not automatically become good evaluators just by virtue of their position, and that all evaluators benefit by training (Bolton, 1980). Lefton et al. (1977) emphasized the importance of training by stating that:

...effective appraisers are made, not born; they're effective because they've learned how to be. Many superiors admit that they don't do performance appraisal because they don't know how. They're probably right. All too many appraisals are messed up by 'appraisers' who know little or nothing about appraising (p. 4).

This training should occur for all appraisers prior to an evaluation system being formally adopted and fully implemented by a school district

(Sweeney and Stow, 1981; McGreal, 1983). This training should include not only information about the purposes and goals of the system, but also substantive skills such as data collection, methods of observation, data analysis, report writing, and teacher remediation techniques (Conley, 1987). When conducted in a systematic fashion, training programs for evaluators have been shown to help reduce certain common rating errors, in particular the tendency to rate employees more leniently than their actual performance would indicate (McIntire, Smith, and Hassett, 1984; Pulakos, 1984), and to increase both the validity and reliability of the evaluator's ratings (Savage, 1983; Wexley and Yukl, 1984; Beebe, 1987).

Studies have also shown that specific training programs were effective in helping teacher evaluators gather meaningful data from classroom observations (Semones, 1987) and in using that data effectively in feedback sessions with the teacher (Faast, 1982). Special training for teacher evaluators has not normally been found in traditional administrator certification programs, but some state level initiatives have emerged to require such training. Recent legislation in Iowa (State of Iowa, 1987) requires all teacher evaluators at the K-12 as well as community college levels to undergo a 30 clock-hour evaluator training program by January of 1989. Also, a recently formed cooperative venture between the Arizona School Administrators Association and Wichita State University is one of a growing number of examples of programs focusing on the training of teacher evaluators (McIntire, Hughes, and Burry, 1987).

In order to train evaluators successfully, it first must be decided what types of data will be collected and analyzed in the process of rating

teachers. Data from classroom observations are the most frequently used source of information, with many different instruments having been created and used in recent years for collecting and coding observation data (Acheson and Gall, 1980). The major trends in types of classroom observation data gathering in the past 25 years include analysis of the interaction between students and teachers, teacher self-analysis and clinical supervision, scripting techniques, and structured checklists (Semones, 1987).

Although classroom observation data from peers, students, and/or supervisors form the majority of the total information on which teachers have been evaluated, a number of other sources of data exist. Authors have referred to "artifacts of teaching" (Savage, 1982; McGreal, 1983), among them lesson plans, tests, reading lists, course outlines, and samples of students' work. These items are especially helpful in making an accurate assessment in areas of teaching competence that are less likely to be observed during a typical instructional episode. Lesson plans, for example, can be an important artifact to assist the evaluator in making a valid assessment of the teacher's planning and organizational skills (Manatt and Stow, 1984). Duke and Stiggins (1986) also contended that the examination of teacher-made tests as an artifact of teaching is a way for the evaluator to determine the degree to which the teacher has linked instruction to assessment.

One other important source of data for use in making evaluations of teacher performance is student achievement. There appears to be a general consensus on the value of collecting student performance data (McGreal,

1983), but studies of the actual linkage of teacher evaluations with student outputs are few. Standardized tests provide one easily accessible source of student achievement data, but the use of these results to form a judgment of teacher competence has both practical and political limitations (Glass, 1974). These problems are illustrated by the pending legal challenge by the St. Louis, Missouri Teachers' Union to that district's use of student scores on the California Achievement Test as one of the measures for evaluating teachers (Rothman, 1987).

The results of teacher-made tests provide a more useful approach both to measuring the teacher's effectiveness and to validating the curriculum (Beebe, 1987). The School Improvement Model Project (Manatt, 1987) studied the relationship between student gain scores on standardized tests and different staff development programs for teachers. A follow-up study by Noriega (1987) analyzed the characteristics of teachers whose students had higher than average gain scores on standardized tests. It is clear from these and other studies that the use of student achievement data, while being an important measure of teacher effectiveness, has been approached with caution by all planners of evaluation systems, and implemented by few local school districts.

Finally, schools need to include in their implementation of appraisal systems an assurance that teachers will receive feedback on a regular basis from the evaluator. Studies are numerous which suggest that immediate and direct feedback from the appraiser to the employee is important both for promoting the validity and reliability of the data and for fostering a climate that is conducive to improvement on the part of

the person being evaluated (Oliver, 1983; Chirnside, 1984). Frequent feedback sessions are also necessary in the process of coaching employees as they implement improvement targets (Manatt, Palmer, and Hidlebaugh, 1976; Fournies, 1978). Wexley and Yukl (1984) reported, however, that frequent and direct feedback is seldom received by those being evaluated. Other studies suggest that both supervisors and evaluatees do not look forward to these appraisal sessions, and that negative outcomes for the appraiser can occur from the giving of honest feedback.

In an extensive discussion on the topic of feedback, Fletcher and Williams (1985) cited several conditions necessary for the implementation of a constructive feedback system, among them:

1. The amount of feedback. Most appraisees appear to be able to deal constructively with two aspects of their performance, but not with more than that in any one appraisal session.
2. Positive feedback. Any criticism should be balanced with reinforcement for positive teacher actions.
3. Focus on performance, not the person. Appraisees are much more willing and able to deal with their actions than with matters relating to their personal characteristics (p. 102).

In their work Teacher Evaluation: Five Keys to Growth, Duke and Stiggins (1986) also suggested that feedback sessions are enhanced if:

1. The supervisor uses specific data and shares that data openly with the teacher in the feedback session.
2. The supervisor links the feedback with prespecified performance standards.
3. The frequency of feedback is sufficient to encourage continued development by the teacher (p. 32).

Human Aspects of Evaluation

While much attention has been focused on the technical aspects of performance appraisal, comparatively few studies exist in the field of education relating to the human factors that affect the quality of appraisal ratings (Ilgen, 1983). Those studies that do exist acknowledge that judging human performance is ultimately an activity based upon a certain amount of subjectivity. David Berliner described the process of appraising teachers in relationship to judging other activities:

Judging teaching is absolutely no different from judging figure skating, poultry, potatoes or cows. Each involves making complex decisions with a good deal of subjectivity (Brandt, 1986, p. 6).

It is this subjectivity, according to Henderson (1984), that produces concern among those being evaluated:

What worries the ratee is that the rater will not measure his or her performance on the actual behaviors demonstrated and results achieved during the rating period, but will instead use a variety of subjective biases to rate performance. In other words the actual rating may be based more on the sex, race, national origin, age or religion of the ratee, or on performance in some past appraisal period, or even on physical or psychological makeup (p. 3).

It is important that designers of teacher performance appraisal systems, especially in career ladder/merit pay systems, understand the research related to well-known human errors in the rating process. These errors, then, can be minimized through designing better instruments, giving raters special training, and motivating raters to appraise accurately (Wexley and Yukl, 1984).

Five different categories of human errors have been selected for analysis, those being (1) leniency/severity/central tendency, (2) the halo effect, (3) rater characteristics, (4) rater position in the organization, and (5) personal bias.

Leniency/severity/central tendency

The definition of "average" performance is subject to human error any time a rating scale is used. Studies exist in which the raters failed to differentiate among levels of performance of ratees by clustering their scores within a narrow range, otherwise known as central tendency error. Doyle (1983) suggested that this error occurs for raters who have an inclination to avoid any extreme, either high or low, on the rating scale. When this error occurs with regularity, all employees rated appear to be "average," thus preventing the appraisal system from differentiating among levels of performance.

Leniency error, and its opposite effect, severity error, have been the subject of many studies. Severity occurs when appraisers concentrate their judgments at the low end of a rating scale, and leniency describes the tendency for the appraiser to rate well above the midpoint of a scale (Saal, Downey, and Lahey, 1980). Numerous studies have shown that, when rating scales are used, a tendency toward leniency exists (Devries et al., 1981; Doyle, 1983; Henderson, 1984; Pulakos, 1984). Henderson (1984) cited the common tendency toward leniency in military personnel ratings, with a typical finding being 95 percent of a unit's officers being rated in a category identified to include only the top five percent.

There is evidence that the purpose of the ratings has an effect on leniency. In the military example cited above, officers are rarely promoted if they are not ranked in the top five percent by their superior, thus leading researchers to speculate that raters inflate their ratings if they are used for promotion purposes. Murphey et al. (1984) hypothesized that raters may use one set of standards for judging another's performance if the rating is used for research purposes and another standard (more lenient) if the results are to be used for administrative decisions such as promotion or demotion.

Halo

Halo is the tendency on the part of evaluators to let their rating of specific criteria on the evaluation instrument be unduly influenced by their overall impression of the ratee (Landy and Farr, 1980; Doyle, 1983; Pulakos, 1984). For example, an evaluator who values planning skills may rate a teacher who is proficient in those skills high on all other criteria as well regardless of the teacher's actual skill level. Another common example cited is one in which the employee is well-liked and gets along well with supervisors and peers, and for this reason is rated highly on all evaluation criteria even though the employee does not perform all aspects of the job at a high level.

Wexley and Yukl (1984) found that halo error can be reduced through two strategies, those being (1) having the evaluator rate all employees on a single criterion before moving on to the next criterion, and not looking back at ratings assigned previously to the teacher, and (2) making the

rating scale benchmarks more specific by using Behaviorally Anchored Rating Scales (BARS), as discussed previously in this chapter.

Rater characteristics

There is some evidence that certain traits possessed by the rater can influence the accuracy of the ratings. Wexley and Yukl (1984) reported that supervisors who are more competent in their own jobs are less likely to produce ratings that have leniency error. Also, supervisors who are more task/production oriented are less lenient in their ratings than those who are primarily oriented to employee relations (Landy and Farr, 1980). The rater characteristic of gender, however, has been studied more frequently than any other rater characteristic. A number of studies indicate that, for the most part, neither the gender of the rater or ratee affects ratings (Nieva and Gutek, 1980; Mobley, 1982; Wexley and Pulakos, 1983; Etaugh and Foresman, 1983; Terborg and Shingledecker, 1983). Landy and Farr (1983) summarized 14 studies, most of them from laboratory or simulation experiments, and all occurring since 1970, which cited similar findings. The only trend noted among some of those studies was a tendency for female raters to assign more lenient ratings than their male counterparts, a tendency noted also by Carroll (1982). Harrington (1984), on the other hand, found that females gave lower ratings than male evaluators when assessing the performance level of a teacher's video-taped lesson.

Concerning an interaction effect between the gender of the rater and the gender of the ratee, there is evidence that, regardless of the gender of the rater, female ratees tended to be rated lower than males who

performed the same type of work (Decotlis and Petit, 1978; Carroll, 1982). These studies, like those previously mentioned, were also in laboratory or simulation settings, and Landy and Farr (1980, 1983) reported that there were no known studies in which the rater and ratee were both actually employees of an organization.

The rater's education and experience have also been studied, but with less frequency than for the effects of gender. Raters with more education and/or experience tend to produce more accurate ratings, according to Landy and Farr (1983).

Rater position

Earlier in this review, references were made to studies of the merits of using peers as evaluators, with a notation that peers tend to produce ratings more lenient than supervisors (Doyle, 1983). Leniency aside, however, Landy and Farr (1983) indicated that no one type of rater appears to be more valid than any other type. Wexley and Yukl (1984), however, contended that the more distant an evaluator is in the organizational structure, the less lenient the ratings tend to be. This finding, along with others, provides support for the use of peers in some aspect of the rating process in teacher performance appraisal systems.

Personal bias

Factors such as an employee's physical attractiveness, race, ethnic background, social standing in the community, personality, preferred teaching style, and other such attributes can distort a rater's evaluations. Some authors refer to these as the "same as me" or "like me"

biases (Henderson, 1984), and the most frequently studied of these is race. In light of equal employment opportunity legislation and the increasing number of minorities in management positions, the impact of rater and ratee race on ratings is of considerable interest to those designing performance appraisal systems for teachers, especially in career ladder/merit pay circumstances (Carroll, 1982).

No studies are known to exist that examine the effect of race within the context of a career ladder/merit pay teacher performance appraisal system. However, studies from other employee evaluation settings suggest that race does indeed produce an effect on ratings. Generally, the literature suggests that blacks will be evaluated less favorably than whites, and that an interaction effect occurs to produce higher ratings for an employee of the same race as the appraiser (Decotiis and Petit, 1978; Landy and Farr, 1980, 1983; Carroll, 1982; Mobley, 1982).

Studies of bias due to an evaluator's preferred teaching style are few. Rucker (1981), however, was able to study the interaction effect between the preferred teaching style of a group of principals with the teachers they evaluated. He hypothesized that those teachers who shared a common style preference with the principal would receive higher ratings. However, no significant differences were found, suggesting that an evaluator's preference for a particular teaching style does not act as a source of bias in the evaluation process.

Summary

Few issues in education are more potentially explosive than teacher evaluation. Most everyone agrees that appraisal of teachers is a

necessary function, but there are vast differences of opinion about the intended purposes of the evaluations, and about the correct procedures to implement teacher performance appraisal systems.

Common wisdom suggests that if the designers of appraisal systems are somehow able to create a foolproof system containing instruments, procedures, and training programs that will always produce valid results, then educators will be able to construct a formula for improving the learning potential of individual students and for improving the effectiveness of schools in general.

However, the use of teacher evaluation as a tool to accomplish these goals brings with it some inherent limitations. The appraisal of human performance is highly susceptible to error based upon factors not directly related to the actual quality level of performance. Through understanding these human factors, and through designing sound appraisal systems that lessen the opportunity for human error, performance appraisal systems that are reliable and free from bias can be constructed. The development of these systems can be the first step toward having an appraisal system that not only lets teachers know where they stand but also allows a district to move toward differentiating its compensation system to account for varying levels of excellence among its teaching staff.

CHAPTER III. METHODS AND PROCEDURES

Introduction

The central purpose of this study was to determine whether or not a performance appraisal system produced results that provided teachers with fair treatment through equal access to the benefits provided by a career ladder advancement system. In particular it was the intent of this study to determine if certain characteristics of the rater, either singly or in combination with certain characteristics of the ratee, had a negative impact on the reliability of the appraisal ratings.

Another related purpose was to analyze the appraisal instrument used by the school district in this study. In particular, the ability of the instrument to produce reliable results among the various criteria was deemed important for investigation in order to control for the effect of the instrument while studying the effects of other variables.

Finally, the ability of a single evaluator to make consistent ratings over time for teachers, and the ability of multiple appraisers to make similar ratings for teachers was of interest to the investigator. More specifically, methods and procedures were developed in this study to answer the following questions:

1. Will there be agreement among the ratings for different criteria used on the appraisal instrument?
2. Will the use of teacher evaluations for promotion in a career ladder system be subject to systematic error due to certain characteristics of the rater?

3. Will teacher evaluation ratings be subject to systematic error due to an interaction effect between certain characteristics of the rater and the ratee?

4. Will there be agreement between appraisal ratings assigned by two different evaluators for a common group of teachers?

5. Can evaluators make consistent ratings on repeated measures of a teacher's performance?

Identification of Research Subjects

The questions posed in this study were best studied through the use of actual teacher evaluation data gathered from raters who understood that (1) the purpose of the ratings was to differentiate among levels of teacher performance, and (2) the results of the ratings would constitute the basis for differentiation in pay for teachers. Therefore, it was necessary to identify a school district which not only was implementing such a pay-for-performance system, but one which also contained the diversity among its rater and teacher population to test all of the hypotheses selected for study.

The Dallas, Texas, Independent School District (DISD) was selected on the basis of its meeting these criteria. This school district sought the services of the School Improvement Model (SIM) at Iowa State University in 1985 to assist in developing an appraisal system to meet the intent of state legislation requiring implementation of a career ladder advancement system for teachers in the State of Texas (Dallas Independent School District, 1985).

During the 1985-86 school year, SIM co-directors Richard P. Manatt and Shirley B. Stow, as well as other members of the SIM staff, provided on-going training sessions in order to improve the skills of evaluators in the school district. This training totaled 48 clock hours and covered topics designed to promote an understanding of the elements of effective teaching and a working knowledge of district evaluation procedures. Through the course of the 1985-86 school year, the data necessary to study the hypotheses in this study became available first to the SIM training staff and later to this investigator through the Department of Human Development and Training in the Dallas Independent School District.

Collection of Data

Data used in this study were collected from formative appraisals completed by evaluators during the 1985-86 school year. Figure 1 shows the Written Record of Observation used by these evaluators. A copy of each completed appraisal form was forwarded to the SIM office at Iowa State University after first being sent by each evaluator to central office staff in DISD. Approximately 34,000 of these completed forms were obtained by this investigator, and it is from this data base that representative samples were drawn to test the hypotheses in this study.

Additionally, it was necessary to obtain information about the demographic characteristics of the evaluators and teachers that relate to the hypotheses in this study. Information about the evaluators' gender, position, level of assignment, education level, years of experience, and race was collected as part of the Dallas Independent School District Test

**WRITTEN RECORD OF OBSERVATION
FORMATIVE APPRAISAL**

The Written Record of Observation (the Formative Appraisal) will be completed by each appraiser after the required formal observations, a minimum of two per year. The appraisers will jointly summarize each of the Individual Written Record of Observation reports into one Written Record of Observation report. The purpose of the formative appraisal is to provide suggestions and recommendations for improvement. Formative appraisals are not cumulative and are not the final evaluation (summative). Additional formative appraisals may be conducted during the year by the principal or designee.

Employee's Name _____ SS# _____

Last _____ First _____ M.I. _____

Teaching Assignment _____ School _____

Years in District _____ Years at this school _____

Principal _____

Appraiser's Name _____

Appraiser's Title and Assignment _____

Rating for each Criterion (O,E,S,B,U)

1. ____ THE TEACHER DEMONSTRATES EFFECTIVE PLANNING SKILLS
2. ____ THE TEACHER IMPLEMENTS THE LESSON PLAN
3. ____ THE TEACHER COMMUNICATES EFFECTIVELY WITH STUDENTS
4. ____ THE TEACHER USES EVALUATION ACTIVITIES APPROPRIATELY
5. ____ THE TEACHER DISPLAYS A THOROUGH KNOWLEDGE OF CURRICULUM AND SUBJECT MATTER
6. ____ THE TEACHER ENSURES STUDENT TIME ON TASK
7. ____ THE TEACHER IMPLEMENTS DISCIPLINE MANAGEMENT PROCEDURES
8. ____ THE TEACHER DEMONSTRATES SENSITIVITY IN RELATING TO STUDENTS
9. ____ THE TEACHER DEMONSTRATES EFFECTIVE INTERPERSONAL RELATIONSHIPS WITH ADULTS

Date of appraisal _____ Appraiser's Signature _____

COMPLETE THIS SECTION IF THIS IS THE SUMMARIZED RECORD OF BOTH APPRAISALS.

Date of conference _____ Conference conducted by _____

Signature of Appraiser _____

Second Appraiser's Signature _____

Teacher's Signature _____

Figure 1. Teacher performance appraisal instrument (formative),
Dallas Independent School District, 1985-86

for Evaluators, an instrument used to determine the extent to which evaluators understood instructional techniques and district evaluation procedures (see Appendix B). Information pertaining to the gender and race of each teacher in the district was obtained from DISD central office staff. Identification numbers were assigned to each evaluator and teacher to insure the anonymity of each subject.

Human Subjects Release

The Iowa State University Committee on the Use of Human Subjects in Research reviewed this project and concluded that the rights and welfare of the human subjects were adequately protected, that risks were outweighed by the potential benefits and expected value of the knowledge sought, that confidentiality of data was assured, and that informed consent was obtained by appropriate procedures.

Methods of Statistical Treatment

Prior to performing statistical tests for each research hypothesis, it was first necessary to develop an appropriate data base from which these tests could be performed. The first step in this process consisted of converting the responses for each of the criteria on the Written Record of Observation from a graphic response mode to a numeric mode. The following procedure was used to make this conversion for each of the criteria rated:

<u>Graphic Response</u>	<u>Numeric Equivalent</u>
Clearly Outstanding	5
Exceeds Expectations	4
Satisfactory	3
Below Expectations	2
Unsatisfactory	1

After this conversion was accomplished, the data were entered, along with other demographic data pertaining to the evaluators and teachers, into the computer by computation center staff at Iowa State University. The revised Statistical Package for the Social Sciences, SPSS^x (Norvvis, 1983) was used to test a number of hypotheses, and other hypotheses using smaller samples were conducted using Introductory Statistics: A Microcomputer Approach (Elsley, 1985).

Sampling procedures

The data base produced through the above-mentioned procedures contained all of the variables necessary to answer the questions posed by this study. However, the data also contained several other variables, some of which not only were outside of the scope of this study but also which posed the potential for contaminating the results of the questions in this investigation. For example, one important factor that could influence evaluation ratings is the context in which it takes place (Joint Committee on Standards for Education Evaluation, pending). Several differences in context existed among the instructional settings of elementary, middle and high schools in DISD, with many more specialized programs existing at the middle and high school levels. Teachers at those levels were generally organized by subject matter taught, with each teacher usually being a specialist in his or her subject area. Elementary teachers, however, were subject matter generalists, with self-contained classroom instruction in a number of subject areas being the normal expectation for most teachers. In order to reduce the risk of failing to account for these and other contextual differences in the evaluation

process, a decision was made to answer all of the questions posed by this study through the use of subjects from only one of the three levels of teachers and evaluators in DISD.

The determination of which level to select for study was based on a desire to eliminate another variable, that being the position held by the appraiser. Evaluators at the high school and middle school levels represented a number of different positions, including principals, assistant principals, deans of instruction, department heads, and central office staff. At the elementary level, however, all of the appraisers were principals, with a teacher's building principal serving as the first appraiser and a principal from another elementary school in the district serving as the second appraiser. The elementary level, therefore, was selected for study because its use of only principals as evaluators reduced the potential for differences in appraisers' ratings being attributed to differences in position held.

Other variables were held constant through the use of formative rather than summative evaluation data. First, formative appraisals were made soon after the evaluator's classroom observation of the teacher (within five working days). Therefore, the formative results were seen as being more sensitive to the teacher's actual classroom teaching performance than the summative results. Summative appraisals were designed to be an amalgamation of data from several sources over an extended period of time. Also, the formative results were more appropriate for studying an evaluator's ratings over time, in that formative evaluations were conducted twice whereas summative evaluations

were conducted only once per year. Formative evaluations also provided the ability to study interrater agreement in that two appraisers conducted each formative appraisal while summative appraisals were completed by only one evaluator.

