INFORMATION TO USERS

This reproduction was made from a copy of a manuscript sent to us for publication and microfilming. While the most advanced technology has been used to photograph and reproduce this manuscript, the quality of the reproduction is heavily dependent upon the quality of the material submitted. Pages in any manuscript may have indistinct print. In all cases the best available copy has been filmed.

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

1. Manuscripts may not always be complete. When it is not possible to obtain missing pages, a note appears to indicate this.

2. When copyrighted materials are removed from the manuscript, a note appears to indicate this.

3. Oversize materials (maps, drawings, and charts) are photographed 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 also filmed as one exposure and is available, for an additional charge, as a standard 35mm slide or in black and white paper format.*

4. Most photographs reproduce acceptably on positive microfilm or microfiche but lack clarity on xerographic copies made from the microfilm. For an additional charge, all photographs are available in black and white standard 35mm slide format.*

***For more information about black and white slides or enlarged paper reproductions, please contact the Dissertations Customer Services Department.**

Hoffman, Kay Trudy

A COMPARISON OF THE EFFICACY OF TWO TYPES OF TEACHER EVALUATION INSTRUMENT FORMATS

Iowa State University PH.D. 1986

A comparison of the efficacy of two types of teacher evaluation instrument formats

by

Kay Trudy Hoffman

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

1986

TABLE OF CONTENTS

PAGE

CHAPTER I. INTRODUCTION 1

Statement of the Problem 4

Purposes of the Study 5

Research Hypotheses 7

Definition of Terms 8

Delimitations of the Study 10

CHAPTER II. REVIEW OF LITERATURE 12

Introduction 12

History and Background 12

Teacher Evaluation Approaches 17

Instrumentation 21
Narrative 25
Checklists 26
Rating scales 27

Criteria Selection..... 34

Formative and Summative Evaluation 39

Further Research Topics 41

Summary 42

CHAPTER III. METHODS AND PROCEDURES 45

Collection of Data 45
Materials 45
In summary 51
Sample 52
Expert panel 57
Procedures 58

Analysis of Data 60

PAGE

CHAPTER IV. FINDINGS 63

Descriptive Data 65

Inferential Statistics 76
Hypotheses 77
Hypotheses testing 78

CHAPTER V. SUMMARY, CONCLUSIONS, AND RECOMMENDATIONS 84

Summary and Conclusions from the Data 84
Results from hypotheses testing 85
Discussion 86

Limitations 89

Recommendations for Further Research 90

BIBLIOGRAPHY 93

ACKNOWLEDGEMENTS 100

APPENDIX A: GRAPHIC RESPONSE MODE (GRM)/INDICATOR INSTRUMENT FORMAT 102

APPENDIX B: DOUBLE SCALE RESPONSE MODE/FORCED INDICATOR RATING INSTRUMENT FORMAT USING A POINT SCALE 104

APPENDIX C: DOUBLE SCALE RESPONSE MODE/FORCED INDICATOR RATING INSTRUMENT FORMAT USING A CONTINUOUS SCALE 106

APPENDIX D: EXPLANATION OF THE RATING SCALE CATEGORIES USED IN THE GRM AND DSRM FORMATS 108

APPENDIX E: INDICATOR EXPLANATION SHEET USED FOR THE GRM INSTRUMENT 110

APPENDIX F: IMPROVEMENT AND STRENGTH AREAS REPORTING FORM 112

APPENDIX G: REGISTRATION CARD FOR DEMOGRAPHIC INFORMATION 114

APPENDIX H: INFORMATION/DIRECTION SHEET 116

LIST OF TABLES

		PAGE
TABLE 1.	Number and percent of participants by job title and job level	54
TABLE 2.	Number and percent of participants by district size and years in supervision	55
TABLE 3.	Number and percent of participants by sex, number of teachers supervised, and previous experience	56
TABLE 4.	Ratings of all participants on the criterion, "Communicates Effectively with Students" (both formats)	65
TABLE 5.	Ratings of the overall criterion, "Communicates Effectively with Students," for GRM/Indicator format and DSRM/Forced Indicator Rating format	66
TABLE 6.	Summary of communication areas identified to improve by the expert panel and by evaluators using the two instrument formats	68a
TABLE 7.	Summary of communication areas identified to strengthen or reinforce by evaluators using the GRM, DSRM, and by the expert panel (first, second and third choices combined)	70
TABLE 8.	A comparison of performance ratings on the criterion and eight indicators by evaluators who used both forms of the DSRM/Forced Indicator Rating instrument format	72
TABLE 8a.	Rating differences of evaluators using DSRM point scale compared to using the DSRM continuous scale (based on frequencies). *N=56	75a