Finally, the difference in format between the formative and summative instrument led to a decision to use only formative ratings. Whereas the formative instrument contained nine criteria (see Appendix B, Written Record of Observation: Formative Appraisal), the summative instrument included one additional criteria, that being "The teacher fulfills employee responsibilities" (see Appendix B, Summative Evaluation Form). This lack of consistency between the instruments prevented a comparative study between the results of the formative and summative phases of the evaluation process.

Questions posed in this study were answered using all of the data for evaluators and teachers at the elementary level, or in some cases were answered using random samples of evaluations submitted for elementary teachers. Random sampling was accomplished in each instance through the use of a computer-generated table of random numbers found in Educational Research: An Introduction (Borg and Gall, 1983).

Statistical analysis

Appropriate research methodologies and statistical tests were selected in order to answer questions posed by this study. The specific means of analysis used to address each question are as follows:

Question 1. Will there be agreement among the ratings for different criteria used on the appraisal instrument?

The method selected for analysis of this question consisted of dividing the appraisal instrument roughly into two halves, with five criteria being considered in one group and four in another. Five of the criteria on the appraisal instrument (see Figure 1) can normally be assessed through observation of teacher behaviors during direct instruction of students, those being:

- Criterion 2. The teacher implements the lesson plan
- Criterion 3. The teacher communicates effectively with students
- Criterion 6. The teacher ensures student time on task
- Criterion 7. The teacher implements discipline management procedures
- Criterion 8. The teacher demonstrates sensitivity in relating to students

The remaining four criteria lend themselves more to the use of artifacts of teaching (lesson plans, tests, etc.) for evaluation purposes.

These criteria are:

- Criterion 1. The teacher demonstrates effective planning skills
- Criterion 4. The teacher uses evaluation activities appropriately
- Criterion 5. The teacher displays a thorough knowledge of curriculum and subject matter
- Criterion 9. The teacher demonstrates effective interpersonal relations with adults

This methodology resulted in a variation of the split-halves method of determining the internal consistency of scores produced by an instrument. This particular method of splitting the instrument, however, also allowed for the investigation of how closely evaluation ratings taken from direct observation of instruction related to scores resulting from examination of lesson plans, tests, worksheets, and other artifacts of teaching.

For this question all of the first semester (formative) evaluations for all of the 3,460 elementary teachers were used for analysis. The

Pearson product-moment procedure was used to develop a correlation coefficient between the scores for each subset of criteria. Also, a correlation coefficient was determined between each subset of criteria and the combined average score for all criteria on the appraisal instrument.

Question 2. Will the use of teacher evaluations for promotion in a career ladder system be subject to systematic error due to certain characteristics of the rater?

The rater characteristics of gender, race, level of training, and years of experience in education were selected for study, with first semester (formative) evaluations by all elementary raters being used to test for differences based upon training and experience. A one-way analysis of variance (ANOVA) was employed to determine differences in mean teacher appraisal scores for these variables, with years of experience divided into six groupings, and training levels divided into five categories. This statistical procedure is roughly equivalent to the student's t-test, but was selected for its ability to make multiple comparisons.

Random samples were drawn from first semester (formative) evaluations at the elementary level to test for differences in mean scores based upon the gender of the appraiser. The two-way analysis of variance (ANOVA) was selected for use because of its ability to make the comparison of differences in score based upon gender of the appraiser (the main effect), and for its ability to answer a subsequent question related to an interaction effect between the gender of the evaluator and teacher.

Similar procedures were used to determine differences in mean appraisal scores based upon the race of the appraiser. Random sampling

procedures were used, followed by application of the two-way analysis of variance (ANOVA) to determine if the race of the appraiser (the main effect) produced differences in mean scores.

Question 3. Will teacher evaluation ratings be subject to systematic error due to an interaction effect between characteristics of the rater and ratee?

The rater-ratee characteristics of gender and race were selected for study, and random sampling procedures were employed among evaluators and teachers at the elementary level to determine the subjects for data analysis. The two-way analysis of variance (ANOVA) procedure was selected for its ability to detect an interaction effect between rater and ratee characteristics.

Question 4. Will there be agreement between appraisal ratings assigned by two different evaluators for a common group of teachers?

Those subjects selected for use in answering this question included all pairs of elementary appraisers (first and second appraisers) who evaluated a common group of at least 25 teachers during the first semester of the 1985-86 school year. For each of these pairs the Pearson product-moment test was used to develop a measure of inter-rater agreement between appraisal scores assigned by each pair of evaluators. Also, the student's t-test was used to determine the significance of difference between the mean teacher appraisal ratings assigned by each pair of raters.

Question 5. Can evaluators make consistent ratings on repeated measures of a teacher's performance?

Teacher appraisal scores were used for all evaluators (both first and second appraisers) at the elementary level who rated the same group of 25

or more teachers for both first and second semester appraisals. The Pearson product-moment test was used to establish a correlation coefficient between each evaluator's first semester and second semester ratings. This procedure served to establish a measure of intrarater agreement for appraisal scores assigned at two different times by the same appraiser.

CHAPTER IV. ANALYSIS AND RESEARCH FINDINGS

The primary purpose of this study was to examine the actual teacher performance appraisal data from a school district using a career ladder advancement system for teachers, in an effort to determine if those data were free from systematic error. Other purposes addressed by this study included studying inter-rater agreement among multiple appraisers and analyzing intra-rater agreement for a single appraiser's ratings of the same teachers over time. This chapter analyzes the data collected from the subjects of the study, those being the teacher evaluators at the elementary level in the Dallas Independent School District during the 1985-86 school year.

This chapter is divided into two sections, those being (1) descriptive data and (2) hypothesis testing. Descriptive data were compiled from responses to a questionnaire accompanying the DISD Test for Evaluators which was administered in April of 1986 to all teacher evaluators in the Dallas Independent School District. Additional demographic information for teachers was obtained from the DISD Department of Human Development and Training. Data for the inferential statistics were collected from the Written Record of Observation (see Figure 1), the formative teacher appraisal instrument used by DISD during the 1985-86 school year.

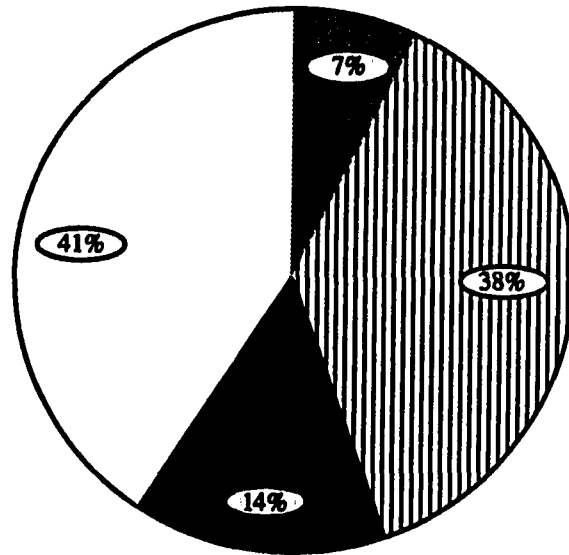
Descriptive Data

Descriptive data, presented in Figures 2 through 7, depict the characteristics of the raters that were selected for analysis in this

study. Figure 2 shows the number and percent of evaluators, by level, who participated in evaluator training during the 1985-86 school year. These evaluators represented the positions of principal, assistant principal, dean of instruction, and department head at the middle and high school levels. At the elementary level a total of 193 evaluators were trained in evaluation techniques, and of that number, 112 (all of them principals) actually evaluated teachers and returned copies of the Written Record of Observation to the SIM office at Iowa State University. Figure 3 shows that the group of elementary teachers selected for analysis in this study comprised the largest number of teachers by level, with a total of 3,460.

Figures 4 through 7 give specific information about the characteristics of the 112 principals at the elementary level who form the data base of evaluators for this study. Figure 4 reveals the distribution of both evaluators and teachers by gender. The majority of the evaluators (70 percent) were males, whereas only 16 percent of the teachers at the elementary level were males. Figure 5 shows the breakdown of evaluators and teachers by race. Evaluators were distributed among blacks (37 percent), whites (42 percent), and Hispanics (21 percent). The majority of the teachers at the elementary level were white (54 percent), with 38 percent being black and eight percent Hispanic.

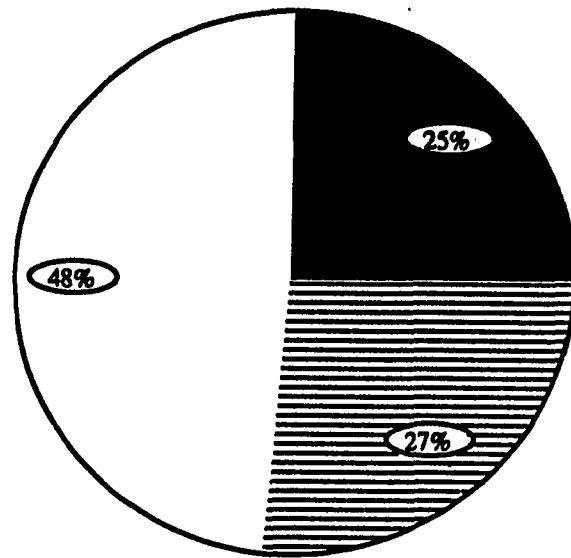
Figure 6 shows that the majority of the group of evaluators were clustered in the top two categories of education levels, with 32 percent of the group possessing a Ph.D. degree and 35 percent having an M.A. + 45 credits.



Evaluators (N=475)

- Elementary School (N=193)
- Middle School (N=65)
- High School (N=180)
- Central Staff (N=37)

Figure 2. Number and percent of teacher evaluators by level, Dallas Independent School District, 1985-86



Teachers (N=7,169)

- Elementary School (N=3,460)
- Middle School (N=1,764)
- High School (N=1,945)

Figure 3. Number and percent of teachers by level, Dallas Independent School District, 1985-86

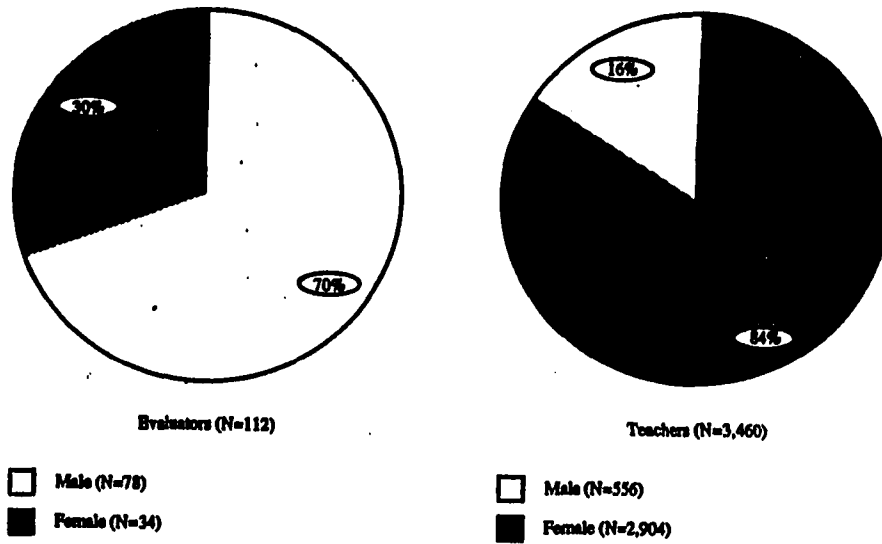


Figure 4. Number and percent of elementary evaluators and teachers by gender, Dallas Independent School District, 1985-86

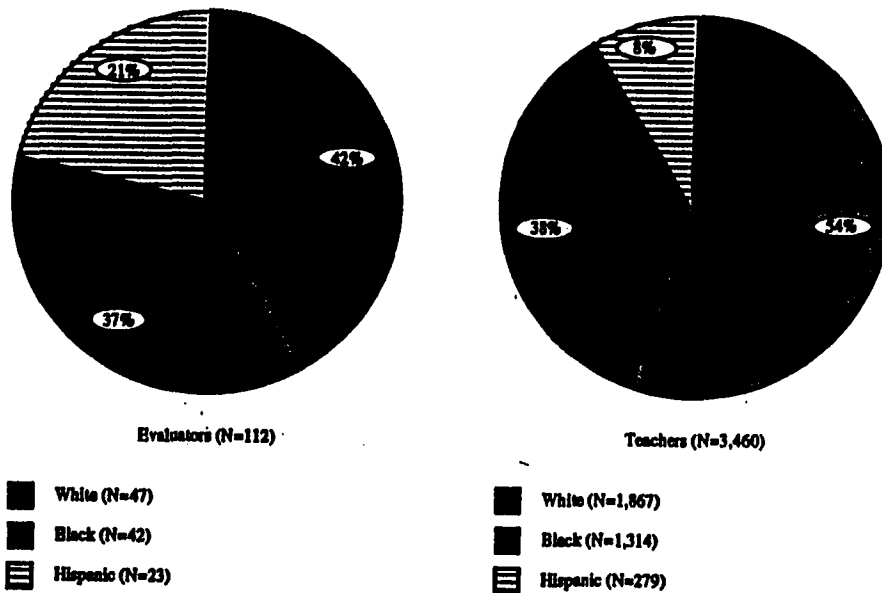
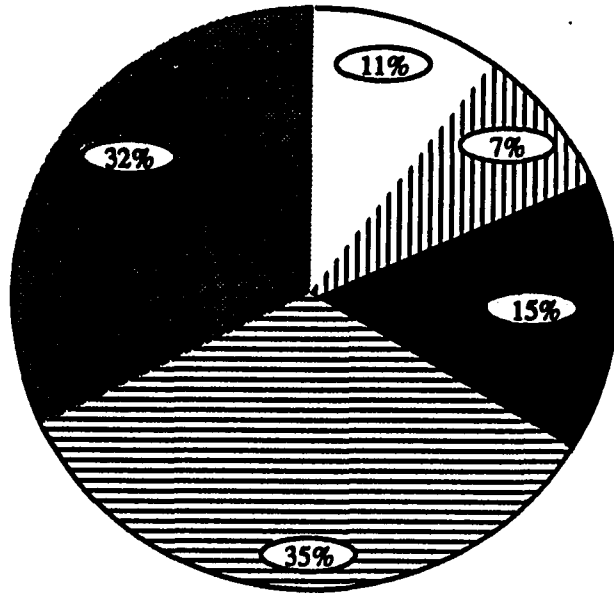


Figure 5. Number and percent of elementary evaluators and teachers by race, Dallas Independent School District, 1985-86



Evaluators' Education Level (N=112)






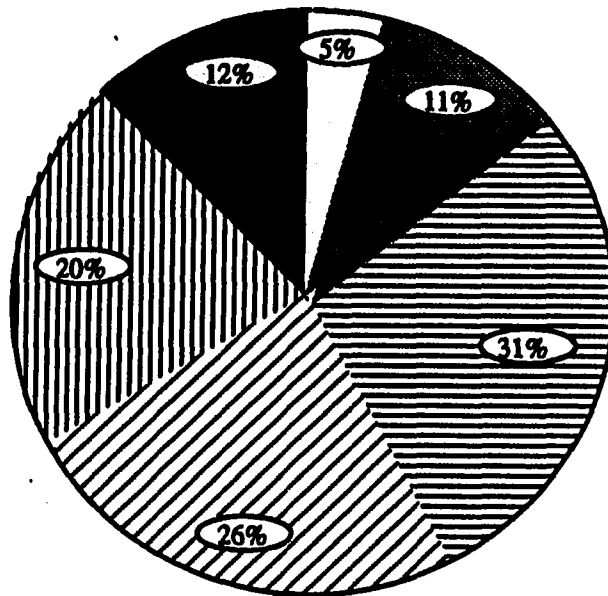
-  MA (N=12)
-  MA + 15 (N=8)
-  MA + 30 (N=17)
-  MA + 45 (N=39)
-  Ph. D (N=36)

Figure 6. Number and percent of elementary evaluators by highest level of education attained, Dallas Independent School District, 1985-86



Evaluators' Experience in Education (N=112)

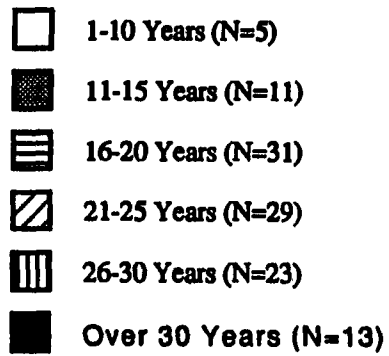


Figure 7. Number and percent of elementary evaluators by total years of experience in education, Dallas Independent School District, 1985-86

The distribution of the group of evaluators by years of experience in education can be noted in Figure 7, with the majority of the evaluators (77 percent) having more than 15 years of experience in education. The single largest group of evaluators (16-20 years of experience) represented 31 percent of the total.

Hypothesis Testing

Each of the questions posed in this study resulted in one or more specific research hypotheses being stated, all of which are stated in the null form. All hypotheses were tested for significance at the .05 level, with all probabilities less than .05 being reported also. Hypotheses are presented and discussed in the order of the questions posed by this study.

Hypothesis 1. There will be no significant positive correlation between an evaluator's ratings for two different subsets of performance criteria on the appraisal instrument.

This hypothesis was formulated to determine the relationship between two sets of scores within the appraisal instrument, thus producing a measure of the instrument's internal consistency. Also, the methodology employed in determining the subsets of criteria allowed for analysis of the relationship between criteria normally judged by an evaluator's collection of data from observation of classroom instruction by the teacher (Criteria 2, 3, 6, 7, and 8) and criteria normally judged by teacher actions occurring outside of the classroom instructional setting (Criteria 1, 4, 5, and 9). A correlation coefficient of .80 or greater was considered significant in testing this hypothesis.

Table 1 shows the mean scores and standard deviations for each of the nine criteria on the appraisal instrument, and Figure 8a depicts the mean

scores graphically. Criterion 5, Knowledge of curriculum and subject matter, received the highest mean score (4.02), with Criterion 4, Evaluation activities, having the lowest mean score (3.73). These calculations were made from first semester formative appraisals completed by 112 appraisers for 3,460 elementary teachers.

Figure 8b depicts the mean scores for the subsets of criteria stated in the hypothesis. The mean score for Group 3 (3.86) represents the average score for all 3,460 teachers taking into account all nine criteria on the instrument. Table 2 reports the correlation coefficients produced through use of the Pearson product-moment test. Correlations were shown

Table 1. Mean scores and standard deviations by criteria for teacher appraisal instrument (N=112 appraisers and 3,460 teachers)

Criteria	Mean	Standard deviation
1. Planning skills	3.94	.73
2. Implements the lesson	3.85	.78
3. Communicates effectively with students	3.97	.69
4. Evaluation activities	3.73	.71
5. Knowledge of curriculum and subject matter	4.02	.72
6. Student time on task	3.95	.71
7. Discipline management	3.88	.73
8. Relating to students	3.95	.70
9. Interpersonal skills with adults	3.80	.73

Criteria

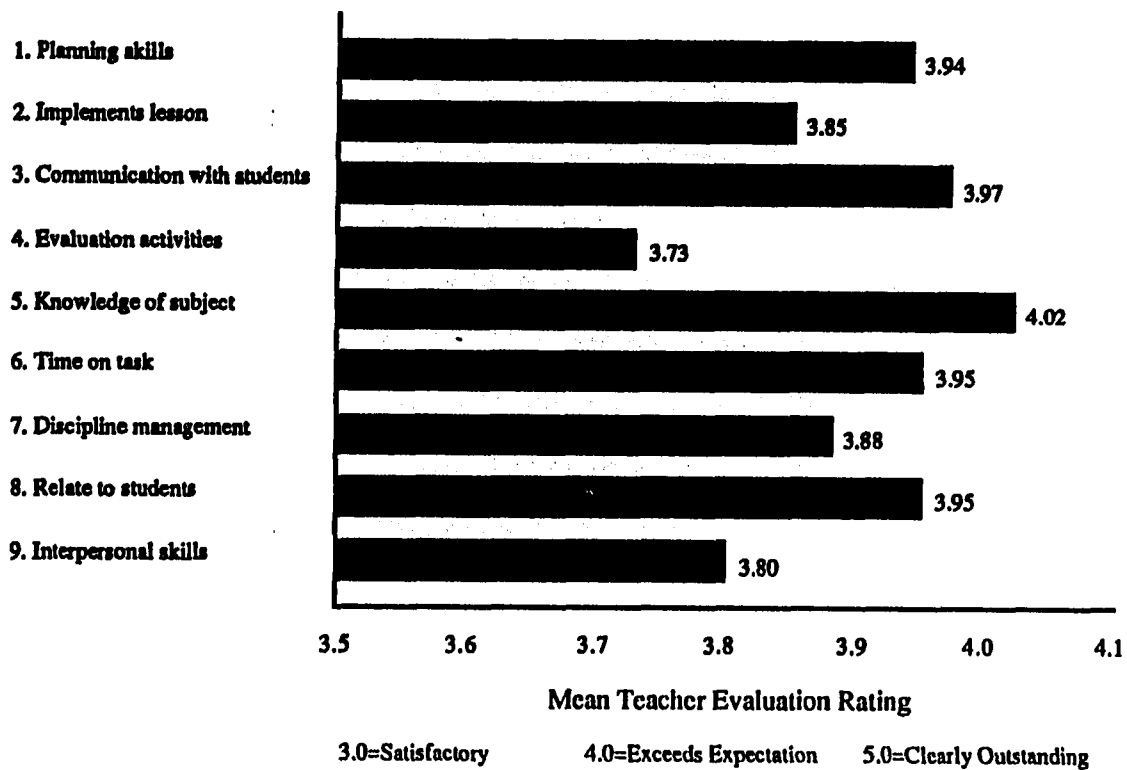


Figure 8a. Mean first semester appraisal scores, by criteria, assigned to 3,460 elementary teachers by 112 evaluators, Dallas Independent School District, 1985-86

Criteria

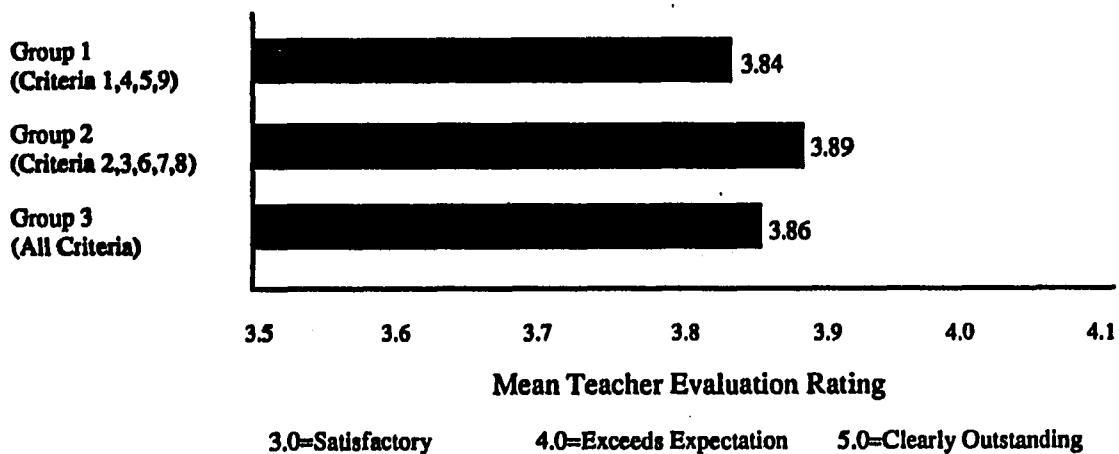


Figure 8b. Mean first semester appraisal scores, by groups of criteria, assigned to 3,460 elementary teachers by 112 evaluators, Dallas Independent School District, 1985-86

to be significant at the .001 level, with Group 1 and Group 2 producing a correlation of .8407. Both subgroups of criteria produced a high correlation with the average scores of the entire instrument, with Group 1 producing a .9476 coefficient and Group 2 producing a .9696 correlation.

Table 2. Pearson product-moment correlation coefficients for subgroups of criteria within teacher evaluation instrument (N=3,460 teachers and 112 appraisers)

Subgroup	Group 1 (Criteria 1,4,5,9)	Group 2 (Criteria 2,3,6,7,8)	Group 3 (All criteria)
Group 1 (Criteria 1,4,5,9)	1.000		
Group 2 (Criteria 2,3,6,7,8)	.8407***	1.000	
Group 3 (All criteria)	.9476***	.9696***	1.000

***Significant at $p < .001$ level.

On the basis of the correlation coefficients being at this high level, coupled with significance levels at .001, the hypothesis of there being no significant positive correlation between subsets of criteria on the appraisal instrument was rejected.

Hypothesis 2. There will be no significant difference in mean teacher evaluation scores based upon the rater characteristics of gender, race, level of training, or years of experience in education.

This hypothesis was written to include those variables most frequently studied by researchers to determine the effects of rater

characteristics on appraisal scores. Each of the variables was isolated and studied using separate statistical tests, with four different sub-hypotheses being stated in the null form.

Hypothesis 2a. There will be no significant difference in mean teacher evaluation scores based upon the gender of the rater.

This hypothesis was tested using formative appraisals submitted by a teacher's building principal for first semester of the 1985-86 school year. Appraisal scores from all of the 34 female principals were used, and a random sample of 34 of the remaining 78 male principals was selected for analysis. For each of the 34 appraisers, two mean teacher appraisal scores, one male and one female, were randomly selected for analysis, resulting in the research design depicted in Figure 9. The effect of race was controlled by having both mean scores selected for each evaluator be from a male and female teacher of the same race.

Figure 10 reveals the mean scores by gender for each of the groups sampled. A two-way analysis of variance (ANOVA) test was applied to determine if the differences in mean scores were significant. On the basis of this analysis, the hypothesis of there being no difference in teacher evaluation scores based upon the gender of the rater was rejected at the .01 level (see Table 3). This was due to the difference between the mean appraisal score of 3.95 for all 68 randomly selected teachers (34 male and 34 female) evaluated by males and the mean score of 3.70 for all 68 randomly selected teachers (34 male and 34 female) evaluated by female appraisers.

Appraisers
(N=68)



Teachers
(N=136)

Figure 9. Research design and number of subjects used to determine differences in teacher performance appraisal ratings based upon gender

Table 3. Analysis of variance of mean elementary teacher appraisal scores by gender

Source	df	Sum of squares	Mean square	F-ratio	F-prob.
Appraisers	1	2.148	2.148	6.777**	.010
Teachers	1	1.103	1.103	3.482	.061
Interaction	1	.081	.081	.254	.990
Error	132	41.827	.317		

**Significant at $p < .01$ level.

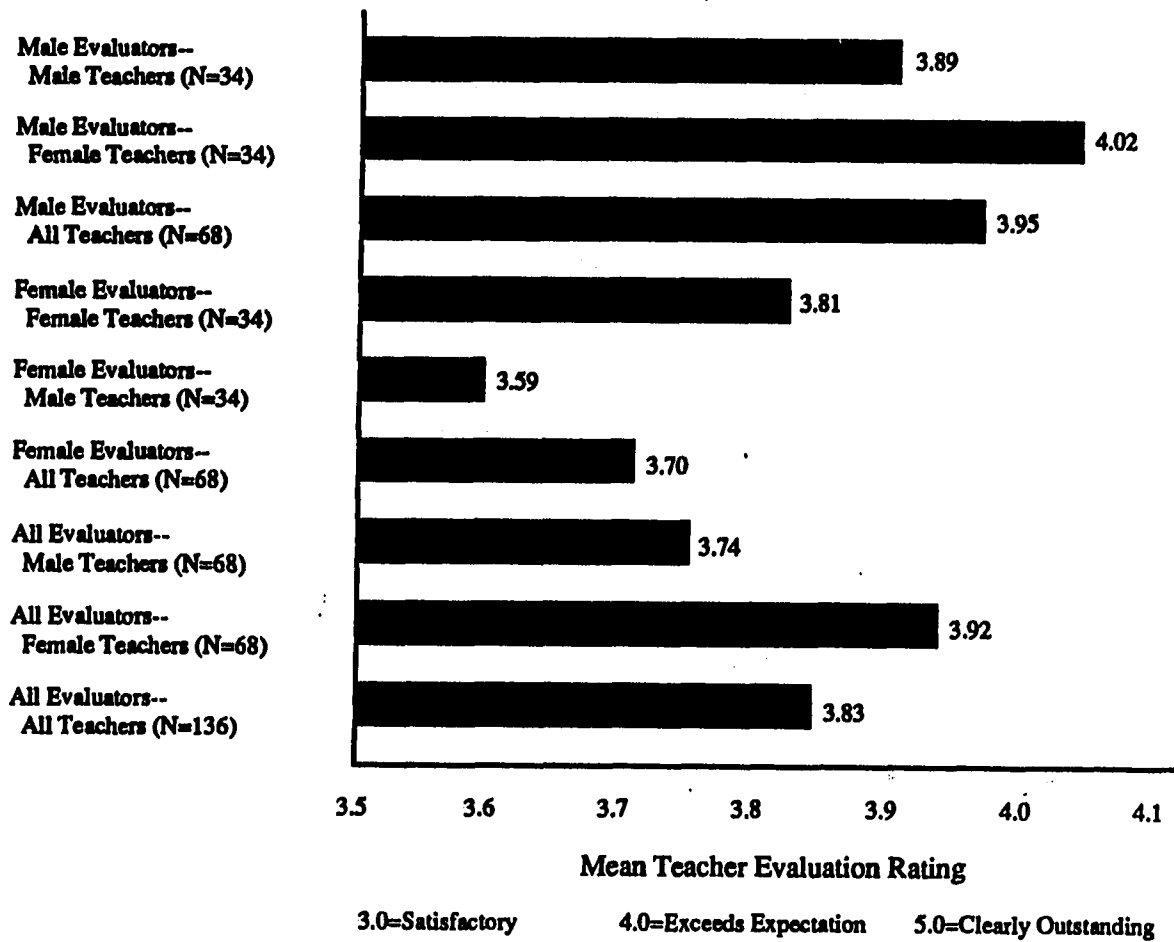
Evaluation Combinations

Figure 10. Mean first semester teacher appraisal scores by gender groups for selected elementary evaluators, Dallas Independent School District, 1985-86

Hypothesis 2b. There will be no significant difference in mean teacher appraisal scores based upon the race of the appraiser.

This hypothesis was tested using formative appraisal data submitted by randomly selected building principals (first appraisers) during the first semester of the 1985-86 school year. To test this hypothesis, random samples of mean teacher appraisal scores were drawn from all 23 Hispanic principals, and random mean scores were drawn from a randomly selected group of 23 of the 42 black principals and from 23 of the 47 white principals. The resulting research design, depicted in Figure 11, called for one random mean teacher appraisal score to be selected for each of three teachers (one Hispanic, one black, and one white) who were evaluated by each principal. The variable of gender was controlled by having all three mean appraisal scores that were selected come from teachers of the same gender.

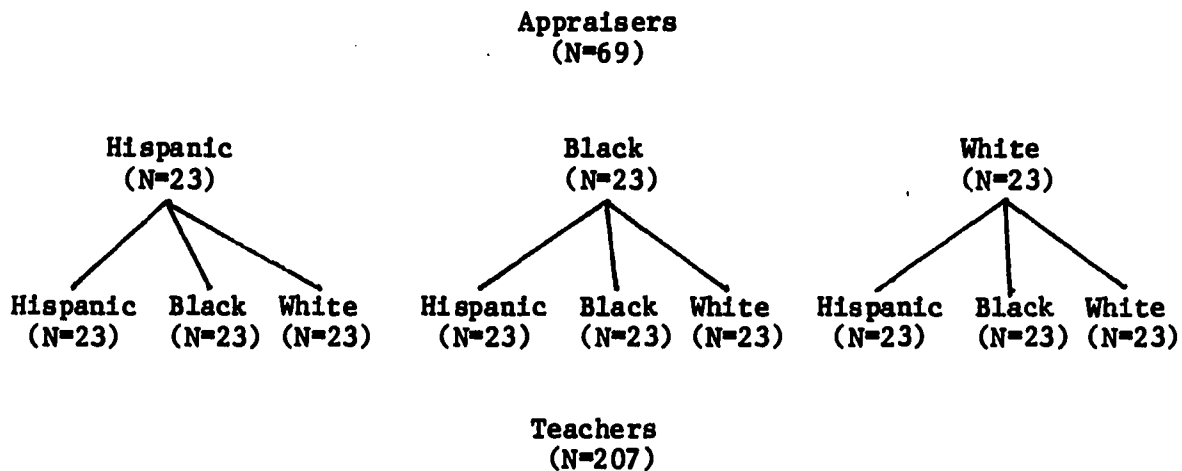


Figure 11. Research design and number of subjects used to determine differences in teacher performance appraisal ratings based upon race

Figure 12 contains the mean scores for appraisers by each gender group sampled. The two-way analysis of variance (ANOVA) test was performed to determine if differences in mean scores were significant.

Table 4 shows an F-ratio of 8.056, indicating that differences in mean scores are significant at the .001 level. This is the result of differences in mean scores between Hispanic (3.70) and white (4.04) evaluators, and between black (3.81) and white (4.04) evaluators. On the basis of this analysis, this hypothesis was rejected at the .001 level.

Hypothesis 2c. There will be no significant difference in mean teacher appraisal scores based upon the level of educational training of the evaluator.

This hypothesis was tested using the mean first semester teacher appraisal scores for all 3,460 elementary teachers in the Dallas Independent School District during the 1985-86 school year. These ratings were submitted by a total of 112 evaluators who were divided into five subgroups based upon educational training. Figure 13 depicts these subgroups and shows the mean appraisal score for each group.

The one-way analysis of variance (ANOVA) test was used to analyze the significance of mean score differences among these five groups. On the basis of this analysis, the hypothesis was rejected at the .001 level, as noted in Table 5. Further analysis revealed that this level of significance was due to the differences in mean scores between those evaluators possessing training at the M.A. level (4.08) and those at the M.A. + 30 (3.83), M.A. + 45 (3.81), and Ph.D. (3.82) levels.

Hypothesis 2d. There will be no significant difference in mean teacher appraisal scores based upon the number of years of experience in education of the evaluator.

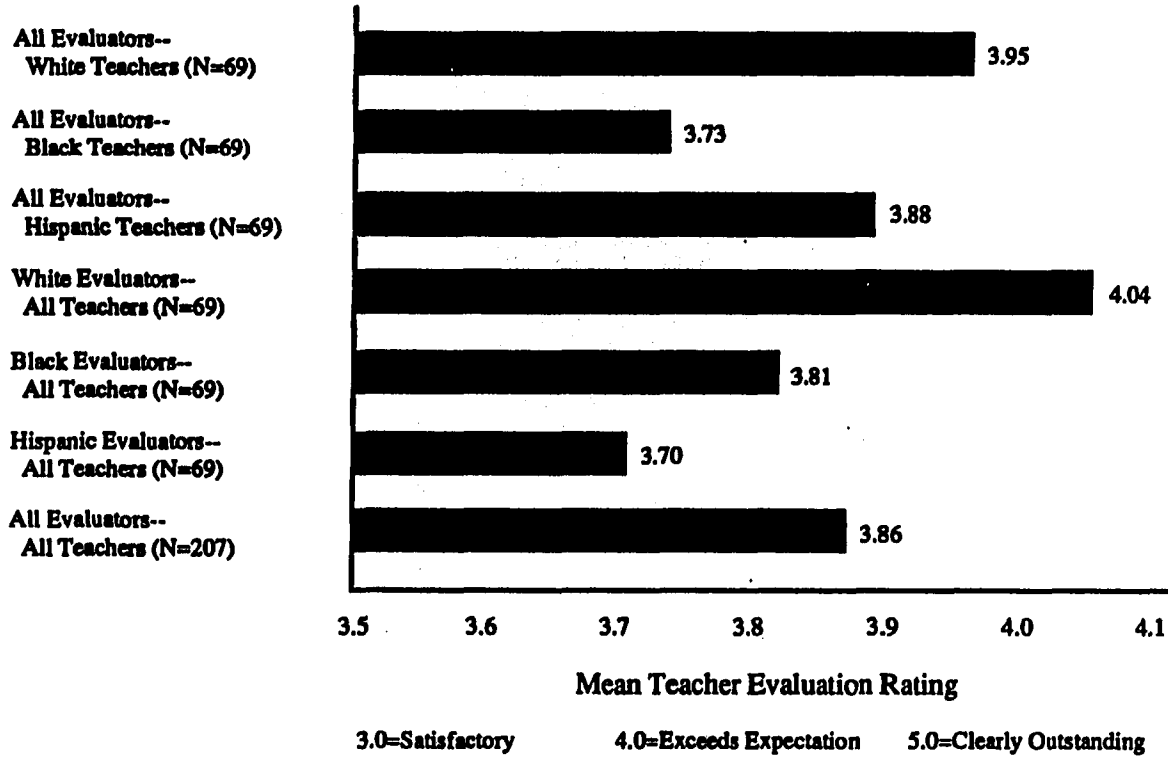
Evaluation Combinations

Figure 12. Mean first semester teacher appraisal scores by race groups for selected elementary evaluators, Dallas Independent School District, 1985-86

Table 4. Analysis of variance of mean elementary teacher appraisal scores by race

Source	df	Sum of squares	Mean square	F-ratio	F-prob.
Appraisers	2	4.161	2.081	8.056***	.001
Teachers	2	1.744	.872	3.377*	.035
Interaction	4	1.543	.386	1.493	.205
Error	198	51.137	.258		

*Significant at $p < .05$ level.

***Significant at $p < .001$ level.

Table 5. Analysis of variance of mean differences in elementary teacher appraisal scores by appraiser's highest level of education

Source	Sum of squares	df	Mean square	F
Between groups	29.154	4	7.289	21.905***
Within groups	1141.590	3455	.333	
Total	1170.744	3459		

***Significant at $p < .001$ level.

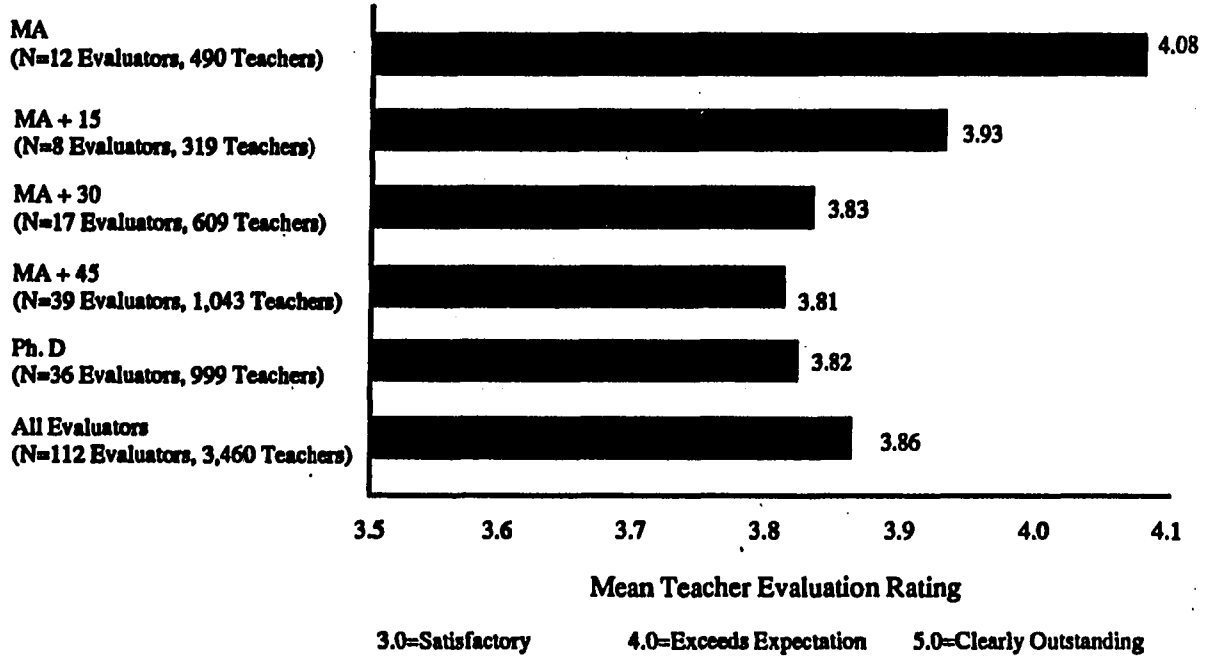
Education Level

Figure 13. Mean first semester teacher appraisal scores by highest level of educational training of evaluator

This hypothesis was tested using the one-way analysis of variance (ANOVA) procedure to determine the significance of differences among mean scores assigned by groups of evaluators based upon their total years of experience in education (teaching and administration). First semester appraisal scores from all 112 evaluators and 3,460 teachers at the elementary level were used in this statistical test. The mean scores by subgroup are shown in Figure 14. Table 6 shows the results of the ANOVA procedure, and on the basis of an F-ratio of 18.376 for the between-group analysis, the hypothesis was rejected at the .001 level of significance. This was attributed to the mean score difference between the 11-15 year group (3.63) and the mean scores for each of the other groups, those being 1-10 years (4.00), 16-20 years (3.86), 21-25 years (3.84), 26-30 years (3.98), and over 30 years (3.88). Significant differences were also noted between the 26-30 year group (3.98) and two other groups, those being 16-20 years (3.86) and 21-25 years (3.84).

Hypothesis 3. There will be no significant difference in mean teacher evaluation ratings due to an interaction effect between the race and gender of the evaluator and the race and gender of the teacher.

The procedure used for this hypothesis was to test gender and race interaction effects separately by using the research design depicted previously in Figures 9 and 11. The interaction effect for gender was tested using a random sample of 68 mean teacher appraisal scores (34 male and 34 female). Table 3 indicates the results of the two-way analysis of variance (ANOVA) test which produced an F-ratio of .254 and an F-probability of .990. Table 4 presents the results of the ANOVA procedure testing the interaction effects of race, which indicate an

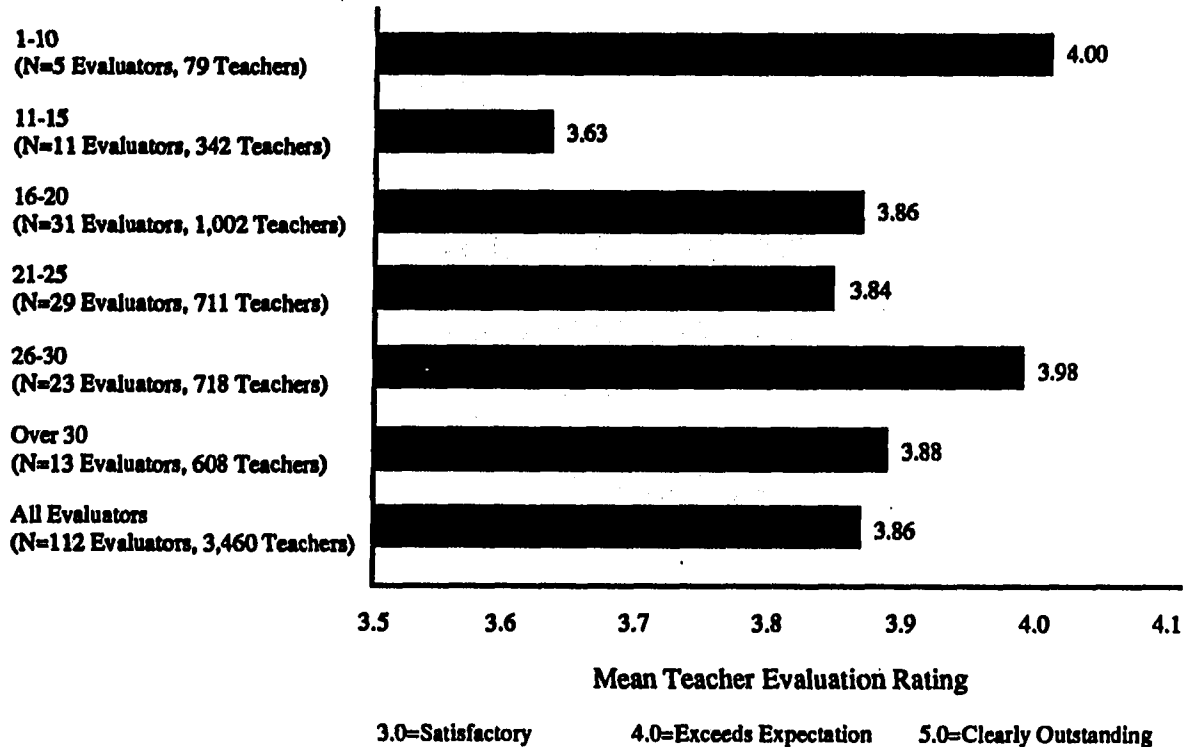
Years in Education

Figure 14. Mean first semester teacher appraisal scores by evaluators' total years of experience in education

Table 6. Analysis of variance of mean differences based upon total years of experience in education of appraiser

Source	Sum of squares	df	Mean square	F
Between groups	30.469	5	6.094	18.376***
Within groups	1145.437	3454	.332	
Total	1175.907	3459		

***Significant at $p < .001$ level.

F-ratio of 1.493 and an F-probability of .205. On the basis of these results, the hypothesis failed to be rejected as it related to interaction effects for both gender and race.