TABLE 9. Analysis of variance using criterion ratings 79

TABLE 10. Analysis of variance of mean differences by group 80

TABLE 11. Summary of (Z) calculations, by format, of areas identified for improvement 81

TABLE 12. Summary of (Z) calculations, by format used, of areas identified for reinforcement 83

CHAPTER I. INTRODUCTION

One of the most discussed, researched, and publicized topics in education today is teacher evaluation. Though evaluation of teachers has occurred since the founding of public schools in the late 1800s, concerns by both the lay public and professionals continue to surface centering on the reliability of evaluative procedures, instruments, and assessment of performance. Many of these concerns focus on the criteria used in evaluating teacher performance, definitions of observable characteristics of effective teaching, instrument subjectivity, and the shortcomings and lack of information about teacher evaluation instruments.

Over the past eight decades, the critical criteria employed in evaluating teachers have constantly shifted - focusing, first, on concerns about school maintenance activities, then teacher behavior inside and outside of the classroom, then materials and content development and classroom performance in general, and, finally, competency and student achievement spawned by state-mandated teacher performance evaluation. While the substance of evaluative criteria has fluctuated throughout the 1980s, it is clear the stress on performance evaluation (including evaluating teacher competency and student achievement) will not be upstaged easily. This may be verified by the fact that, by 1984, thirty-four states and the District of Columbia had enacted

statutes requiring state or locally developed teacher evaluation systems (Wise, Darling-Hammond and Pease, 1982). Furthermore, states such as Florida, Kentucky, and North Carolina require the use of a detailed coding system to evaluate beginning teachers.

Because of this press for effective teacher evaluation, particularly in the last ten years, educators have voraciously sought to define effective teaching. Research on what constitutes "good teaching" is not scarce. Such experts as Denham and Lieberman (1981), Hunter and Russell (1977), Rosenshine (1970, 1979), Brophy and Evertson (1974, 1976), Brophy (1978), Good and Power (1976), Medley (1979), McGreal (1983), Stallings (1977), Popham (1974, 1975), Borich (1977), Dunkin and Biddle (1974), Glass (1977), Redfern (1972, 1980), Iwanicki (1981), Stow and Sweeney (1981), and Manatt, Palmer and Hidlebaugh (1976), and Manatt (1981) have provided researched evidence that teaching behaviors do make a difference in student achievement and, consequently, are an issue to be reckoned with in teacher evaluation. Seemingly missing in the review of both study and opinion on teacher evaluation, however, is research on instrument design and the scale used in summative evaluation.

Numerous studies have been conducted on performance evaluation instruments - specifically, rating scales. But these studies have found application primarily in business and industry where performance is product-specific, thus, more

amenable to rating types of formats. Evaluation of teaching performance does not lend itself as easily to objective instrument development as Carfield and Walter (1984) found after examining one hundred twenty-seven teacher evaluation forms. They concluded that "many teacher evaluation forms are poorly constructed, too vague, and subjective."

In summary, primary concerns in teacher evaluation have centered upon areas other than instrumentation such as its purpose and identification of observable characteristics of effective teaching. Conspicuously absent in the research of the past decade are studies examining the efficacy of teacher evaluation instrument format - specifically, instrument reliability and validity and the ability of instrument format to assist in the differentiation of qualitative levels of teacher performance.

Time and again, authorities in evaluation, such as Popham (1975) and Dunkleberger (1982), asserted that most teacher evaluation instruments fail to identify and improve teaching behaviors. Borich (1977), a noted authority on teacher performance evaluation, accentuated the need for further research in teacher evaluation instrumentation when he stated, "There is a pressing need to develop performance evaluation instruments which are valid and reliable."

Statement of the Problem

State statutes and legal opinions have narrowly defined the focus of teacher performance evaluation -- to improve instruction and to assist in making decisions related to the continued employment of a teacher (Wise, Darling-Hammond & Pease, 1982). Evaluation instruments and procedures must be valid and reliable to achieve both. Yet, despite the presence of objective criteria, we know little about instrumentation. Further, while research studies have documented teacher behaviors which directly affect student performance and achievement within the classroom, the forms used to record these behaviors appear to be poorly constructed. Given the importance of teacher evaluation for improving instruction and assessing accountability, coupled with the lack of significant guiding research on instrument format, a study focusing on teacher performance evaluation instrumentation and scale development, comparing summative evaluation instrument formats, was warranted. Such investigative research will assist others in making decisions and drawing conclusions about the influence of instrument format in validating ratings of teacher performance.