Hypothesis 4. There will be no significant degree of agreement between mean teacher evaluation ratings assigned by the first appraiser and those assigned by the second appraiser.

This hypothesis was tested using all pairs of first and second appraisers at the elementary level who evaluated a common group of 25 or more teachers during the first semester of the 1985-86 school year. Two subhypotheses were developed to facilitate the use of different statistical treatments.

Hypothesis 4a. There will be no significant positive correlation between mean teacher evaluation ratings assigned by first and second appraisers.

Table 7 reveals the results produced by application of the Pearson product-moment test for each pair of evaluator's mean teacher appraisal scores for the teachers they evaluated independently. A total of 19 of the 27 pairs (70 percent) obtained a correlation coefficient at the level

Table 7. Teacher appraisal score correlations between first and second appraisers

Sample	Number of teachers appraised	First appraiser mean score	Second appraiser mean score	r	r ²	t
1	39	4.15	4.16	.95	.90	17.34***
2	27	4.28	4.15	.92	.84	11.42***
3	32	3.93	3.75	.93	.87	13.90***
4	33	4.59	4.64	.92	.84	12.64***
5	26	4.13	3.99	.94	.88	13.44***
6	26	3.83	3.55	.87	.76	8.64***
7	25	4.26	3.90	.54	.29	3.07**
8	25	3.79	3.49	.84	.70	7.33**
9	27	3.83	3.52	.92	.85	11.71***
10	27	3.75	3.74	.54	.29	3.22**
11	27	3.37	3.40	.98	.96	24.20***
12	30	4.13	3.70	.61	.37	4.07***
13	25	3.52	3.42	.79	.62	6.08***
14	27	4.40	4.31	.88	.78	9.33***
15	27	4.55	3.89	.80	.65	6.74***
16	25	3.91	3.44	.69	.47	4.54***
17	29	4.25	3.94	.86	.74	8.83***
18	25	3.60	3.56	.87	.77	8.72***
19	30	4.26	4.25	.87	.76	9.31***
20	27	3.76	3.48	.87	.75	8.71***
21	29	3.58	3.51	.82	.67	7.44***
22	25	3.92	3.88	.82	.67	6.86***
23	32	3.62	3.53	.66	.44	4.84***
24	34	4.07	3.82	.69	.47	5.35***
25	31	4.36	4.22	.73	.54	5.82***
26	25	3.97	3.75	.81	.66	6.72***
27	31	4.17	4.02	.95	.91	17.03***

**Significant at $p < .01$.

***Significant at $p < .001$.

considered significant for testing this hypothesis (.80 or higher). All of the tests were significant at the $p < .01$ level.

Table 8 shows the results produced from conducting the Pearson product-moment test to the mean scores assigned by each pair of appraisers. This produced a combined average correlation coefficient of .863 for the 27 pairs of scores, and a coefficient of determination of .750. On the basis of these analyses, the subhypothesis of there being no significant positive correlation between ratings of first and second appraisers was rejected.

Table 8. Correlation between first semester mean teacher appraisal scores assigned by selected evaluators

Number of pairs sampled	Mean score first appraisers	Mean score second appraisers	r	r ²	t
27	4.000	3.814	.863	.75	8.54***

***Significant at $p < .001$.

Hypothesis 4b. There will be no significant difference between mean teacher evaluation ratings assigned by first and second appraisers.

Table 9 displays the results of the student's t-test for mean teacher appraisal scores for the same 27 pairs of scores used in Table 7. Only six of the 27 pairs of evaluators' scores (22 percent) showed mean scores that were significantly different at the .05 level. Table 10 reveals that there was no significant difference between the combined average mean scores of first and second appraisers as the T-ratio of 2.128 produced an

Table 9. T-test analysis for significance of differences in mean teacher appraisal ratings for selected elementary evaluators

Sample	Number of teachers appraised	First appraiser mean score	Second appraiser mean score	t	F-prob. two-tail
1	39	4.15	4.16	-.053	.99
2	27	4.28	4.15	1.018	.63
3	32	3.93	3.75	.930	.999
4	33	4.59	4.64	-.459	.999
5	26	4.13	3.99	1.136	.522
6	26	3.83	3.55	1.940	.110
7	25	4.26	3.90	4.394***	.001
8	25	3.79	3.49	2.826**	.010
9	27	3.83	3.52	2.202	.060
10	27	3.75	3.74	.022	.999
11	27	3.37	3.40	-.128	.999
12	30	4.13	3.70	4.687	.999
13	25	3.52	3.42	.677	.999
14	27	4.40	4.31	.741	.999
15	27	4.55	3.89	4.848***	.001
16	25	3.91	3.44	3.243**	.010
17	29	4.25	3.94	1.871	.126
18	25	3.60	3.56	.228	.999
19	30	4.26	4.25	.093	.999
20	27	3.76	3.48	2.375*	.040
21	29	3.58	3.51	.666	.999
22	25	3.92	3.88	.303	.999
23	32	3.62	3.53	.755	.999
24	34	4.07	3.82	2.455*	.032
25	31	4.36	4.22	1.147	.510
26	25	3.97	3.75	1.931	.112
27	31	4.17	4.02	1.346	.360

*Significant at $p < .05$.

**Significant at $p < .01$.

***Significant at $p < .001$.

Table 10. Two-tailed T-test analysis for mean teacher appraisal scores assigned by first and second appraisers

	First appraisers	Second appraisers	t	df
N	27	27		
Mean	4.000	3.814	2.128	52
SD	.321	.321		

f-probability of .072. Based on these results, the subhypothesis of there being no significant difference between mean teacher evaluation ratings for first and second appraisers failed to be rejected.

On the basis of the combined analysis of the relationship between the mean scores (correlation tests) and of the differences in mean scores (t-tests), the hypothesis of there being no significant high rate of agreement between mean teacher evaluation ratings for first and second appraisers was rejected.

Hypothesis 5. There will be no significant positive correlation between an evaluator's first and second semester mean teacher appraisal ratings.

This hypothesis was formulated to compare an evaluator's ratings of the same teachers for repeated formative appraisals. The subjects for this hypothesis consisted of all appraisers, including both first and second appraisers, who evaluated a common group of at least 25 teachers during both first and second semesters during the 1985-86 school year.

The Pearson product-moment procedure was selected for its ability to produce a correlation coefficient between appraisal scores, thus

establishing a coefficient of stability for the appraisal instrument. Table 11 indicates that 33 of the 54 appraisers' scores (61 percent) had a correlation coefficient of .80 or higher between their first and second semester evaluations of the same group of teachers. Table 12 shows the results of the Pearson product-moment test for its application to mean first and second semester scores for all teachers evaluated by the 54 evaluators who appraised 25 or more teachers. This test produced a combined average correlation coefficient of .85 between the first and second semester appraisals. On the basis of these two tests, the hypothesis of there being no significant positive correlation between an evaluator's scores for repeated measures was rejected.

Table 11. Teacher appraisal score correlations between an appraiser's first semester and second semester evaluations

Appraiser	Number of teachers appraised	Mean score first semester	Mean score second semester	r	r ²	t
1	33	3.99	4.16	.84	.70	8.47***
2	32	3.39	4.02	.55	.31	3.63**
3	32	3.29	3.72	.79	.63	7.11***
4	30	4.33	4.46	.91	.83	11.53***
5	34	3.88	4.21	.90	.82	11.93***
6	32	4.62	4.66	.94	.88	14.50***
7	25	4.20	4.36	.87	.75	8.37***
8	42	3.59	4.19	.88	.78	11.81***
9	43	3.95	4.07	.86	.74	10.76***
10	26	3.73	4.08	.77	.59	5.81***
11	29	4.00	4.09	.94	.88	13.83***
12	32	3.92	3.91	.61	.37	4.18***
13	43	3.34	3.69	.71	.51	6.46***
14	38	3.92	4.15	.91	.83	13.28***
15	31	4.63	4.79	.91	.83	11.98***
16	45	3.92	4.17	.78	.61	8.23***
17	28	4.27	4.46	.86	.75	8.74***
18	39	3.57	3.90	.73	.54	6.57***
19	28	3.93	4.21	.86	.74	8.52***
20	41	3.82	4.20	.69	.47	5.88***
21	29	3.55	3.81	.36	.13	2.00
22	50	3.61	4.24	.85	.72	11.06***
23	37	3.78	3.98	.90	.80	11.89***
24	33	4.11	4.23	.95	.89	16.19***
25	40	4.10	4.27	.74	.55	6.79***
26	33	3.46	3.71	.86	.73	9.19***
27	27	4.04	4.50	.96	.91	16.07***
28	29	3.74	3.97	.93	.87	13.21***
29	29	4.07	4.39	.80	.64	6.87***
30	33	3.66	4.27	.66	.44	4.95***
31	30	3.72	4.03	.86	.74	8.82***
32	35	4.32	4.75	.69	.47	5.40***
33	47	3.26	4.04	.57	.33	4.70***
34	31	4.09	4.12	.91	.82	11.52***
35	39	3.98	4.28	.57	.32	4.17***

**Significant at $p < .01$.

***Significant at $p < .001$.

Table 11. Continued

Appraiser	Number of teachers appraised	Mean score first semester	Mean score second semester	r	r ²	t
36	34	4.14	4.45	.84	.70	8.71***
37	27	3.61	3.81	.83	.69	7.38***
38	39	4.18	4.64	.40	.16	2.68**
39	31	3.72	4.14	.68	.46	4.95***
40	26	3.48	3.85	.55	.30	3.22**
41	33	4.42	4.61	.89	.80	10.98***
42	37	4.12	4.14	.81	.66	8.30***
43	33	3.90	4.20	.80	.64	7.39***
44	35	4.04	4.08	.83	.70	8.68***
45	30	3.62	3.76	.76	.58	6.16***
46	32	3.75	4.27	.74	.54	5.98***
47	26	3.33	3.64	.91	.82	10.44***
48	32	3.92	4.30	.80	.63	7.18***
49	25	3.94	4.23	.85	.72	7.73***
50	44	4.33	4.46	.87	.76	11.42***
51	41	4.31	4.70	.65	.42	5.33***
52	43	3.87	4.17	.75	.56	7.16***
53	36	3.66	3.76	.81	.66	8.14***
54	49	3.73	4.15	.83	.69	10.27***

Table 12. Correlation between mean teacher appraisal scores assigned by selected appraisers for first and second semesters

N	First semester \bar{X}	First semester S.D.	Second semester \bar{X}	Second semester S.D.	r	r ²	t
54	3.89	.33	4.18	.28	.85	.72	11.54***

***Significant at $p < .001$ level.

CHAPTER V. SUMMARY, CONCLUSIONS, AND RECOMMENDATIONS

Summary

The primary purposes of this study were to (1) determine whether or not teacher evaluation scores were affected by rater biases, and (2) determine the reliability of scores for multiple appraisers and for single appraisers for repeated evaluations of the same teachers. A related purpose of the study was to conduct this analysis in a school district using teacher evaluation ratings to determine career ladder advancements. In essence, the study attempted to establish the degree to which the instrument and procedures used by the school district offered all of its teachers fairness and equal access to the rewards of a career ladder advancement system.

Evaluations from the Dallas, Texas, Independent School District were used in this study, with the data base consisting of approximately 34,000 completed copies of the Written Record of Observation. These instruments were collected during the second year (1985-86 school year) of that district's participation with the School Improvement Model Project at Iowa State University. Sampling procedures were used to further define the population of subjects used for each of the research hypotheses, and the analysis of the data in each case resulted in findings relating to the major goals of the study.

Internal consistency of the appraisal instrument

Question 1. Will there be agreement among the ratings for different criteria used on the appraisal instrument?

There was a high rate of agreement among the scores for different criteria on the teacher performance appraisal system. Evaluators rated teachers at a mean level of 3.86 on a 1.0 to 5.0 point scale for all criteria combined. The highest rated criterion was Knowledge of Subject Matter (4.02), and the lowest rated criterion was Evaluation Activities (3.73). The mean score for the subset of five criteria directly observable during classroom instruction was 3.89, and the average for the four criteria more likely to be assessed through examination of artifacts of teaching was 3.84. This resulted in a correlation coefficient of .8407 occurring between scores of these two subsets of criteria, and correlations of .9476 and .9696 respectively between the scores for each of the subsets of criteria and the average score (3.86) for all criteria combined.

Effect of rater characteristics on teacher appraisal scores

Question 2. Will the use of teacher evaluations for promotion in a career ladder system be subject to systematic error due to certain characteristics of the rater?

There was a significant difference in mean teacher appraisal scores associated with the gender of the evaluator. Female evaluators rated a sample of teachers at an average score of 3.70, and the average for the sample of teachers evaluated by males was 3.95. There was also a significant difference in scores based upon the race of the rater. Hispanic evaluators had the lowest average rating (3.70) and whites had the highest (4.04). A significant difference also was noted between ratings by black evaluators (3.81) and white evaluators (4.04).

Significant differences in mean teacher evaluation ratings were also found based upon the level of training and years of experience of the evaluator. Evaluators with an M.A. degree produced an average teacher appraisal score of 4.08 on a five-point scale, which was significantly more lenient than the mean scores for those evaluators at the M.A.+30 (3.83), M.A.+45 (3.81), and Ph.D. (3.82) levels.

For the variable of experience, those evaluators with 11-15 years experience in education (teaching and administration combined) produced a mean teacher appraisal score of 3.63, which was significantly more severe than each of the other experience groups, those being 1-10 years (4.00), 16-20 years (3.86), 21-25 years (3.84), 26-30 years (3.98), and over 30 years (3.88). Also, the 26-30 year group had a significantly more lenient mean score (3.98) than the 16-20 (3.86) and 21-25 (3.84) year groups.

Question 3. Will teacher evaluation ratings be subject to systematic error due to an interaction effect between characteristics of the rater and ratee?

Gender and race were selected to determine if those characteristics of the evaluator produced differences in mean appraisal scores based upon an interaction effect with the gender and race of the teacher being evaluated. Tests were conducted separately for these variables and no significant interaction effect was found for either gender or race characteristics.

Inter-rater agreement

Question 4. Will there be agreement between appraisal ratings assigned by two different evaluators for a common group of teachers?

A high rate of agreement was found between the scores of first and second appraisers who were sampled in this study. Seventy percent of the 27 pairs of evaluators attained a correlation coefficient of .80 or higher. Further analysis showed a high degree of common variance (75 percent) for the scores of the two appraiser groups. Twenty-two percent of the pairs of evaluators had mean appraisal scores that were significantly different from each other, and no significant difference was found between the overall mean score for first appraisers (4.000) and second appraisers (3.814).

Intra-rater agreement

Question 5. Can evaluators make consistent ratings on repeated measures of a teacher's performance?

Appraisers were found to be able to make consistent ratings over time as measured through correlation coefficients between their first and second semester ratings of a common group of teachers. Of the 54 evaluators sampled, 61 percent had a correlation coefficient of .80 or greater between their ratings. The coefficients ranged between .36 and .96, and all but one (.36) were significant. The average correlation coefficient for the group was .85, with an average of 72 percent of the variances being in common for the group of scores.

Conclusions

The analyses of the data point to several conclusions relating to the internal consistency of the appraisal instrument used, the effect of rater/ratee characteristics on appraisal scores, the ability of multiple

appraisers to agree on ratings for the same teacher, and the ability of evaluators to make consistent ratings over time.

1. The instrument used was consistent in producing ratings among two subsets of criteria. From these findings it appears justified to conclude that ratings of teaching performance criteria have a high degree of correlation with each other whether the evaluator gathered data from classroom observation or from other sources.

2. Male evaluators rated teachers significantly higher than female evaluators. However, the lack of interaction effect between rater and ratee gender suggests that the gender of the teacher does not affect an individual evaluator's rating of a teacher's performance.

3. Minority evaluators appear to be more severe than white evaluators in their assignment of teacher performance appraisal ratings. This holds true regardless of the race of the teachers evaluated in the target population.

4. Evaluators with higher levels of educational training tend to assign more severe teacher appraisal scores than evaluators with less training.

5. While some significant differences were noted in teacher appraisal scores based upon the experience level of the evaluator, these differences do not occur in a linear fashion. The findings lead one to suspect a curvilinear relationship between experience level and appraisal scores. Those evaluators with the least experience are likely to start their careers giving lenient ratings, followed by stricter ratings in

mid-career, and then returning to more lenient ratings after several years of experience in education.

6. Teacher performance appraisal ratings do not vary significantly between two different appraisers. In essence it makes little difference in terms of the end result if a teacher is appraised by his or her building principal (first appraiser) or by a principal from another building in the school district (second appraiser).

7. If a teacher is evaluated more than once in a school year by the same appraiser, it is highly likely that subsequent appraisals will have a substantial positive correlation to the first appraisal. In essence the first appraisal score for a teacher appears to be a very accurate predictor of future appraisal scores.

Limitations

1. All teacher evaluation data to be analyzed in this study came from a single large urban school district with a significant minority student and staff population, and generalizations cannot be made outside that population.

2. All evaluation data were gathered during a single school year, thus limiting the investigator's ability to analyze the stability of evaluation ratings beyond that time frame.

3. Variables not considered in this study may have an undetected direct or interaction effect on evaluation ratings of teachers.

4. All teacher evaluators were required to be involved in an on-going training program during the year that evaluation data were collected. This requirement prevented the establishment of a control

group, and therefore no attempt was made to analyze the effect of this training program on the evaluation ratings.

5. Data analyzed for all hypotheses were drawn from teacher evaluations at the elementary level (grades K-6). While this procedure helped hold several variables constant, it limited the investigator's ability to analyze the significance of differences in evaluation scores based upon the level of students taught by the teacher.

6. Due to the lack of consistency between the formative and summative evaluation instruments, the data analyzed included only formative appraisals made by teacher evaluators. The summative instrument differed from the formative in its addition of a criterion called "Employee Responsibilities." Also, the summative instrument was completed by only the first appraiser (principal), whereas the formative appraisals were completed by multiple appraisers. Therefore, the use of only formative appraisals was deemed most appropriate for answering questions posed by this study. While this procedure allowed for a consistent focus on performance ratings over time (first semester and second semester ratings within the same school year), it did, however, prevent the analysis of the relationship between an evaluator's formative rating and the end-of-the-year (summative) rating.

Discussion

This study has attempted to add to the body of knowledge relating to bias and reliability in appraisal ratings. Despite the difficulties of taking a very large data base and defining it through sampling techniques, much has been learned in this study about the effect of an evaluator's

gender, race, level of training, and experience in education as they relate to teacher appraisal ratings. Additionally, a substantial amount of evidence was presented showing the ability of the appraisal instrument to produce reliable scores across raters and across time. However, a number of conditions were present in this study which make it appropriate to caution readers regarding going beyond the research literature presented in making generalizations.

Consider, for example, the appraisal instrument used by evaluators in this study. Appraisers rated nine different criteria using a five-part graphic response mode. The instrument, unlike others used by many school districts, was designed as a "low inference" tool for appraisers. Each response indicator was coupled with a detailed descriptor suggesting specific teacher actions for that level of performance (see Appendix B). These descriptors were developed through the "critical incident" technique described by Wexley and Yukl (1984) in their discussion of Behaviorally Anchored Rating Scales (BARS). The use of the BARS system has the advantage of focusing on specific behaviors, thus reducing the extent to which ratings are affected by personal bias (Borman, 1977). Evaluation instruments requiring a higher degree of inference on the part of the rater may well produce ratings with higher levels of error due to personal bias than those obtained in this study.

The role of rater training is also an important factor to acknowledge in presenting conclusions of this study. The use of low inference appraisal instruments, coupled with intensive training of evaluators, can produce reliable results, as shown in this study. Also, one could

speculate, on the basis of several studies of the effects of rater training (Savage, 1983; Pulakos, 1984; Wexley and Yukl, 1984; Beebe, 1987), that over time the differences in appraisal ratings between evaluator groups would become even smaller.

The sampling procedures selected also merit discussion. The use of subjects at the elementary level, coupled with the specific descriptors of the appraisal instrument, may have influenced the level of the ratings. For example, the lowest rated criterion (Criterion 4, Evaluation Activities) contained many descriptors related to tests and other formal ways of evaluating progress that may occur more frequently at the secondary level than at the elementary level. Therefore, it would be unwise to automatically assume that similar results would have occurred at other levels for this criterion or others on the appraisal instrument. Evaluators in this study may, in the absence of actually observing teacher behaviors, have assigned ratings based more on their own personal biases, a tendency also noted by Nieva and Gutek (1980).

The selection of a large urban school district as the subject for this study presented a unique opportunity to study the effect of rater characteristics on appraisal scores. Only in a large urban school district could one find in sufficient numbers the evaluators and teachers for each group of rater characteristics studied. Studying a single school district of this size resulted in a number of advantages for the researcher, not only in the gathering and processing of the data, but also in the making of generalizations of the findings to other similar school organizations. However, the nature of the data base, coupled with the

sampling procedures, merits further discussion in order to clarify the conclusions relating to rater characteristics.

First, in examining the effects of gender it should be noted that, although a sufficiently large number of female evaluators (34) were present to provide integrity for the statistical tests used, females still represented only 30 percent of the total number of evaluators. Among teachers, however, the vast majority (84 percent) were females. While statistically significant differences in mean teacher appraisal scores were found between male evaluators (3.95) and female evaluators (3.70), the effect of the gender imbalance should be noted. One might conclude that if this similar imbalance exists in other schools (a highly likely probability), then the generalizability of the findings of this study would be enhanced. However, it is possible that in schools in which the gender of the evaluators is more balanced, one would be more likely to find ratings closer together for males and females (Peck, 1978).

A more detailed analysis of the data base revealed the extent to which elementary teaching was a female endeavor in the subject school district. In many of the 112 elementary schools in Dallas, only two or three male teachers were employed. Although not a topic for this study, it is of interest to note that these male teachers were rated lower on the average (3.74) by all evaluators than were female teachers (3.92). Further analysis revealed that the highest rated random sample of teachers was the group of female teachers evaluated by male evaluators (4.02), whereas the lowest mean appraisal score (3.59) was recorded for male teachers evaluated by female evaluators. With the commonly held

perception of elementary school teaching being sex role congruent only for women, it is possible that all male elementary teachers' ratings in this study suffered from the tendency of evaluators to assign lower ratings to employees in sex-role incongruent positions, as found by Nieva and Gutek (1980) and Carroll (1982).

The mean scores cited above helped contribute to the finding of no significant interaction effect between the gender of the appraiser and the teacher, although parts of the "like-me" bias are present that normally would reveal itself in evaluators giving higher ratings to teachers of the same sex. One would expect the pattern that was discovered in the appraisals of the female evaluators in this study, that being a higher mean rating for female teachers (3.81) than for males (3.59). However, for male evaluators a similar trend was found, with higher mean ratings given to female teachers (4.02) than to male teachers (3.89). These findings are similar to those of Harrington (1984) but dissimilar to results obtained by Landy and Farr (1983). Both of these research efforts, however, were conducted in laboratory settings using simulated samples of employee work performance.

In this study actual performance ratings were used, and females gave significantly stricter ratings. There could be several reasons for these results. Some researchers contend that more competent evaluators give stricter ratings (Nieva and Gutek, 1980; Landy and Farr, 1983; Wexley and Yukl, 1984), and a case could be made that the females in the study might be more competent. In particular, females come to their first principalship with more years of teaching experience and more experience

in curriculum development, according to Erickson (1985), who studied female principals in Montana over a two-year period of time. These experiences may make female appraisers more knowledgeable about effective teaching practices and give them the skills needed to assign and defend lower ratings.

An evaluator's style of management could also have played a part in the rating process. Landy and Farr (1980) found task-oriented evaluators to give stricter ratings than employee-relations oriented evaluators. The practical implication of these findings in terms of career ladder advancement for teachers is of importance to researchers and to school districts with a similar mixture of male and female teachers and evaluators. One is led to conclude, based upon the findings of this study, that teachers (male or female) evaluated by males have a greater possibility of receiving higher ratings (and thus have a greater chance for career ladder advancement) than teachers evaluated by females. On the other hand, the findings of no significant interaction effect between the gender of the appraiser and the teacher suggests that, regardless of the gender of the individual evaluator, male and female teachers have an equal opportunity for career ladder advancement when compared with all teachers appraised by that particular evaluator.

As with the findings related to the effects of rater gender, the analysis of effects due to race revealed significant differences based upon the race of the appraiser. However, contrary to findings by Mobley (1982), no interaction effect was found between the race of the evaluator and the race of the teacher. In essence one is led to conclude that

minority evaluators had stricter standards for all teachers regardless of race. One is tempted to speculate that these differences resulted from biases held by the rater, but the lack of interaction effect for rater/ratee race effect seems to suggest otherwise. One would have expected, based upon previous research studies (Decotlis and Petit, 1978; Carroll, 1982; Mobley, 1982; Landy and Farr, 1983), that raters of the same race would have assigned higher scores to teachers of the same race. The fact that this did not happen suggests other reasons for differences in appraisal scores based upon the race of the evaluator.

The possibility exists, of course, that these differences are reflective of true differences in teacher performance. If this were true, however, one would then conclude that minority principals were assigned to buildings in which the level of teaching competence was generally lower than for teaching staffs supervised by white principals. A more likely reason may be found in interaction effects not controlled or tested for in this study. The variable of gender was controlled by using only teachers of the same gender for each evaluator, but other variables were not accounted for, among them the level of training and the experience of the evaluator, the leadership style of the evaluator, and the quality and frequency of interactions between the rater and ratee.

Unlike the mixed results of tests for score differences based upon gender and race, a clear trend was found for different levels of educational training. Scores generally became slightly lower as the education level of the evaluator increased. This suggests that the training of these evaluators, either in the amount or the quality of the

training, contributed to lower teacher performance scores being assigned. The differences, however, were not significant among all levels, and once the M.A.+30 level was reached, differences in mean scores decreased dramatically. The mean score at the M.A.+30 level was 3.83, while the M.A.+45 was 3.81, and the score at the Ph.D. level was 3.82.

It is of interest to note the high percentage of principals with Ph.D. degrees, when compared to percentages found in other similar school systems in the country. This was the result of many principals having met the requirements of degree-granting institutions prior to the beginning of the School Improvement Model training project in the Dallas Independent School District. In essence, then, it appears that additional training can be an effective means of combating leniency error by evaluators, as well as being a means of helping remove significant differences among the ratings of the evaluators. This is especially true, according to Pulakos (1984), if the training includes orientation to specific evaluation procedures and techniques.

Unlike the mean teacher appraisal scores by training level, the results of ratings by experience levels yielded no similar pattern or trend. While several significant differences were noted among various experience levels, no clear linear relationship was noted, a finding similar to a study by Harrington (1984). For instance, evaluators in the 11-15 year category had the lowest mean score (3.63), and it was significantly different from the mean score (3.88) for evaluators with over 30 years experience. One is tempted to speculate, on the basis of these results, that appraisal scores will become more lenient as the

evaluator's experience increases, a finding contradictory to the previous discussion of evaluations becoming less lenient as the evaluator's training level (and quite possibly his or her experience level as well) increases. However, one also must note the significant difference among the 26-30 year group (3.98) and the over 30 group (3.88). This finding, along with the 4.00 mean score for the least experienced group of 1-10 years, leads to the conclusion that evaluators may go through different phases in their careers, with the mid-career years being the ones most likely to produce the most severe ratings.

Analysis of interrater agreement in this study showed that two pairs of raters were frequently in close agreement in their evaluations of a common group of teachers. For the 27 pairs sampled, first appraisers rated teachers at an average of 4.00 on a five-point scale, and second appraisers rated teachers at an average of 3.814. These scores correlated at a substantial level ($r=.863$), with 75 percent of their variances being in common. Only 22 percent of the pairs of scores were different at the .05 level, and no significant difference was found between the overall mean evaluation scores of first and second appraisers. On the basis of these findings, the interrater reliability of the scores produced by the evaluation system was affirmed. However, the reason for this high level of agreement may extend beyond the favorable merits of the evaluator training program or the internal consistency of the appraisal instrument. It is possible that, contrary to the assumed adherence to district procedures, the two evaluators communicated with each other and agreed upon evaluation ratings prior to making their individual appraisals.

It is of interest to note that the "second appraiser" in this sample was not a peer of the teacher in a true sense of the word. While the second appraiser was not a supervisor of the teacher, the second appraiser was, in all instances, the principal of another elementary building in the same school district. Previous studies showing that appraisal scores tend to become more severe the farther away the appraiser is from the appraisee in the organizational hierarchy (Decotiis and Petit, 1978; Doyle, 1983) would suggest higher scores being assigned by second appraisers if they were peers of the raters. However, in the case of the sample used in this study, the building principal was actually closer to the teacher than the second appraiser. Put differently, in the system used in the Dallas Independent School District at the elementary level, it was a teacher's own building principal who continued to work and interact with the teacher before, during, and after the appraisal ratings were made. The second appraiser, on the other hand, could make his or her observations and ratings of the teacher, report them to the teacher's building principal (first appraiser) and disengage from the evaluation process without any further face-to-face contact with the teacher. It is of interest, then, to note that the mean score for the sample of second appraisers (3.814) was lower than that for first appraisers (4.000). In essence, then, the results of this study remain consistent with studies showing the tendency for leniency to be more evident in the ratings of evaluators who have an on-going relationship with those who are appraised (Landy and Farr, 1983).

Concerning the ability of an appraiser to produce consistent ratings over time, this study found that a high degree of agreement existed

between the first and second set of mean teacher appraisal scores, a finding similar to that of McNeil and Popham (1973). Sixty-one percent of the pairs of scores had a correlation coefficient of .80 or higher, and a correlation coefficient of .85 was found for the relationship between the mean first and second appraisal scores for the 54 evaluators. It is of interest to note that the mean appraisal score for the 54 evaluators increased between the first semester (3.89) and the second semester (4.18). This points out another trend that may have practical importance for career ladder advancement of teachers, that being the tendency of subsequent appraisals to be more lenient than the first appraisal. This elevation of scores is a logical happening if the appraisal system has as one of its purposes the improvement of instructional skills. This element was present in the DISD system, and therefore it seems possible that evaluators may have been predisposed to giving higher second appraisal ratings regardless of true differences in performance over the first evaluation.

The question posed by this study was not, however, the significance of differences between the first and second appraisals, but rather the degree of agreement between the scores, a procedure frequently used to determine reliability (Rowley, 1976). This question relates to the practical issue of how often a teacher must be appraised in order for a stable measure of the teacher's performance to be established. The results of this study suggest, for the most part, that the first appraisal of a teacher is a highly reliable predictor of future appraisals. This finding supports the contention that the appraisal instrument used in this

study produced reliable results, but also suggests that, if time does not exist to make multiple formative appraisals, the assigning of a single formative appraisal score for a teacher is likely to be a good predictor of subsequent ratings of that teacher.

Impact on career ladder implementation

Results obtained in this study have practical implications for those who design and implement career ladder systems for teachers. One important factor yet to be discussed is the question of how many teachers would actually be promoted based upon the results of the evaluations reported in this study. While state and local planners usually refrain from discussing quotas for promotion, the underlying assumption of most career ladder systems is that not all of the teachers in a school system will be promoted. For example, those involved in the design of Tennessee's Master Teacher Plan felt that only about 15 percent of the state's teachers were Master Teachers (Pate-Bain, 1983).

The school district referred to in this study used both the results of performance appraisals as well as other criteria (see Appendix E) to determine qualification for advancement. An overall performance appraisal of "exceeds expectations" was necessary to advance beyond the entry level step, and an overall appraisal of "clearly outstanding" was required to advance to the top step of the career ladder. To achieve these levels teachers needed to receive a certain minimum rating for each performance criteria (see Appendix B, Summative Evaluation Form). Translated into the statistical format used in this study, a teacher needed to achieve a minimum mean appraisal score of at least 3.77, on a five-point scale, for

advancement to Levels Two and Three, and 4.77 to be considered for Level Four.

Table 12 shows that a sample of 54 appraisers rated teachers at an average of 3.89 for first semester and 4.18 for second semester. By using the standard deviation of .28 for the second semester formative appraisal, z-score analysis indicates that 92.65 percent of a normally distributed population of scores would be 3.77 or higher and 1.79 percent of the scores would be 4.77 or higher. The projection of these percentages to the 3,460 teachers in this study would mean that 3,206 teachers (92.65 percent of the total) would qualify for career ladder advancement to Levels Two and Three, but only 62 teachers (1.79 percent) would qualify for Level Four.

Readers of these statistics should be mindful of their tentative nature, since these projections are based upon formative appraisals, whereas the scores from the summative evaluation are the ones actually used to determine career ladder advancement (see Appendix E). However, these figures are useful in examining the degree to which a sample of teachers scored in relation to promotion standards. Also, based upon the findings in this study of a high correlation between the first and second formative appraisal ratings, one could anticipate that the projections of career ladder advancement based upon the second formative appraisal will have a high correlation with the actual results of the summative appraisal.

Career ladder developers at the state and local levels may view these findings with mixed reactions. First, having almost all of the teachers

being rated as "exceeding expectations" may lead them to question the leniency of the evaluators as well as the leniency of the standards for promotion to Levels Two and Three. The level of leniency found in this study may also lead to difficulties in funding the increased salaries for this number of teachers and may be viewed by skeptics as merely a salary escalator under a different name. However, career ladder advocates interested in demonstrating that an evaluation system can be responsive to the need to promote only a smaller number of outstanding teachers to the top level of a career ladder system will be encouraged by the findings of this study showing only 1.79 percent of the total qualifying for promotion to Level Four.

Teachers also may have mixed reactions to these findings. As stated previously in this study, teacher organizations have traditionally opposed pay-for-performance systems, with one reason being the potential for such systems to promote disunity within the organization through the use of unequal compensation patterns (Lieberman, 1985). However, the high percentage of teachers qualifying for promotion to Levels Two and Three found in this study may lessen their concern. They still may express concern, however, that the small number of teachers not promoted are the victims of circumstances unrelated to their actual teaching performance, such as rater bias. Teachers may also be concerned about the small percentage of the total teaching staff who received ratings that would qualify them for advancement to the top level of the career ladder. Even those teachers who speak favorably about career ladders may make a case

that the potential for being promoted to this level is so small that teachers will not be motivated to strive for it.

Both career ladder planners and teachers in those schools using career ladders should find comfort in the high degree of both inter- and intrarater reliability found in this study. State planners and local officials should reasonably expect instruments and procedures similar to those used in this study to produce consistent results over time. Without this degree of reliability the instruments and procedures would surely be less defensible legally and could justifiably come under attack by teachers and their professional organizations.

The results found in this study related to rater bias, however, remain troublesome. States with heterogeneous populations of evaluators and teachers still face difficulties in demonstrating that evaluations reflect only true levels of employee performance and not systematic error due to biases held by the rater. The data in this study show that minority teachers received lower ratings than white teachers. In practical terms the "average" white teacher achieved the minimum level for advancement to the second step of the career ladder (3.77 mean rating) whereas the "average" black teacher did not (see Figure 12). Also, it was easier to achieve the career ladder cutoff if a teacher were evaluated by a white evaluator than if he or she were evaluated by a black or Hispanic rater. Over the course of time these biases, if left unchecked, could result in proportionately fewer minority teachers being advanced to the higher levels of the career ladder. Likewise, the apparent difference in

standards held by white and minority evaluators could be an easy target for teachers who are not in agreement with the results of evaluations.

Similar concern should surround the results relating to gender bias. Previous discussion centered around the possible reasons for female elementary teachers being rated higher than males, and for female evaluators rating teachers lower than their male counterparts. The practical implication remains, however, that ratings with this level of bias are open to criticism for putting certain groups at a disadvantage in competing for career ladder advancement.

Finally, differences were also noted in evaluation scores based upon the education level and years of experience of the evaluator. Unlike the variables of gender and race, however, these variables are alterable during an evaluator's career, and therefore can be viewed as less of a threat to the integrity of an evaluation system. The results of this study support the use of trained and experienced evaluators as a way of reducing the potential for bias in evaluation ratings.

Recommendations for Practitioners

In addition to revealing findings of interest to researchers, the results of this study suggest that certain practices be adhered to by those involved with implementing teacher performance appraisal systems, especially those using a career ladder advancement system for compensating teachers.

1. Continued use should be made of appraisal instruments, such as that used in the 1985-86 school year by the Dallas Independent School District, which contain criteria validated by research as those that

possess the ability to discriminate among varying levels of teacher performance.

2. Training of evaluators is essential in order to enhance the reliability of appraisals, and should be an on-going process in districts using evaluation ratings as the basis for promotion. The ability of raters to develop a common understanding of effective teaching practices is important so that teachers can be similarly assessed by different raters.

3. The use of female and minority evaluators should be encouraged in school districts implementing teacher performance appraisal systems. Although some evidence presented in this study points to the continuance of systematic error due to the effects of race and gender, an equal or greater amount of evidence is presented to demonstrate that training and experience have a positive effect in producing reliable appraisal ratings. Therefore, over time the differences noted due to gender and race will likely become less, while the positive effects of having evaluators be balanced by gender and race will continue.

4. The use of multiple appraisers is recommended in the implementation of teacher performance appraisal systems. A high degree of interrater agreement can be achieved through the use of an effective training program and through the use of an appraisal instrument containing research-based criteria.

5. Teacher performance appraisals need to be based upon factual information gathered over time by trained evaluators. Evaluators should give feedback to teachers periodically so that their strengths are

positively reinforced and their areas of growth are made known to them. Therefore, it is recommended that separate formative and summative phases be incorporated into appraisal systems, as existed in the Dallas Independent School District's system.

Recommendations for Further Research

1. The study should be replicated in other pay-for-performance districts as a way of supporting the findings of this study. It may also be of interest to researchers to replicate the study in districts not using a pay-for-performance system. This would allow for analysis of differences in ratings based upon the purpose of evaluating teachers.

2. This study focused on evaluations made only at the elementary level (grades K-6). Future research efforts should be broadened to include teachers and evaluators at all grade levels in a school system. This procedure would allow for analysis of differences in ratings based upon the level of assignment in a school system.

3. This study tested the effects of rater training on appraisal scores through the use of each rater's highest level of formal education attained. Subsequent researchers should focus on the appraiser's actual knowledge of the elements of effective teaching and knowledge of the district's policies and procedures relating to teacher evaluation. The testing of evaluators, and the matching of those scores with actual teacher evaluation scores, could add support for the results of this study.

4. Additional studies should be conducted which focus on the psychometric quality of evaluation data submitted by raters in positions

other than those used in this study. In particular, the study of self, peer, and student appraisals would assist in determining the reliability of ratings produced by the assessment instrument.

5. Future research efforts should address the quality of the interactions between the rater and ratee. In this study raters indicated something about the level of teacher performance in their evaluations, but they also may have projected how well they liked or disliked particular teachers. Further research efforts should be undertaken which are able to account for this variable.

6. While this study was able to determine the stability level of appraisers' ratings over the course of one school year, studies in the future should measure stability of ratings over an even longer time period. Such longitudinal studies should also determine the relationship between formative and summative appraisals as a means of determining the predictive validity of formative appraisals.

BIBLIOGRAPHY

- Acheson, K., and Gall, M. Techniques in the Clinical Supervision of Teachers. New York, NY: Longman, 1980.
- Alexander, L. "Ten Ways to Honor Teachers." American School Board Journal 172, 1 (1985): 33-34.
- Allen, T. "Identifying Behaviors of the Master Teacher." Ph.D. Dissertation, Iowa State University, Ames, IA, 1986.
- American Association of School Administrators. Effective Teaching: Observations from Research. Arlington, VA: Author, 1986.
- American Psychological Association. Standards for Educational and Psychological Tests. Washington, D.C.: Author, 1985.
- Anania, J. "The Influence of Instructional Conditions on Student Learning and Achievement." Evaluation in Education: An International Review Series 7, 1 (1983): 2-106.
- Anderson, L. "Staff Development and Instructional Improvement: Response to Robbins and Wolfe." Educational Leadership 44, 5 (1987): 64-65.
- Andrews, H. Evaluating for Excellence. Stillwater, OK: New Forums Press, 1985.
- Astuto, T., and Clark, D. Merit Pay for Teachers: An Analysis of State Policy Options. College of Education, Kansas State University, Manhattan, KS, 1985.
- Ban, J., and Soudah, J. "A New Model for Professionalizing Teacher Evaluation." Peabody Journal of Education 56, 1 (1978): 24-32.
- Bartko, J. "On Various Intraclass Correlation Reliability Coefficients." Psychological Bulletin 83, 5 (1976): 762-765.
- Beal, J., Foster, C., and Olstad, R. University of Washington Teacher Assessment System. Teacher Education Research Center, University of Washington, Seattle, WA, 1985.
- Beebe, R. "Developing Sound Performance Appraisal Procedures." National Association of Secondary School Principals Bulletin 71, 497 (1987): 96-101.
- Bell, T. "The Peer Review Model for Managing a Career Ladder/Master Teacher/Performance Pay Program for Elementary and Secondary Schools." Washington, D.C.: Department of Education, 1983.

- Bellon, J. "Evaluator Competencies Needed for Evaluating Teachers and Teaching." Thresholds in Education 10 (1984): 22-24.
- Benzley, J., Kauchak, D., and Peterson, K. "Peer Evaluation: An Interview Study of Teachers Evaluating Teachers." Paper presented at American Educational Research Association. Chicago, IL: American Educational Research Association, 1985.
- Berliner, D. "The Half-Full Glass: A Review of Research on Teaching." In Using What We Know About Teaching, pp. 51-57. Edited by P. Hosford. Alexandria, VA: Association for Supervision and Curriculum Development, 1984.
- Bernardin, J., and Beatty, R. Performance Appraisal: Assessing Human Behavior at Work. Boston, MA: Kent Publishing Company, 1984.
- Blumberg, A. Supervisors and Teachers: A Private Cold War. Berkeley, CA: McCutcheon Publishing Corp., 1974.
- Bolton, D. Evaluating Administrative Personnel in School Systems. New York, NY: Teachers College Press, 1980.
- Borg, W., and Gall, M. Educational Research: An Introduction. New York, NY: Longman, Inc., 1983.
- Borman, W. "Consistency of Rating Accuracy and Rating Errors in the Judgment of Human Performance." Organizational Behavior and Human Performance 20 (1977): 238-252.
- Boyer, E. High School. New York, NY: Harper and Row, 1983.
- Brandt, R. "On the Expert Teacher: A Conversation with David Berliner." Educational Leadership 44, 2 (1986): 4-9.
- Brandt, R. "Proceed With Caution." Educational Leadership 44, 7 (1987a): 3.
- Brandt, R. "On Teacher Evaluation: A Conversation with Tom McGreal." Educational Leadership 44, 7 (1987b): 20-24.
- Brookover, W. "Characteristics of Effective Schools." Paper presented at the Iowa State University Conference on Making Good Schools Better, Ames, IA, June 29, 1987.
- Brophy, J. "Teacher Behavior and Student Achievement." In Handbook of Research on Teaching, 3rd ed. Edited by M. Whittrock. New York, NY: Macmillan, 1986, 328-375.