Purposes of the Study

The purposes of this study were to examine the efficacy of two types of evaluation instrument formats to determine 1) which instrument format assisted evaluators in making valid ratings of a teacher's performance on a specified criterion, 2) which instrument format led to greater agreement among evaluators in their ratings (inter-rater reliability), 3) if format assisted raters in identifying teaching behaviors to improve and to reinforce, and 4) if the use of a continuous scale resulted in evaluator ratings different from those made using a point scale. Expert panel ratings of performance on a videotaped lesson and identification of teaching behaviors to improve and reinforce were considered to be the "correct" responses. The instruments used in this study were the Graphic Response Mode/Indicator (GRM/Indicator), and two forms of an instrument format entitled Double Scale Response Mode/Forced Indicator Rating (DSRM/Forced Indicator Rating). One of the DSRM formats used a four point rating scale and the other format used a continuum on which to record ratings. The Graphic Response Mode/Indicator format included a specific performance criterion, four rating categories (a point scale), and brief written statements to describe performance at each of the rating levels. In addition, a separate page contained eight indicators, which described effective performance on the criterion, and were to be used for reference purposes by

participants who used the GRM/Indicator instrument. The Double Scale Response Mode/Forced Indicator Rating format included a specific performance criterion, four rating categories, and the eight performance indicators (listed on the same page as the stated criterion) which were to be rated before the evaluators using this instrument format rated performance on the specified criterion. (One format of this instrument used a four point rating scale on which to record ratings; the other used a continuum.)

Evaluators participating in this study were randomly divided into two groups - one using the GRM/Indicator format, the other using the DSRM/Forced Indicator Rating formats. All evaluators were asked to identify performance or teaching behavior areas to improve and to reinforce; an Improvement and Strength Areas Reporting Form was given to both groups of evaluators for the purpose of recording the identified improvement and reinforcement areas.

Evaluators were given the two types of instrument formats for the purpose of collecting information/data to assess the following:

1. The efficacy of the two instrument formats by comparing ratings on the specified criterion.
2. Inter-rater reliability of ratings of evaluators who used the two different instrument formats.
3. Which instrument format assisted raters in identifying areas for growth or improvement in

the teaching performance related to the specified criterion.

4. Which instrument format assisted raters in identifying strengths or areas to reinforce in the teaching performance related to the specified criterion.
5. If the use of a continuous rating scale resulted in ratings different from those on the point scale and, if so, in what direction those ratings were drawn.

Research Hypotheses

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

1. The mean score ratings on the criterion by evaluators who used the DSRM/Forced Indicator Rating format will be significantly closer to those of the expert panel mean score ratings than the mean score ratings by those who used the GRM/Indicators (validity).
2. There will be significantly less variance in ratings of performance on the specified criterion among evaluators who used the DSRM/Forced Indicator Rating format than those who used the GRM/Indicator format (inter-rater reliability).
3. Identified job improvement targets by evaluators who used DSRM/Forced Indicator Rating formats will be significantly closer to those of the expert panel than those who used the GRM/Indicator format.

4. Identified reinforcement areas by evaluators who used DSRM/Forced Indicator Rating formats will be significantly closer to those of the expert panel than those who used the GRM/Indicator format.

Definition of Terms

1. CLASSROOM EVALUATION: The appraisal of teacher performance within a classroom setting.
2. INSTRUMENT: The tool used to record collected data on teacher performance based on a series of classroom observations.
3. ADMINISTRATOR, SUPERVISOR, EVALUATOR, RATER: Any person responsible through authority, power, or position for assessing teacher performance.
4. CRITERION: An identified, specific area of teacher performance upon which evaluation is conducted.
5. INDICATORS: Descriptors of effective performance on a specified criterion.
6. STANDARD: The measure used as a comparison when judging the quality, quantity, or value of a specified criterion.
7. DISCRIMINATE: The ability to show or distinguish differences in teacher performance.
8. JOB IMPROVEMENT TARGETS: Observed teaching behaviors or techniques determined by the rater as needing improvement in order for the teacher to achieve acceptable standards.

9. REINFORCEABLE AREAS: Specific, observed, effective teaching practices which should be maintained or expanded within the classroom setting as determined by the rater.

10. GRAPHIC RESPONSE MODE/INDICATOR: A format using brief statements which explain or define the criterion to assist in rating performance at various rating levels; indicators (descriptors of effective performance on the criterion) are provided on a separate page to assist in rating performance on the specified criterion.