- Brophy, J. "Using Observation to Improve Your Teaching." (Occasional Paper No. 21). East Lansing, MI: Institute for Research on Teaching, 1979.
- Cardy, R., and Kehoe, J. "Rater Selective Attention Ability and Appraisal Effectiveness: The Effect of Cognitive Style on the Accuracy of Differentiation among Ratees." Journal of Applied Psychology 69, 4 (1984): 589-594.
- Carnegie Forum on Education and the Economy. A Nation Prepared: Teachers for the 21st Century. New York, NY: Author, 1986.
- Carroll, S. Performance Appraisal and Review Systems: The Identification, Measurement, and Development of Performance in Organizations. Glenview, IL: Scott, Foresman & Company, 1982.
- Chirnside, C. "Ten Commandments for Successful Teacher Evaluation." National Association of Secondary School Principals Bulletin 68, 475 (1984): 42-43.
- Conley, D. "Critical Attributes of Effective Evaluation Systems." Educational Leadership 44, 7 (1987): 60-64.
- Cruickshank, D., and Applegate, J. "Reflective Teaching as a Strategy for Teacher Growth." Educational Leadership 38, 7 (1981): 553-554.
- Cuccia, N. "Systematic Observation Formats: Key to Improving Communication in Evaluation." National Association of Secondary School Principals Bulletin 68, 469 (1984): 31-38.
- Cummings, C. Peering in on Peers. Snohomish, WA: Snohomish Publishing Co., 1985.
- Cummings, C. Teaching Makes a Difference. Snohomish, WA: Snohomish Publishing Co., 1984.
- Dallas Independent School District. Teacher Appraisal System Handbook, 1985-86. Dallas Independent School District, Dallas, TX, 1985.
- Darling-Hammond, L., Wise, A., and Pease, S. "Teacher Evaluation in the Organizational Context: A Review of the Literature." Review of Educational Research 53 (1983): 285-328.
- Decotiis, T., and Petit, A. "The Performance Appraisal Process: A Model and Some Testable Propositions." Academy of Management Review 3 (July 1978): 635-646.
- Devries, D., Morrison, A., Shullman, S., and Gerlach, M. Performance Appraisal on the Line. New York, NY: John Wiley and Sons, 1981.

- Dornbusch, S. The Collegial Evaluation Program: A Manual for the Professional Development of Teachers. Stanford Center for Research and Development in Teaching, Stanford University, Palo Alto, CA, 1976.
- Doyle, K. Evaluating Teaching. New York, NY: D. C. Heath and Company, 1983.
- Duckett, W. The Competent Evaluator of Teaching. Bloomington, IN: Phi Delta Kappa, 1985.
- Duke, D., and Stiggins, R. "Evaluating the Performance of Principals: A Descriptive Study." Educational Administration Quarterly 21, 4 (1985): 71-98.
- Duke, D., and Stiggins, R. Teacher Evaluation: Five Keys to Growth. Washington, D.C.: National Education Association, 1986.
- Dunnette, M., and Fleishman, E. Human Performance and Productivity: Human Capability Assessment. Hillsdale, NJ: Lawrence Erlbaum Associates, 1982.
- Eichel, E., and Bender, H. Performance Appraisal: A Study of Current Techniques. New York, NY: American Management Association, 1984.
- Ellis, E. "Peer Observation: A Means for Supervisory Acceptance." Educational Leadership 36, 6 (1979): 423-426.
- Elsay, F. Introductory Statistics: A Microcomputer Approach. Monterey, CA: Brooks Cole Publishing Co., 1985.
- Erickson, H. "Conflict and the Female Principal." Phi Delta Kappan 67, 4 (1985): 288-291.
- Etaugh, C., and Foresman, E. "Evaluations of Competence as a Function of Sex and Marital Status." Sex Roles 9, 7 (1983): 759-765.
- Faast, D. "An Analysis of the Effectiveness of Training Teacher Evaluators in the Specific Steps in the Process of Clinical Supervision and Teacher Performance Evaluation." Unpublished Ph.D. Dissertation, Iowa State University, Ames, IA, 1982.
- Fletcher, C., and Williams, R. Performance Appraisal and Career Development. Brookfield, VT: Brookfield Publishing Company, 1985.
- Foley, W. "On Evaluation and the Evaluation of Teachers." Iowa City, IA: University of Iowa, Institute for School Executives, 1981.
- Fournies, F. Coaching for Improved Work Performance. New York, NY: Van Nostrand Reinhold Company, 1978.

- Frick, T., and Semmel, M. "Observer Agreement and Reliabilities of Classroom Observational Measures." Review of Educational Research 48, 1 (1978): 157-184.
- Furtwengler, C. "Lessons from Tennessee's Career Ladder Program." Educational Leadership 44, 7 (1987): 66-69.
- Glass, G. "Teacher Effectiveness." In Evaluating Educational Performance. Edited by H. Wahlberg. Berkeley, CA: McCutcheon, 1974.
- Goldsberry, L. "Reality--Really? A Response to McFaul and Cooper." Educational Leadership 41, 7 (1984): 10-11.
- Good, T., and Brophy, J. Looking in Classrooms. 3rd ed. New York, NY: Harper and Row, 1984.
- Goodlad, J. A Place Called School. New York, NY: McGraw-Hill, Inc., 1984.
- Greenfield, William D. "Research on School Principals." In The Effective Principal: A Research Summary, pp. 14-19. Reston, VA: National Association of Secondary School Principals, 1982.
- Grossnickle, D., and Cutter, T. "It Takes One to Know One--Advocating Colleagues as Evaluators." National Association of Secondary School Principals Bulletin 68, 469 (1984): 56-60.
- Guba, E., and Lincoln, Y. Effective Evaluation: Improving the Usefulness of Evaluation Results Through Responsive and Naturalistic Approaches. San Francisco: Jossey-Bass, Inc., 1981.
- Harrington, L. "The Association of Administrators' Learning Styles, Background Differences, and Teacher Appraisal Judgments." Ph.D. Dissertation, Iowa State University, Ames, IA, 1984.
- Harris, B. "Resolving Old Dilemmas in Diagnostic Evaluation." Educational Leadership 44, 7 (1987): 46-49.
- Henderson, R. Performance Appraisal. Reston, VA: Reston Publishing Company, Inc., 1984.
- Henderson, R. Performance Appraisal: Theory to Practice. Reston, VA: Reston Publishing Company, Inc., 1981.
- Heneman, R., and Wexley, K. "The Effects of Time Delay in Rating and Amount of Information Observed on Performance Rating Accuracy." Academy of Management Journal 26, 7 (1983): 677-686.

- Hoffman, K. "A Comparison of the Efficacy of Two Types of Teacher Evaluation Instrument Formats." Ph.D. Dissertation, Iowa State University, Ames, IA, 1986.
- Holdzkom, D. "Appraising Teacher Performance in North Carolina." Educational Leadership 44, 7 (1987): 40-44.
- Holley, Freda M. "From Edith Bunker to J. R. Ewing: The Evaluator Plays All the Roles." In The Competent Evaluator of Teaching, pp. 39-64. Edited by W. Duckett. Bloomington, IN: Phi Delta Kappa, 1985.
- Hopfengardner, J., and Walker, R. "Collegial Support: An Alternative to Principal-Led Supervision of Instruction." National Association of Secondary School Principals Bulletin 68, 47 (1984): 35-40.
- Hunter, M. "Knowing, Teaching and Supervising." In Using What We Know About Teaching, pp. 169-192. Edited by P. Hosford. Alexandria, VA: Association for Supervision and Curriculum Development, 1984.
- Ilgen, D. "Gender Issues in Performance Appraisal: A Discussion of O'Leary and Hansen." In Performance Measurement and Theory, pp. 219-230. Edited by F. Landy, S. Zedek, and J. Cleveland, Hillsdale, NJ: Lawrence Earlbaum Associates, 1983.
- Issler, K. "A Conception of Excellence in Teaching." Education 103, 4 (1983): 338-343.
- James, L., Demaree, R., and Wolf, G. "Estimating Within-Group Interrater Reliability With and Without Response Bias." Journal of Applied Psychology 69, 1 (1984): 85-98.
- Joint Committee on Standards for Educational Evaluation. (Pending). Standards for Evaluations of Educational Personnel. Kalamazoo, MI: College of Education, The Evaluation Center.
- Joyce, B., and Showers, B. Power in Staff Development Through Research on Training. Alexandria, VA: Association for Supervision and Curriculum Development, 1983.
- Joyce, B., and Weil, M. Models of Teaching. 3rd ed. Englewood Cliffs, NJ: Prentice-Hall, Inc., 1986.
- Judkins, M. "Identifying Discriminating Items for the Student Evaluation of Teachers." Ph.D. Dissertation, Iowa State University, Ames, IA, 1987.
- Keefe, J. "How Do You Find the Time?" In Rethinking Reform: The Principal's Dilemma, pp. 31-38. Edited by H. Walberg and J. Keefe. Reston, VA: National Association of Secondary School Principals, 1986.

- Keppel, F. "A Field Guide to the Land of Teachers." Phi Delta Kappan 68, 1 (1986): 18-24.
- Krajewski, R. "No Wonder it Didn't Work! A Response to McFaul and Cooper." Educational Leadership 41, 7 (1984): 11.
- Landy, F., and Farr, J. "Performance Rating." Psychological Bulletin 87 (January 1980): 72-107.
- Landy, F., and Farr, J. The Measurement of Work Performance: Methods, Theory, and Applications. New York, NY: Academic Press, 1983.
- Landy, F., Zedeck, S., and Cleveland, J. Performance Measurement and Theory. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc., 1983.
- Latham, G., and Wexley, K. Increasing Productivity Through Performance Appraisal. Reading, MA: Addison-Wesley Publishing Company, 1981.
- Lefton, R., Buzzotta, V., Sherberg, M., and Karraker, D. Effective Motivation Through Performance Appraisal. New York, NY: John Wiley & Sons, 1977.
- Lempesis, C. "Peer Observation Improves Teacher Performance." National Association of Secondary School Principals Bulletin 68, 471 (1984): 155-156.
- Levine, R., Solomon, M., Hellstern, G., and Wollmann, H. Evaluation Research and Practice: Comparative and International Perspectives. Beverly Hills, CA: Sage Publications, 1981.
- Lieberman, M. "Educational Specialty Boards: A Way Out of the Merit Pay Morass?" Phi Delta Kappan 67, 2 (1985): 103-108.
- Lightfoot, S. The Good High School. New York, NY: Basic Books, Inc., 1983.
- Manatt, R. "Competent Evaluators of Teaching: Their Knowledge, Skills and Attitudes." In The Competent Evaluator of Teaching, pp. 9-41. Edited by W. Duckett. Bloomington, IN: Phi Delta Kappa, 1985.
- Manatt, R. "Lessons from a Comprehensive Performance Appraisal Project." Educational Leadership 44, 7 (1987): 8-14.
- Manatt, R., and Stow, S. Clinical Manual for Teacher Performance Evaluation. Ames, IA: Iowa State University Research Foundation, 1984.
- Manatt, R., and Stow, S. Developing and Testing a Model for Measuring and Improving Educational Outcomes of K-12 Schools: Technical Report. School Improvement Model, College of Education, Ames, IA, 1986.

- Manatt, R., Palmer, K., and Hidlebaugh, E. "Evaluating Teacher Performance with Improved Rating Scales." National Association of Secondary School Principals Bulletin 60, 401 (1976): 21-24.
- Martin, E. "Teacher Evaluation: A Selected Review of the Recent Literature." Journal of Classroom Interaction 18, 2 (1983): 16-19.
- McCormick, K. "Alexander Outlines His Master Teacher Scheme." The American School Board Journal 170, 10 (1983): 32.
- McFaul, S., and Cooper, J. "Peer Clinical Supervision: Theory vs. Reality." Educational Leadership 41, 7 (1984): 4-9.
- McGreal, T. Successful Teacher Evaluation. Arlington, VA: Association for Supervision and Curriculum Development, 1983.
- McGreal, T. "Teacher Performance Appraisal." In Instructional Leadership Handbook, pp. 76-77. Edited by J. Keefe and J. Jenkins. Reston, VA: National Association of Secondary School Principals, 1984.
- McIntire, R., Hughes, L., and Burry, J. "The Training and Certifying of Teacher Appraisers," Educational Leadership 44, 7 (1987): 62.
- McIntire, R., Smith, D., and Hassett, C. "Accuracy of Performance Ratings as Affected by Rater Training and Perceived Purpose of Rating." Journal of Applied Psychology 69, 1 (1984): 147-156.
- McNeil, J., and Popham, J. "The Assessment of Teacher Competence." In Second Handbook of Research on Teaching, pp. 218-244. Edited by R. Travers. Chicago, IL: Rand McNally & Co., 1973.
- Medley, D., Coker, H., and Soar, R. Measurement-Based Teacher Evaluation. New York, NY: Longman, Inc., 1984.
- Miller, G. "The Magical Number Seven, Plus or Minus Two: Some Limits on Our Capacity for Processing Information." Psychological Review 63 (1956): 81-97.
- Mobley, W. "Supervisor and Employee Race and Sex Effects on Performance Appraisals: A Field Study of Adverse Impact and Generalizability." Academy of Management Journal 25, 3 (1982): 598-606.
- Murnane, R., and Cohen, D. "Merit Pay and the Evaluation Problem: Why Most Merit Pay Plans Fail and a Few Survive." Harvard Education Review 56 (1986): 1-17.
- Murphey, K., Balzer, W., Kellam, K., and Armstrong, J. "Effects of the Purpose of Rating on Accuracy in Observing Teacher Behavior and Evaluating Teaching Performance." Journal of Educational Psychology 76, 1 (1984): 45-54.

- Murray, F. "Goals for the Reform of Teacher Education: An Executive Summary of the Holmes Group Report." Phi Delta Kappan 68, 1 (1986): 28-32.
- National Association of Secondary School Principals. NASSP Newsleader 34, 6 (1987a): 2.
- National Association of Secondary School Principals. NASSP Newsleader 34, 7 (1987b): 16.
- National Commission on Excellence in Education. A Nation at Risk: The Imperative for Education and Reform. Washington, D.C.: U.S. Government Printing Office, 1983.
- National Study of School Evaluation. A Self-Directed Program for Developing Teacher and Administrator Evaluation Procedures. Falls Church, VA: Author, 1984.
- Nieva, V., and Gutek, B. "Sex Effects on Evaluation." Academy of Management Review 5, 2 (1980): 267-276.
- Noriega, T. "The High Gain Teacher." Ph.D. Dissertation, Iowa State University, Ames, IA, 1987.
- Northfield Public Schools. Staff Development: The Northfield Model. Northfield Public Schools, Northfield, MN, 1983.
- Norvsis, M. User's Guide SPSS^x. New York, NY: McGraw-Hill, Inc., 1983.
- Odden, A. "Sources of Funding for School Reform." Phi Delta Kappan 67, 5 (1986): 335-340.
- Oliver, B. "Desirable Qualities in Teacher Performance Appraisal Systems." The Teacher Educator 18, 3 (1983): 26-31.
- Olson, L. "Performance-Based Pay Systems for Teachers Are Being Re-examined." Education Week 6, 29 (1987): 1-13.
- Parish, J. "Excellence in Education: Tennessee's Master Plan." Phi Delta Kappan 64, 10 (1983): 722-724.
- Pate-Bain, H. "A Teacher's Point of View on the Tennessee Master Teacher Plan." Phi Delta Kappan 64, 6 (1983): 725.
- Patten, T. A Managers Guide to Performance Appraisal. New York, NY: Free Press, 1982.
- Peck, T. "When Women Evaluate Women, Nothing Succeeds Like Success: The Differential Effects of Status upon Evaluations of Male and Female Professional Ability." Sex Roles 4, 2 (1978): 205-213.

- Persell, C., and Cookson, P. "The Effective Principal in Action." In The Effective Principal: A Research Summary. Reston, VA: National Association of Secondary School Principals, 1982.
- Plumlee, L. Improving Performance Evaluation Procedures. New York, NY: American Management Association, 1983.
- Popham, J. Educational Evaluation. Englewood Cliffs, NJ: Prentice-Hall, Inc., 1975.
- Pulakos, E. "A Comparison of Rater Training Programs: Error Training and Accuracy Training." Journal of Applied Psychology 69, 4 (1984): 581-88.
- Redfern, G. Evaluating Teachers and Administrators: A Performance Objectives Approach. Boulder, CO: Westview Press, 1980.
- Rothman, R. "Using Pupil Scores to Assess Teachers Criticized as Unfair." Education Week 6, 36 (1987): 1-18.
- Rowley, G. "The Reliability of Observational Measures." American Educational Research Journal 13, 1 (1976): 51-59.
- Rucker, D. "A Study of Principals' Teaching Style Preference and Teachers' Teaching Style as a Source of Bias in Teacher Evaluation." Ph.D. Dissertation, Iowa State University, Ames, IA, 1981.
- Saal, F., Downey, R., and Lahey, M. "Rating the Ratings: Assessing the Psychometric Quality of Rating Data." Psychological Bulletin 88, 2 (1980): 413-428.
- Sarbin, T., Taft, R., and Bailey, D. Clinical Inference and Cognitive Theory. New York, NY: Holt, Rinehart and Winston, 1960.
- Savage, D. "Teacher Evaluation: The Need for Effective Measures." Learning 12 (1983): 54-56.
- Savage, J. "Teacher Evaluation Without Classroom Observation." National Association of Secondary School Principals Bulletin 66, 458 (1982): 41-45.
- Semones, M. "Developing a Data Gathering Technique to be Used in Classrooms." Unpublished Ph.D. Dissertation, Iowa State University, Ames, IA, 1987.
- Shanker, A. "Separating Wheat from Chaff." Phi Delta Kappan 67, 2 (1985): 108-109.
- Shanker, A. "Teachers Must Take Charge." Educational Leadership 44, 1 (1986): 12-14.

- Shavelson, R., Webb, N., and Burstein, L. "Measurement of Teaching." In Handbook of Research on Teaching, pp. 50-91. Edited by M. C. Wittrock. New York, NY: Macmillan Publishing Company, 1986.
- Simon, A., and Boyer, G. Mirrors for Behavior: An Anthology of Classroom Observation Instruments. Philadelphia, PA: Research for Better Schools, Inc., 1970.
- Smith, B., Peterson, D., and Micceri, T. "Evaluation and Professional Improvement Aspects of the Florida Performance Measurement System." Educational Leadership 44, 7 (1987): 16-18.
- Soar, R. "Teacher Evaluation: A Critique of Currently Used Methods." Phi Delta Kappan 65 (1983): 239-246.
- Sofer, S. "Peer Teachers as Mirrors and Monitors: Second Year Evaluation Report." ERIC ED 260 098, 1985.
- Spring Hill Center. Building a Culture of Achievement: Recommendations for Improving K-12 Education in Minnesota. Wayzata, MN: Author, 1986.
- Stallings, J. "What Do We Mean by Quality in Education." In Rethinking Reform: The Principal's Dilemma, pp. 61-69. Edited by Walberg and Keefe. Reston, VA: National Association of Secondary School Principals, 1986.
- State of Iowa. House Bill 499. Des Moines, IA: Author, 1987.
- State of Tennessee. Comprehensive Education Reform Act of 1984. Nashville, TN: Author, 1984.
- Stow, S., and Sweeney, J. "Developing a Teacher Performance Evaluation System." Educational Leadership 38 (1981): 538-541.
- Strahan, R. "More on Averaging Judge's Ratings: Determining the Most Reliable Composite." Journal of Consulting and Clinical Psychology 48, 5 (1980): 587-589.
- Sweeney, J., and Stow, S. "Performance Improvement: A People Program." Education 101, 3 (1981): 267-269.
- Terborg, J., and Shingledecker, P. "Employee Reactions to Supervision and Work Evaluation as a Function of Subordinate and Manager Sex." Sex Roles 9, 7 (1983): 813-824.
- Thompson, J. "On Models of Supervision in General and on Peer-Clinical Supervision in Particular." ERIC ED 192 462, 1979.

- Travers, R. "Criteria of Good Teaching." In Handbook of Teacher Evaluation. Edited by J. Millman. Beverly Hills, CA: Sage Press, 1981.
- Vance, R., Winne, P., and Wright, E. "Longitudinal Examination of Rater and Ratee Effects in Performance Ratings." Personnel Psychology 36, 3 (1983): 609-620.
- Wexley, K., and Pulakos, E. "The Effects of Perceptual Congruence and Sex on Subordinates' Performance Appraisals of Their Managers." Academy of Management Journal 26, 4 (1983): 666-676.
- Wexley, K., and Yukl, G. Organizational Behavior and Personnel Psychology. Homewood, IL: Richard D. Irwin, Inc., 1984.
- Wise, A. "Teacher Evaluation: A Study of Effective Practices." Santa Monica, CA: The Rand Corporation. ERIC ED 246 559, 1984.
- Wise, A., and Darling-Hammond, L. "Teacher Evaluation and Teacher Professionalism." Educational Leadership 42, 4 (1985): 28-31.
- Zahorik, J. "Teaching: Rules, Research, Beauty, and Creation." Journal of Curriculum and Supervision 2, 3 (1987): 275-284.

ACKNOWLEDGMENTS

Many people provided support for this study, but leaders in the Dallas Independent School District were of utmost importance. In particular the contributions of then general superintendent Linus Wright and Dr. Sandra Berg, Director of Training and Development, were considerable as they provided leadership in implementing the teacher appraisal system and in sharing evaluation data with the School Improvement Model staff at Iowa State University.

In addition, the writer acknowledges the significant amount of support provided by family members in their patience and understanding throughout this project. Also, the project certainly would have faltered without the guidance and support of Dr. Richard Manatt. In his role as major professor, he provided the technical support needed to conceptualize and structure the investigation, and the inspiration needed to bring the project to completion.

Appreciation is also extended to the writer's doctoral committee, who in addition to Dr. Manatt, consisted of Dr. Gary Phye, Dr. Anthony Netusil, Dr. Detroy Green, and Dr. Ruth Swenson. Their collective insight helped shape the investigation in the initial phases, and their individual suggestions strengthened the study as it evolved.

In addition, appreciation is extended Katy Rice, Kevin Stow, Peggi Bevins, and Mary Peterson for preparing the data for computer entry. Beth Ruiz provided advice on computer management of the data, and Grant Nauman assisted in the display of graphics used in the study. Also, Bonnie Trede

117b

helped throughout the project with her efficient processing of the initial drafts and the final copy.