11. DOUBLE SCALE RESPONSE MODE/FORCED INDICATOR RATING FORMAT (point scale): A format providing indicators, descriptors of effective performance on a criterion, on the same page as the criterion and which must be rated before performance on the criterion is rated. A four point rating scale was provided.

12. DOUBLE SCALE RESPONSE MODE/FORCED INDICATOR RATING (continuous scale): A format providing indicators, descriptors of effective performance on a criterion, on the same page as the criterion which must be rated before performance on the criterion is rated. Although four points on the scale were provided, evaluators could choose any point on the continuous line to record the rating.

Delimitations of the Study

Limits on the application and generalizability of the findings are due to several delimiting factors.

1. Prior training and experience of the participants may have had a bearing on familiarity with evaluation procedures and terminology.
2. Participants were selected from a ten-district area in a midwestern state and may, therefore, be expected to have greater congruence in goals, expectations, background and/or philosophies regarding evaluation concepts then might have been expected with participants randomly selected from a broader sample.
3. Participants were employed in districts which, in the last two years, experienced mandated merit pay procedures. Those procedures were eliminated recently. This may have left attitudes about or impressions related to teacher evaluation.
4. The expert panel was comprised of three professors of educational administration from Iowa State University. They may have engendered a more compatible philosophy, value system, and expectations for performance than might have

been expected had the panel members been randomly selected from a broader sample.

5. The videotape used for evaluation in this study depicted a single teacher functioning at a specific grade level and within a specific subject area. Consequently, the results may not be generalizable to other levels and disciplines.
6. The lack of specific rating category limits in the Double Scale Response Mode/Forced Indicator Rating instrument format using a continuous scale limited the usefulness of the scale for analysis.

CHAPTER II. REVIEW OF LITERATURE

Introduction

Chapter II presents the review of literature as it relates to six areas central to teacher evaluation - history and background, teacher evaluation approaches, instrumentation, criteria selection, formative and summative evaluation, other aspects of the literature particularly germane to the study, and final a summary.

History and Background

An understanding of the instrumentation, format and scales used in the evaluation of teachers is enhanced by a brief review of the history of teacher evaluation. As a supervisory activity, teacher evaluation has existed in some form or another since the early 1900s (DiRocco & Igoe, 1977). The first form of evaluation conducted with public school teachers in the United States involved an in-class observation by a supervisor visiting the subordinate teaching staff member at least once a year for the purpose of control and inspection. Following the observation, the supervisor prepared a written report based on criteria reflecting completion of specified school duties (many of which centered on building maintenance) and teacher behavior exhibited both

inside and outside of the classroom. The public's perceptions of both teacher and school operation - important issues to community members - were addressed, however subjectively, and noted by the supervisor (Lamb & Swick, 1975).

From the early 1900s to the late 1940s, the focus of teacher evaluation shifted from operational emphasis to performance issues due to the influence of dramatically changing trends. These trends included an increase in student numbers as cities and towns grew in population, technological advancements which impacted every American, and a national awareness that structure and formalized operation were needed in all organizations - including schools - for effective management. Consequently, supervision and evaluation benchmarks placed emphasis on identifying those procedures which would ensure sound teacher performance of particular educational tasks such as task analysis, behavior management, and teaching to objectives, thus, providing teachers with guidance leading to improvement of these particular tasks (Lucio & McNeil, 1979). Philosophies of education - emphasizing improvement in teaching techniques, materials selection, facility design, and curriculum development - began to emerge. Evaluation instruments were developed to aid supervisors in describing teacher and student behaviors. As early as 1925, the rating scale was the most commonly used form of recording teacher performance in the classroom (Spears, as cited in McLaughlin, 1982). And, by the 1930s,

though rating instruments were still used, they were constructed to describe, more closely, classroom behavior based on observation of identified criteria (Reemers, as cited in McLaughlin, 1982).

By the 1950s, however, technology and world competition demanded a closer scrutiny of "what" was being taught. Overcrowded classrooms, shortages of well-qualified teachers and half-day school sessions became common creating concerns over whether or not teachers were performing effectively in addressing student needs - particularly in content areas (Shepherd & Ragan, 1982). Racial segregation, the advent of Sputnik, and automation even more strongly directed public attention to education leading to cries for sweeping reform from California to New York. To meet these demands, the "new" issues in supervision and evaluation centered around the use of objectives, joint teacher-supervisor responsibilities, and differentiated supervision - subjecting teaching performance to the scrutiny of principals, department heads, and/or powerful interest groups (Lucio & McNeil, 1979).