APPENDIX A. TEACHER APPRAISAL SYSTEM

	<u>Page</u>
Philosophy of Education	119
Philosophy of Instruction	122
Philosophy of Professional Employee Evaluation	124
Teacher Appraisal System: Purpose	124
Teacher Appraisal System: Procedures	125

Philosophy of Education

The Dallas Independent School District is committed to providing and structuring resources to enable each student to develop toward his or her maximum potential. A complementary belief is that each student has the responsibility to make full utilization of the resources provided. The District is committed to offering a full array of options for students kindergarten through twelfth grade. Each program of study is important; within each option teachers and students alike are striving for excellence.

DISD is a dynamic, changing, and growing school organization which recognizes and serves the needs of the different cultural groups in the community with the twin thrusts of equity and excellence of all program offerings. These philosophic premises suggest and support the following educational goals:

Intellectual Discipline:

1. Encourage the development and use of higher levels of thinking (e.g., critical, analytical, independent) including the use of the scientific method as a problem solving process for life situations.
2. Maintain respect for individual differences and adapt instructional programs for individuals and subgroups according to their needs and abilities.
3. Maintain high academic standards for all so that the purposes of the community of Dallas, the State of Texas, and the nation are well served.

4. Provide a school climate and classroom atmosphere in which individual creativity is fostered, expressed, and recognized.
5. Encourage all students to cultivate aesthetic interests as well as to strive for proficiency in the practical disciplines.
6. Help all students to recognize the intrinsic value of continuing their education and to acquire those tools necessary for the effective use and selection of life-time learning resources.

Moral and Ethical Values:

1. Provide an opportunity to develop an appreciation of the aesthetic, religious, and moral values arising from the age-old efforts of humankind to relate to the universe and humanity.
2. Encourage all students to reflect on the values of the community, state, and nation, affirming those qualities which they find credible and seeking change where they discover a need.

Citizenship and Civic Responsibility:

1. Develop an understanding of each person's rights and responsibilities in a democratic society and of the need to be punctual, diligent, and competent in the performance of the obligations incurred as members of the community and citizens of the state, nation, and world.
2. Develop a critical respect for authority and leadership.

Competence in Social Relations:

1. Provide social experiences which assist students in attaining maturity as they cope with childhood and adolescence with a developing set of values, appreciations, and tastes.

2. Provide social experiences which have relevance to adult living.
3. Encourage the development of responsible social behavior, balancing leisure time between self-satisfying activities and those that are helpful to society.

Career and Economic Awareness:

1. Promote an awareness of the many and diverse career opportunities available to all students in acquiring skills and knowledge which will expedite achievement of their post high school career ambitions.
2. Help students to develop an understanding of and a method for planning for those economic resources believed to be necessary for personal and/or family security and welfare.

Self-Realization:

1. Guide students towards a better understanding of themselves so that they may select achievable goals leading to meaningful, rich lives.
2. Help each individual overcome sometimes debilitating stereotypes such as handicapping conditions, gender, race, and socio-economic background.

Personal Health:

1. Provide those activities and experiences which will serve as a foundation for a lifelong program of physical fitness and health.
2. Provide a school climate in which feelings of security, personal worth, and accomplishment can flourish, thus promoting mental and emotional well-being.
3. Emphasize the importance of personal hygiene.

The district believes that excellence, high expectations, and potential of students must all interact in ways that are mutually beneficial to children, parents in the district, staff, and society. Therefore, staff development, task-oriented leadership and consideration and caring for all employees of the district is imperative to the fulfillment of this educational philosophy.

Philosophy of Instruction

The major role of the teacher in the Dallas Independent School District is to provide effective instruction which will facilitate learning. To guide and assist student learners in mastery of prescribed objectives, the teacher uses techniques such as modeling, demonstrating, probing, and questioning. This primary role requires recognition of individual differences, helping students develop productive behaviors, and fostering parental and community involvement in, and support for, the educational experiences of all students. Other roles of the teacher include serving as a role model, advisor, curriculum planner, positive peer, and a productive member of the educational team. Everything a teacher does and says becomes a part of the modeling for youth; therefore, teachers are expected to be good examples.

The District employs a structured, content-oriented process of instruction commonly referred to as the "Six Steps." The Six Steps of Successful Teaching are (1) teacher finds what students know about the subject, (2) teacher tells students what will be learned and why, (3) teacher demonstrates what will be learned, (4) students practice what is

being learned, (5) students apply what has been learned, and (6) teacher evaluates what has been learned.

The Six-Step instructional process is intended to provide dependable, high-impact teaching while also encouraging innovation and flexibility. Ideally, a teacher should assess students' needs and abilities and maximize instruction through a variety of classroom approaches and strategies. The District's methodology requires that classroom instruction move through a series of activities beginning with teaching the whole group on-level (teaching the specified content for a grade or course) to individualizing techniques at the application step.

The primary aim of instruction in the District is to guide every child toward the maximum of his or her educational potential. With this in mind, the Board of Education has targeted a goal for 1989, vis., to have 85 percent of all students perform at or above grade level on standardized achievement tests. Intermediate goals are specified annually in individual school improvement plans.

The teacher's instructional effectiveness in the Dallas Independent School District shall be determined by both teaching performance and positive employee behaviors. Instructional effectiveness will include (but not be limited to) classroom management, teaching techniques, fair and equitable grading procedures, teacher/student rapport, instructional planning, variety in methods and classroom activities, and student achievement. Effective employee behaviors will include following District and building policies and procedures, punctuality, and good work attendance.

Philosophy of Professional Employee Evaluation

The primary purpose of evaluation is to improve administration and instruction through assessment, communication, and motivation. Those persons charged with responsibility for performance evaluation will strive for valid, reliable, and objective assessments of all evaluatees.

Evaluation will encompass the educational professional's ability to establish a learning environment and perform in such a manner which contributes to the achievement of district goals. Evaluation must assure accountability; thus it must protect students from marginal performance on the part of teachers and administrators.

Evaluation should be motivating, comparative, and objective. Formative evaluation is to help all professionals improve performance while summative evaluation enables the board and administrative cabinet to make better decisions.

Central to all evaluation, measurement, and rating is the theme of student growth and achievement. Evaluation will be on-going, with continual formal and informal data gathering conducted by designated district personnel within the established guidelines. Evaluation information and criteria over time will set standards which will validate the district's teacher/administrator selection process. Evaluation information will facilitate the career planning and professional development of teachers and administrators.

Teacher Appraisal System: Purpose

The implementation of the appraisal system has three main purposes:

- (1) to improve the quality of instruction,

- (2) to provide direction to staff for professional growth, and
- (3) to provide information to serve as the basis for sound and defensible career ladder and employment decisions.

Teacher Appraisal System: Procedures

The following information outlines the procedures that will be followed during the 1985-86 school year.

Each teacher shall have at least two appraisals¹ during the school year unless unusual circumstances² intervene.

Teachers with an overall rating of below expectations will have:

1. Applicable policies regarding the evaluation process provided prior to students reporting.
2. A pre-observation individual conference, first observation, and post conference conducted prior to the end of the second grading period but no sooner than the third week of school.
3. A re-cycle of #2 by May 1.

Teachers new to the building will have:

1. Applicable policies regarding the evaluation process provided prior to students reporting.

¹ Appraisals are for the purposes of gathering data (formative), are conducted twice a year, and are not the final evaluation.

² Unusual circumstances are defined as absences of teacher that preclude correct number of observations and appropriate and/or scheduled conferences.

2. A pre-observation group conference, first observation, and post conference conducted prior to the end of the second grading period but no sooner than the third week of school.

3. A re-cycle of #2 by May 1.

Teachers reporting to the building after the orientation to the evaluation process will have an individual conference regarding the evaluation process.

All other teachers will have:

1. Applicable policies regarding the evaluation process provided prior to students reporting.

2. A first observation and post conference conducted prior to the end of the first semester.

3. A re-cycle of #2 by May 1.

Each teacher should be familiar with the explanations for training, observations, appraisals, conferences, and the summative evaluation.

TRAINING:

Teachers:

The building principal or designated supervisor shall acquaint each teacher/employee under his/her supervision with the evaluation procedures and with the instruments to be used.

Appraisers:

All appraisers must receive training.

OBSERVATIONS:**Formal Observations:**

...A formal observation consists of a minimum of 30 uninterrupted minutes. Each teacher will have a minimum of two formal observations unless unusual circumstances intervene.

...The principal must make one of the formal classroom observations.

...Whenever a teacher is formally observed, notes using the "Classroom Observation Form" must be taken so that suggestions will be based on facts. A copy of the "Classroom Observation Form" will be left for the teacher by the observer.

...Additional formal observations may be scheduled at the discretion of the evaluator or the request of the teacher.

Informal Observations:

...Informal observations and input without restrictions from persons familiar with the teacher's work such as supervisors, department chairs, or persons designated to provide assistance shall be used to assist in getting a total picture of the teacher's performance. Data gathered from informal observations must be shared verbally or in writing with the teacher.

APPRAISALS:**Teacher Self-Appraisal:**

...Teacher self-appraisal shall be completed prior to the first conference. The self-appraisal will be shared and discussed at the first conference. The self-appraisal is used for personal goal setting and is

not used by the evaluator as a basis for determining a teacher's overall performance for the year.

Two Formative Appraisals:

...A 2x2 (two formative appraisals each conducted by two appraisers) is required each year.

Names of Appraisers:

...Each teacher will be given the name/s of his/her appraisers.

Written Record of Observation (Formative Appraisal):

...The Written Record of Observation (the formative appraisal) will be completed by each appraiser after the required formal observation/s. The appraisers will jointly summarize each of the individual Written Record of Observation reports into one Written Record of Observation report. The purpose of the formative appraisal is to provide suggestions and recommendations for improvement. Formative appraisals are not cumulative and are not the final evaluation (summative).

Final Evaluation (Summative Report):

...All of each teacher's appraisals will be summarized into one final evaluation report--the Summative Evaluation. The Summative Evaluation will be completed by the principal except in unusual cases. Principals with 75 or more teachers may have his/her designee complete the final evaluation (Summative) and conduct the final evaluation conference. Principals are responsible for completing the Summative Evaluation and for

conducting the conference of teachers whose overall evaluation is less than satisfactory.

CONFERENCES:

Pre-Observation:

Individual:

Pre-observation (individual) conferences may be held with teachers evaluated less than satisfactory at any time prior to the observation.

Group:

Pre-observation (group) conferences may be held with teachers new to the building at any time prior to the observation.

Conferences--(First Formative Appraisal):

...The first formative appraisal conference will be concluded within five working days of the formal observation. The teacher's self-appraisal shall also be shared and discussed at this time.

...After discussion of the "Written Record of Observation," the record shall be signed and dated by both parties. A copy will be given to the teacher. The teacher's signature does not necessarily indicate agreement with the formative appraisal/observation but rather signifies awareness of the content.

Summative Evaluation:

...A minimum of one diagnostic and prescriptive conference will be held to discuss the final evaluation (Summative). The diagnostic/prescriptive conference will be held prior to May 1 of each

school year. The purpose of such conference is to advise the teacher of necessary improvements to move to the next level of performance by the close of the following school year in order to achieve career ladder advancement.

...The conference for the second formative appraisal may be combined with the final evaluation conference (Summative Evaluation).

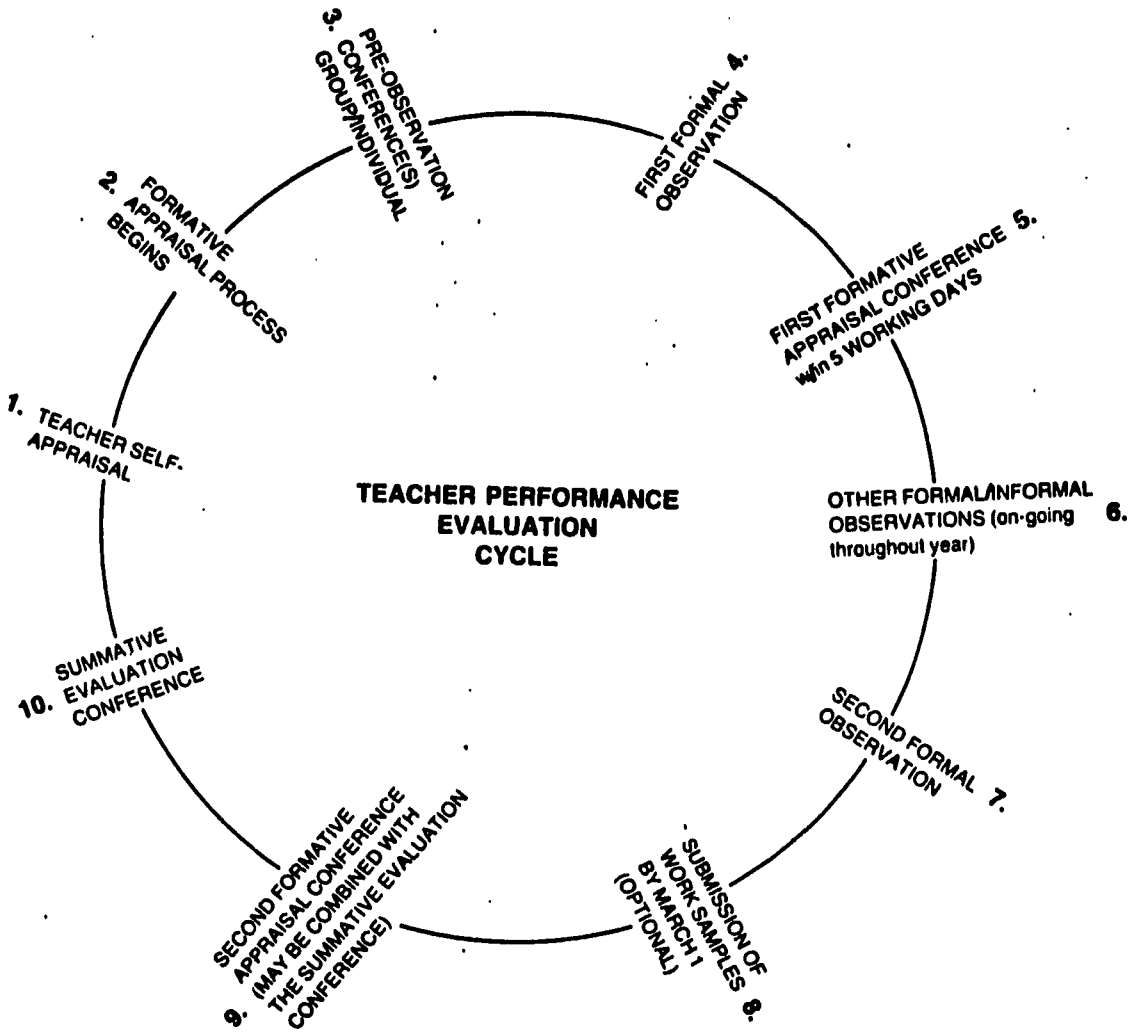
...Teachers may submit work samples or other input for consideration by March 1. All submitted work samples will be stamped (dated) by the evaluator. The input will be considered as a part of the Summative Evaluation.

...If ratings fall to a level that would preclude maintenance or advancement on the career ladder, then, at the written request of the teacher, both appraisers shall be present at the diagnostic/prescriptive conference. Written records pertinent to the evaluation must be available.

...After discussion of the Summative Evaluation, the evaluation shall be signed and dated by both parties. A copy will be given to the teacher. The teacher's signature does not necessarily indicate agreement with the Summative Evaluation but rather signifies awareness of the content.

...Before the Summative Evaluation becomes a part of the teacher's permanent file, she/he will have ten working days upon receipt of the evaluation to include a written response for clarification or to add information or opinion. As a professional courtesy, a copy should be sent to the principal. This response becomes a permanent part of the summative evaluation.

PERFORMANCE EVALUATION PROCESS CYCLE



APPENDIX B. INSTRUMENTS

	<u>Page</u>
DISD Test for Evaluators: Demographic Information	133
Classroom Observation Form	134
Written Record of Observation: Formative Appraisal	135
Written Record of Observation, Formative Appraisal: Criteria and Descriptors	136
Summative Evaluation Form	146

DALLAS INDEPENDENT SCHOOL DISTRICT
TEST FOR EVALUATORS

Directions: Complete the following items on side 1 of the answer sheet by filling in the appropriate circle on the answer sheet. Use only a number 2 lead pencil.

Name--Print last name first, space between names, fill in circles beneath.

Sex--Fill in "M" or "F."

Identification Number--Fill in social security number, fill in circles below also.

Special Codes:

K - Current Position

- | | |
|-------------------------|---------------------|
| 1 = Principal | 4 = Department Head |
| 2 = Assistant Principal | 5 = Central Staff |
| 3 = Dean of Instruction | 6 = Other |

L - Level of Assignment

- | | |
|----------------|-------------------|
| 1 = Elementary | 3 = High School |
| 2 = Middle | 4 = Central Staff |

M - Education--My most advanced degree is:

- 1 = BA/BS
- 2 = BA/BS plus 15 semester hours
- 3 = BA/BS plus 30 semester hours
- 4 = BA/BS plus 45 semester hours
- 5 = MA/MS
- 6 = MA/MS plus 15 semester hours
- 7 = MA/MS plus 30 semester hours
- 8 = MA/MS plus 45 semester hours
- 9 = Ph.D./Ed.D

N - Total years of experience in teaching/administration

- | | |
|-----------|-------------|
| 1 = 1-10 | 4 = 21-25 |
| 2 = 11-15 | 5 = 26-30 |
| 3 = 16-20 | 6 = Over 30 |

O - Years in current building assignment

- | | |
|----------|-------------|
| 1 = 1-4 | 4 = 16-25 |
| 2 = 5-8 | 5 = Over 25 |
| 3 = 9-15 | |

P - Racial/Ethnic

- | | |
|---------------------|--------------|
| 1 = American Indian | 4 = Hispanic |
| 2 = Asian | 5 = White |
| 3 = Black | |



Classroom Observation

Date: _____

Observer: _____

Teacher: _____

Subject: _____

Time	Observations
Total Time Observed	
Time from total that is actual engaged time	Summary of Observed Lesson
	<hr/> <i>Observer Signature</i>

WRITTEN RECORD OF OBSERVATION
FORMATIVE APPRAISAL

The Written Record of Observation (the Formative Appraisal) will be completed by each appraiser after the required formal observations, a minimum of two per year. The appraisers will jointly summarize each of the Individual Written Record of Observation reports into one Written Record of Observation report. The purpose of the formative appraisal is to provide suggestions and recommendations for improvement. Formative appraisals are not cumulative and are not the final evaluation (summative). Additional formative appraisals may be conducted during the year by the principal or designee.

Employee's Name _____ SS# _____

Last _____ First _____ M.I. _____

Teaching Assignment _____ School _____

Years in District _____ Years at this school _____

Principal _____

Appraiser's Name _____

Appraiser's Title and Assignment _____

Rating for each Criterion (O,E,S,B,U)

1. _____ THE TEACHER DEMONSTRATES EFFECTIVE PLANNING SKILLS
2. _____ THE TEACHER IMPLEMENTS THE LESSON PLAN
3. _____ THE TEACHER COMMUNICATES EFFECTIVELY WITH STUDENTS
4. _____ THE TEACHER USES EVALUATION ACTIVITIES APPROPRIATELY
5. _____ THE TEACHER DISPLAYS A THOROUGH KNOWLEDGE OF CURRICULUM AND SUBJECT MATTER
6. _____ THE TEACHER ENSURES STUDENT TIME ON TASK
7. _____ THE TEACHER IMPLEMENTS DISCIPLINE MANAGEMENT PROCEDURES
8. _____ THE TEACHER DEMONSTRATES SENSITIVITY IN RELATING TO STUDENTS
9. _____ THE TEACHER DEMONSTRATES EFFECTIVE INTERPERSONAL RELATIONSHIPS WITH ADULTS

Date of appraisal _____ Appraiser's Signature _____

COMPLETE THIS SECTION IF THIS IS THE SUMMARIZED RECORD OF BOTH APPRAISALS.

Date of conference _____ Conference conducted by _____

Signature of Appraiser _____

Second Appraiser's Signature _____

Teacher's Signature _____

WRITTEN RECORD OF OBSERVATION, FORMATIVE APPRAISAL:
CRITERIA AND DESCRIPTORS

Criterion I: DEMONSTRATES EFFECTIVE PLANNING SKILLS

Unsatisfactory:

1. Does not select long-range goals.
2. Does not write instructional objectives.
3. Does not use curriculum guides, texts, and materials adopted by the District to plan.
4. Does not plan for use of appropriate DISD Steps of Successful Teaching.

Below Expectations:

1. Consistently selects inappropriate long-range goals.
2. Writes instructional objectives that are not at the correct level of difficulty.
3. Selects learning content which is incongruent with the prescribed curriculum.
4. Plans for use of DISD Steps of Successful Teaching inconsistently.

Satisfactory:

1. Selects appropriate long-range goals.
2. Writes instructional objectives at the correct level of difficulty.
3. Selects learning content which is congruent with prescribed curriculum.
4. Plans for use of appropriate DISD Steps of Successful Teaching.

Exceeds Expectations:

1. Selects appropriate instructional objectives that are related to the long-range goals.
2. Plans review techniques and guided practice activities for the established instructional objectives.
3. Includes teaching methods and procedures congruent with curriculum guides, texts, and materials adopted by the District.
4. Plans appropriate time allotment for DISD Steps of Successful Teaching.

Clearly Outstanding:

1. Consults student files when selecting long-range goals to guide proper selection of instructional materials.
2. Utilizes both formative and summative evaluation procedures that reflect selected instructional objectives.
3. Includes a variety of teaching methods and procedures congruent with learning styles.
4. Plans for use of DISD Steps of Successful Teaching to meet group/individual needs.

Criterion II: IMPLEMENTS THE LESSON PLAN**Unsatisfactory:**

1. Does not state instructional objectives.
2. Does not use an organized series of instructional events.
3. Does not involve all students in class activities.
4. Does not provide feedback to students.

Below Expectations:

1. States instructional objectives inconsistently.
2. Uses an organized series of instructional events inconsistently.
3. Involves only high achieving students in class activities.
4. Lacks consistency in providing feedback to students.

Satisfactory:

1. States instructional objectives.
2. Uses an organized series of instructional events.
3. Involves all students in class activities.
4. Provides feedback to students.

Exceeds Expectations:

1. States instructional objectives and explains their importance.
2. Uses an organized series of instructional events which includes a smooth transition from one activity to another.
3. Involves all students by using techniques which check for their understanding.
4. Suggests study techniques as feedback, i.e., supplementary reading, use of library, peer tutoring.

Clearly Outstanding:

1. Serves as a resource to others in writing instructional objectives.
2. Uses an organized series of instructional events which emphasize lesson closure.
3. Involves all students within a class period by using a variety of instructional methods.
4. Provides feedback to students that encourages them to explore the concept further.

Criterion III: COMMUNICATES EFFECTIVELY WITH STUDENTS**Unsatisfactory:**

1. Is not clear when communicating with students.
2. Does not provide structuring comments to clarify the tasks.
3. Does not equitably distribute response opportunities.
4. Does not use a variety of verbal and nonverbal techniques.

Below Expectations:

1. Inconsistently is clear when communicating with students.
2. Provides structuring comments to clarify the tasks inconsistently.
3. Inconsistently distributes response opportunities.
4. Uses a variety of verbal and nonverbal techniques inconsistently.

Satisfactory:

1. Is clear when communicating with students.
2. Provides structuring comments to clarify the tasks.
3. Equitably distributes response opportunities among students.
4. Uses a variety of verbal and nonverbal techniques.

Exceeds Expectations:

1. Is clear when communicating with students and uses probing techniques.
2. Provides structuring comments that offer positive reinforcement.
3. Equitably distributes response opportunities and promotes active participation.
4. Uses a variety of verbal and nonverbal techniques to help the lesson proceed smoothly.

Clearly Outstanding:

1. Serves as a model for students in the use of language and manner of speaking to others.
2. Serves as a resource to others as to how to provide structuring comments.
3. Provides opportunity for students to develop skills in effective communication.
4. Motivates students by using a variety of verbal and nonverbal techniques when responding to questions or answers.

Criterion IV: USES EVALUATION ACTIVITIES APPROPRIATELY**Unsatisfactory:**

1. Does not use tests which reflect objectives that are taught.
2. Does not provide feedback on tests.
3. Does not check and return assignments in a timely manner.
4. Does not provide written feedback to students that helps them learn from checked assignments.

Below Expectations:

1. Inconsistently uses tests which reflect objectives that are taught.
2. Inconsistently provides feedback on tests.
3. Inconsistently checks and returns assignments.
4. Inconsistently provides written feedback to students regarding checked assignments.

Satisfactory:

1. Uses tests which reflect objectives that are taught.
2. Provides feedback on tests by giving written comments as well as points or scores.
3. Checks and returns assignments in a timely manner.
4. Provides written feedback to students regarding checked assignments.

Exceeds Expectations:

1. Uses tests which reflect objectives that are taught by using a combination of essay and objective items.
2. Reviews tests with students.
3. Assesses transfer of learning through assignments given.
4. Uses a variety of evaluation activities to ensure student progress.

Clearly Outstanding:

1. Uses pre- and post-tests to monitor student progress.
2. Makes opportunities for one-to-one conferences in regard to tests.
3. Asks students to evaluate their assignments.
4. Uses results from evaluation activities to modify instruction for group/individuals to ensure student progress.

Criterion V: DISPLAYS A THOROUGH KNOWLEDGE OF CURRICULUM AND SUBJECT MATTER**Unsatisfactory:**

1. Does not designate the purpose of the topic or activity.
2. Does not use curriculum guides or texts adopted by the District.
3. Does not identify subset of skills that are essential for accomplishing the instructional objective(s) of the lesson.
4. Does not have sufficient knowledge of content to meet the needs of students.

Below Expectations

1. Inconsistently explains topics or activities in context.
2. Inconsistently uses the curriculum guide or texts adopted by the District.
3. Inconsistently identifies subset of skills that are essential for accomplishing the instructional objective(s) of the lesson.
4. Provides instruction which inconsistently meets the needs of students.

Satisfactory:

1. Designates the purpose of the topic or activity.
2. Uses District adopted curriculum guides which include curriculum density.
3. Identifies the subset of skills that are essential for accomplishing the instructional objective(s) of the lesson.
4. Demonstrates sufficient knowledge of content to meet the needs of students.

Exceeds Expectations:

1. Relates specific topics or activities to content area.
2. Integrates concepts that require the use of skills learned in other content areas.
3. Uses a logical sequence of content to teach the lesson.
4. Provides instruction according to the Learner Standards.

Clearly Outstanding:

1. Serves as a resource in helping others to designate the purpose of the topic or activity.
2. Maintains curriculum alignment.
3. Demonstrates a knowledge of scope and sequence of curriculum and subject matter.
4. Serves as a resource in helping others to select content to meet the needs of students.

Criterion VI: ENSURES STUDENT TIME ON TASK

Unsatisfactory:

1. Does not manage time efficiently.
2. Does not organize students for effective instruction.
3. Does not establish procedures for students to follow on completion of tasks.
4. Does not devote class time to instructional activities.

Below Expectations:

1. Is inconsistent in the management of time.
2. Is inconsistent in organizing students for effective instruction.
3. Inconsistently establishes procedures for students to follow up on completion of tasks.
4. Inconsistently devotes class time to instructional activities.

Satisfactory:

1. Demonstrates effective time management skills.
2. Organizes students for effective instruction.
3. Establishes procedures so students know what to do upon completing a task.
4. Devotes class time to instructional activities.

Exceeds Expectations:

1. Minimizes management and transition time.
2. Guides/monitors concept/skill practice during class time.
3. Minimizes the time students need to wait for help to complete a task.
4. Focuses instructional activities on lesson objectives.

Clearly Outstanding:

1. Serves as a resource for using time management skills.
2. Maintains a classroom climate which ensures learning.
3. Reinforces students who spend time on task.
4. Provides options for students in fulfilling assignments.

Criterion VII: IMPLEMENTS DISCIPLINE MANAGEMENT PROCEDURES**Unsatisfactory:**

1. Does not communicate parameters for student classroom behavior.
2. Does not manage discipline problems in accordance with District policy.
3. Does not demonstrate positive relationships with students.
4. Does not define the limits of acceptable behavior and the consequences of misbehavior.

Below Expectations:

1. Inconsistently communicates parameters for student classroom behavior.
2. Manages discipline problems in accordance with District policy inconsistently.
3. Demonstrates positive relationships with students inconsistently.
4. Inappropriately defines the limits of acceptable behavior and the consequences of misbehavior.

Satisfactory:

1. Communicates parameters for student classroom behavior.
2. Manages discipline problems in accordance with District policy.
3. Demonstrates positive relationships with students.
4. Defines the limits of acceptable behavior and the consequences of misbehavior.

Exceeds Expectations:

1. Communicates parameters for student classroom behavior and rewards desired behavior.
2. Uses positive reinforcement to shape behavior.
3. Demonstrates positive relationships with students while promoting self-discipline.
4. Demonstrates alternative strategies when defining the limits of acceptable behavior.

Clearly Outstanding:

1. Anticipates problems and has a plan for dealing with the potential major problems.
2. Uses voice control, cues, hand signals, eye contact, and/or other techniques to establish desired behaviors.
3. Serves as a resource to others in learning how to implement discipline management procedures.
4. Implements management procedures that result in positive classroom climate.

Criterion VIII: DEMONSTRATES SENSITIVITY IN RELATING TO STUDENTS**Unsatisfactory:**

1. Does not exhibit a willingness to listen.
2. Does not make an effort to know each student as an individual.
3. Does not demonstrate awareness of the needs of all students.
4. Does not show respect for individuals.

Below Expectations:

1. Occasionally exhibits a willingness to listen.
2. Inconsistently makes an effort to know each student as an individual.
3. Occasionally demonstrates awareness of the needs of all students.
4. Inconsistently shows respect for individuals.

Satisfactory:

1. Exhibits a willingness to listen.
2. Makes an effort to know each student as an individual.
3. Demonstrates awareness of the needs of all students.
4. Shows respect for individuals.

Exceeds Expectations:

1. Uses active listening skills when working with students.
2. Makes an effort to know each student as an individual and provides opportunities for individual differences.
3. Demonstrates awareness of the needs of all students by adapting the content for a pluralistic society.
4. Shows respect for individuals by modeling proper behavior.

Clearly Outstanding:

1. Exhibits a willingness to listen to replies while providing constructive feedback.
2. Uses knowledge of individual students to capitalize on strengths and plans for students to use their strengths.
3. Serves as a resource for adapting the content for a pluralistic society.
4. Acknowledges the rights of others to hold differing views or values.

Criterion IX: DEMONSTRATES EFFECTIVE INTERPERSONAL RELATIONSHIPS WITH ADULTS**Unsatisfactory:**

1. Does not demonstrate cooperative behaviors with administrators, consultants, community members, and/or other teachers.
2. Does not demonstrate acceptance of the pluralistic and multi-cultural nature of the school, the District, and/or the community when performing daily tasks.
3. Does not demonstrate acceptance of different ethnic and/or cultural points of view.
4. Does not demonstrate by language and behavior a sensitivity to sex-role stereotyping.

Below Expectations:

1. Inconsistently demonstrates cooperative behaviors with administrators, consultants, community members, and/or other teachers.
2. Inconsistently demonstrates acceptance of the pluralistic and multi-cultural nature of the school, the District, and/or the community when performing any task.
3. Inconsistently demonstrates acceptance of different ethnic and/or cultural points of view.
4. Inconsistently demonstrates by language and behavior a sensitivity for sex-role stereotyping.

Satisfactory:

1. Demonstrates cooperative behavior with administrators, consultants, community members, and other teachers.
2. Demonstrates acceptance of the pluralistic and multi-cultural nature of the school, the District, and the community when performing daily tasks.
3. Demonstrates acceptance of different ethnic and/or cultural points of view.
4. Demonstrates by language or behavior a sensitivity to sex-role stereotyping.

Exceeds Expectations:

1. Fosters cooperation among administrators, consultants, community members, and other teachers.
2. Aids parents and other community members to value the pluralistic and multicultural nature of the school, the District, and the community.
3. Displays a knowledge of different ethnic and/or cultural points of view.
4. Influences others, through language and behavior, to become sensitive to sex-role stereotyping.

Clearly Outstanding:

1. Assumes a leadership role in creating cooperation among administrators, consultants, community members, and other teachers.
2. Brings parents and other community members together in ways to build upon the pluralistic and multi-cultural nature of the school, the District, and the community.
3. Participates actively to enhance the ethnic and/or cultural heritage of the school, the District, and the community.
4. Assumes a leadership role in eliminating sex-role stereotyping.



**SUMMATIVE
(FINAL EVALUATION)**

CONFIDENTIAL

Employee's Name _____
Last First MI Social Security No.

Teaching Assignment _____ School _____ Principal _____

Years of service in this school _____ Years of service in DISD _____

Total years in teaching profession _____

DEFINITIONS OF PERFORMANCE RATINGS

- CLEARLY OUTSTANDING (O):** 7 or more criteria rated Clearly Outstanding and no criterion rated below Exceeds Expectations.
- EXCEEDS EXPECTATIONS (E):** 7 or more criteria rated Exceeds Expectations or above with no criterion rated below Satisfactory.
- SATISFACTORY (S):** 7 or more criteria rated Satisfactory or above with no criterion rated as Unsatisfactory.
- BELOW EXPECTATIONS (B):** 4 criteria rated Below Expectations but no more than 3 criteria rated as Unsatisfactory.
- UNSATISFACTORY (U):** 4 or more criteria rated as Unsatisfactory.

DISTRIBUTION OF CRITERIA RATINGS

RATING	NUMBER	NOTES FOR OVERALL PERFORMANCE RATINGS:
CLEARLY OUTSTANDING (O)	_____	1. A BELOW EXPECTATIONS rating on any of the 1st nine (9) criteria means the best possible overall performance rating is SATISFACTORY . (See notes 3 and 4 regarding Criterion X.) 2. An UNSATISFACTORY rating on any of the 1st nine (9) criteria means the best possible overall performance rating is BELOW EXPECTATIONS . (See notes 3 and 4 regarding Criterion X.) 3. For an overall rating of CLEARLY OUTSTANDING, EXCEEDS EXPECTATIONS OR SATISFACTORY , descriptors a, b, c must be checked YES on criterion ten (10). 4. One No check for either a, b, c on criterion ten (10) means the best possible overall performance rating is BELOW EXPECTATIONS ; two or more No checks for a, b, c means the best possible overall performance rating is UNSATISFACTORY .
EXCEEDS EXPECTATIONS (E)	_____	
SATISFACTORY (S)	_____	
BELOW EXPECTATIONS (B)	_____	
UNSATISFACTORY (U)	_____	
OVERALL RATING: _____		

(Must include Criterion X — notes 3 and 4)

Recommendation of Principal

- Recommended for re-employment
- Below expectations
- Not recommended for re-employment

A formal conference was held on (date) _____ with my evaluator.

I acknowledge that the contents of the evaluation were discussed. I understand that my signature below does not necessarily mean that I agree with the evaluation. I also understand that I have the right to discuss my status with the Assistant Superintendent — Elementary/Secondary of the Dallas Independent School District.

Signed comments are attached by principal/evaluator _____ and/or teacher _____.

Date _____ Teacher's Signature _____

Evaluator's Signature _____

EMPLOYEE RESPONSIBILITIES

Criterion X: THE TEACHER FULFILLS EMPLOYEE RESPONSIBILITIES

The checklist for criterion X is to be completed when determining the overall performance rating on the final evaluation (summative).

	YES	NO
a. Follows applicable District policies in a professional manner that promotes operational efficiency in the school.	()	()
b. Follows administrative directives in a professional manner that promotes operational efficiency in the school.	()	()
c. Utilizes applicable policies and procedures to resolve issues and conflicts in a manner that promotes operational efficiency of the school.	()	()
d. Attends staff meetings.	()	()
e. Serves on staff committees and participates in school activities.	()	()
f. Maintains a continuous effort to improve professionally, through workshops, publication of articles, seminars, college courses, in-service training, and professional readings.	()	()
g. Maintains a condition of health that enables the teacher to meet the professional expectations of the District.	()	()
h. Provides accurate data to school and District as requested for management purposes.	()	()
i. Keeps the principal informed with respect to the needs of the classroom.	()	()
j. Communicates school policies to students and parents.	()	()
k. Other (specified by local school principal at beginning of school year).	()	()

EXPLANATIONS

To have an overall performance rating of **CLEARLY OUTSTANDING**, **EXCEEDS EXPECTATIONS**, or **SATISFACTORY** on the final evaluation (summative), descriptors a, b, and c must be marked **YES**. Also, the definitions on page one must be met.

Any **NO** on descriptors, d-k, requires that the principal provide directives on how to receive a yes.

**SUMMATIVE
(FINAL EVALUATION)**

SUMMARY OF RATINGS

PERFORMANCE RATINGS

O = CLEARLY OUTSTANDING E = EXCEEDS EXPECTATIONS S = SATISFACTORY

B = BELOW EXPECTATIONS U= UNSATISFACTORY

- _____ 1. THE TEACHER DEMONSTRATES EFFECTIVE PLANNING SKILLS
- _____ 2. THE TEACHER IMPLEMENTS THE LESSON PLAN
- _____ 3. THE TEACHER COMMUNICATES EFFECTIVELY WITH STUDENTS
- _____ 4. THE TEACHER USES EVALUATION ACTIVITIES APPROPRIATELY
- _____ 5. THE TEACHER DISPLAYS A THOROUGH KNOWLEDGE OF CURRICULUM AND SUBJECT MATTER
- _____ 6. THE TEACHER ENSURES STUDENT TIME ON TASK
- _____ 7. THE TEACHER IMPLEMENTS DISCIPLINE MANAGEMENT PROCEDURES
- _____ 8. THE TEACHER DEMONSTRATES SENSITIVITY IN RELATING TO STUDENTS
- _____ 9. THE TEACHER DEMONSTRATES EFFECTIVE INTERPERSONAL RELATIONSHIPS WITH ADULTS

EMPLOYEE RESPONSIBILITIES CHECKLIST

Criterion X: THE TEACHER FULFILLS EMPLOYEE RESPONSIBILITIES

Descriptors a, b, and c are checked as follows:

	YES	NO
a	()	()
b	()	()
c	()	()

Note to evaluator

Use the summary on this page to complete page one.
Please read notes, 1-4, on page one carefully.

APPENDIX C. LETTER OF COMMUNICATION



November 4, 1986

Dr. Richard Manatt
 Director
 SIM Projects
 College of Education
 Iowa State University
 E005 Lagomarcino Hall
 Ames, Iowa 50011

Dear Dick,

I have read David Peterson's dissertation/thesis outline. Yes, the information regarding inter-rater reliability and the effects of our training efforts will be of interest to us. Therefore, the data obtained as a part of our joint project may be used as outlined. Naturally, we will want a separate report.

Mr. Wright does expect a report regarding the summative evaluation - i.e.; numbers of ratings for each category and whether or not training produced a more competent evaluation. I'm making the assumption that Dave's analysis serves this purpose.

By the way, AASA filming went as I expected - fine! Jerry Melton did come. See you in a few weeks.

Sincerely,

Sandra Lanier-Berg
 Manager, Training
 and Development
 Human Resources

SLB/acg

cc: Dr. Deberie Gomez

'84-'85

100
 YEARS
 ★ ★ ★ ★

Dallas Independent
 School District

Linus Wright
 General Superintendent

3700 Ross Avenue
 Dallas, Texas 75204
 (214) 824-1620

APPENDIX D.

DISD SUMMARY OF EVALUATOR TRAINING, 1985-86

DALLAS INDEPENDENT SCHOOL DISTRICT
SUMMARY OF EVALUATOR TRAINING, 1985-86

<u>Topic Area</u>	<u>Clock Hours of Instruction</u>
Teacher Effectiveness Research	3.0
Conferencing Strategies	8.5
Professional Growth Plans	4.0
Data Gathering and Analysis	13.5
Effective Teaching Strategies	4.0
DISD Criteria, Descriptors, Procedures	10.0
Learning Styles	2.0
School Climate	<u>3.0</u>
Total	48.0 Clock hours

APPENDIX E.

CRITERIA FOR CAREER LADDER PLACEMENT AND ADVANCEMENT

Each teacher shall be assigned to a career ladder level based on PERFORMANCE, EXPERIENCE, JOB-RELATED EDUCATION, ADVANCED ACADEMIC TRAINING, AND JOB ASSIGNMENTS.

LEVEL ONE

ENTRY

1. Level 1 Certificate;

MAINTENANCE

1. Continued satisfactory performance during first 2 years or nonrenewal of contract.

CERTIFICATION

1. Completion of probationary year with satisfactory performance in all categories.

Valid for 3 years and renewable once with 6 semester hours or 90 advanced academic training hours or combination.

LEVEL TWO

ENTRY

1. Level 2 Certificate; and
2. Exceeds expectations in previous year prior to consideration of Level II placement; and
3. Either: Bachelor's degree
No evaluation lower than satisfactory for the most recent three-year period
3 years creditable classroom teaching experience
9 semester or 135 training hours or combination¹

-----OR-----

Master's degree or Doctorate degree in designated area
No evaluation lower than satisfactory for the most recent two-year period
2 years creditable classroom teaching experience

CERTIFICATION

1. Level 1 Certificate
and
2. Bachelor's and 3 years experience, or master's and two years experience, or doctorate and 1 year experience;
and
3. District recommendation

¹See note at end of section.

LEVEL 2 (Continued)**MAINTENANCE**

1. At least satisfactory performance every year.
2. Teacher to be reassigned to level 1 if performance is below expectations.

Valid for 5 years and renewable with 6 semester or 90 academic training hours or combination¹

LEVEL THREE**ENTRY**

1. Level 3 Certificate; and
2. Exceeds expectations for 3 of prior 4 years with no lower than satisfactory in other year; and
3. Five years teaching experience at level 2; and
4. 6/90 training hours¹

-----OR-----

1. Level 3 Certificate; and
2. Clearly outstanding for 2 of prior 3 years with no lower than satisfactory in other year; and
3. Three years teaching experience at level 2; and
4. 3/45 training hours¹

CERTIFICATION

1. Level 2 Certificate; and
2. Bachelor's and 8 years experience, or master's and 5 years experience, or doctorate and 3 years experience; and
3. District recommendation

MAINTENANCE

1. Better than satisfactory performance at least 1 of every 2 consecutive years and never below satisfactory.
2. Teacher to be reassigned to level 1 if performance is below expectations.
3. Teacher to be reassigned to level 2 if teacher has satisfactory or below performance at level 3 for two consecutive years.

Valid for 5 years and renewable with 6 semester or 90 academic training hours or combinations¹

LEVEL FOUR (Master Teacher)**ENTRY**

1. Master Teacher Certificate; and
2. Clearly outstanding in 2 of prior 3 years with at least satisfactory in other year; and
3. Three years teaching experience at level 3; and
4. Satisfactory performance on Master Teacher Exam; and
5. 6/90 training hours¹

-----OR-----

1. Master Teacher Certificate; and
2. Clearly outstanding for 3 consecutive years; and
3. Two years experience at level 3; and
4. Satisfactory performance on Master Teacher Exam; and
5. 3/45 training hours¹

MAINTENANCE

1. Clearly outstanding performance for 2 of every 3 years and not below satisfactory in other year; and
2. Teach in classroom at least 60% of day; and
3. Two Master Teacher duties every 3 years² and
4. 3/45 training hours¹

-----OR-----

1. Clearly outstanding each year; and
2. 60% teaching time; and
3. Two Master Teacher duties every 3 years²

Valid for life

Teacher to be assigned to level 3 if any of the above requirements are not met.

CERTIFICATION

1. Level 3 Certificate; and
2. Bachelor's and 11 years experience, or master's and 8 years experience, or doctorate and 5 years experience (in approved program of study); and
3. District recommendation

NOTE:

Teachers who are demoted on the career ladder must re-qualify for entry into the higher level under performance standards. If the district determines that extraordinary personal circumstances caused the lower rating and performance is clearly outstanding in the next year, the teacher may be reinstated.

¹Professional Training Hours: In all cases, the requirements for professional training hours are specified as higher education coursework (semester hours) or advanced academic training hours (inservice or other), or a combination of both for an equivalent ratio of one semester hour for every fifteen academic training hours.

²Master Teacher Duties: Master Teacher duties shall be defined by the State Board of Education and shall include supervising student teachers; team leader, mentor, or department chairman; conducting and advanced academic training; or assessing Master Teacher candidates.

Beginning September 1, 1984, fifty (50) percent of the coursework or training must be in the area of subject taught/certification unless the evaluation identifies a specific need in another area.