Meanwhile, the tools for measuring teacher behavior and student-teacher interaction experienced a simultaneous revolution in the 1930s, 1940s, and into the 1950s, witnessed by a plethora of well-documented studies and instruments developed to measure these interactions - 1934, Pupil-Teacher Rapport Scale; 1945, Anderson and Brewer Scale; and 1949, Withall Climate Index (Walberg, 1974). Instruments for

evaluating teachers continued to evolve with the most sophisticated versions emerging in the 1960s for the purpose of assessing teacher influence in the classroom. The most frequently used observation instrument at that time was the Flanders Interaction Analysis System which distinguished between "direct and indirect" teaching influence. This instrument emerged as the most noteworthy "point of departure" in complex instrument development in teacher evaluation because of this emphasis on teacher influence (Walberg, 1974).

The 1960s saw social issues, human rights, protests, and sporadic violence, dominating legal, social, and educational fronts. Again, schools responded - forced primarily, by federal legislation and the demand for improved teacher assessment techniques. Evaluators were encouraged and, frequently, directed - by law - to document performance in the classroom, though the results were, by today's standards, relatively unsophisticated. Such a renewed emphasis on stringent evaluation measures was frequently viewed by teachers and their organizations as threatening to job security and performance further complicating performance evaluation efforts.

As the nation slowly recovered from the dramatic events of the sixties, the 1970s ushered in an era stressing educational accountability which led to research and data collection relating to teacher performance evaluation - heavily focusing on criteria for evaluation. Menne (1972),

Borich (1977), Rosenshine (1979), Popham (1974), Brophy and Evertson (1974), Good & Power (1976) and Manatt, Palmer and Hidlebaugh (1976) conducted a wealth of studies researching effective teaching behaviors within the classroom as well as validating the criteria used to assess the effects of teacher behavior on student achievement. The conclusions reached in the majority of these studies indicated that teacher behavior does impact student learning and that certain behaviors have greater impact than others. These findings influenced supervisors in their approach to evaluation; it seemed likely that teacher evaluation procedures should include specific data collection to determine qualitative levels of performance based on validated criteria.

The aforementioned studies provided the impetus for the events of the next decade. The 1980s produced a series of well-publicized, volatile, national reports whose conclusions and recommendations had sweeping implications for the content and context of teacher evaluation. The Report of the President's National Commission on Excellence in Education (Coleman, 1983), "A Quest for Common Learning" (Boyer & Levine, 1981), Twentieth Century Fund Task Force Report (1981) and several state task force reports were highly critical and pointed to bold, new directions for the educational and public communities. This forced the reexamination of program content priorities, teaching strategies, and evaluative techniques and outcomes.

In summary, teacher evaluation has been marked by a progression of trends, events, and emphases. The first evaluative procedures focused on periodic observation of how well a teacher performed particular duties. The wave of technological advancements was the next trend leading to increased concern for improved performance. Rating scales were frequently used to record assessments of teacher performance in the 1940s, 1950s and into the 1960s. Foci of these rating scales changed from how teachers interacted with students to what teachers were doing in the classroom. Social reforms of the 1960s and 1970s began the renewed national concern for effective schooling which culminated in a multitude of research studies on effective teaching behaviors. As data were translated into implications for action, various committees published national reports demanding strict accountability efforts on the part of educational leaders in state and local arenas. These mandates have had direct impact on the need to develop teacher evaluation instruments that are reliable, valid, and can discriminate between various levels of teacher performance.

Teacher Evaluation Approaches

The history and background of teacher evaluation on the national level has colored and influenced evaluation approaches of individual states and districts. While national

events influenced approaches to evaluation, in reviewing teacher evaluation literature, the approach any district uses in teacher evaluation is markedly affected by district beliefs concerning teaching processes (example: rationalistic or naturalistic, Stephens, 1976), evaluation purposes, supervisory roles, and evaluation models (Manatt, Palmer & Hidlebaugh, 1976; Redfern, 1980). Most approaches developed in the past eight years can be characterized as viewing evaluation as an activity which "functions to inform decisions about the pursuit of stable, consensual, programatic and institutional goals" (Floden & Weiner, 1978). The approach a district uses in teacher evaluation determines, to a large extent, the type of instrument developed. According to Haefele (1980), a dozen approaches to teacher evaluation (*five of which specifically employ a district-developed instrument) were found to be the most common:

1. Teacher competence is measured by performance of the teacher's classes on standardized tests given at the end of the year. Year-end performance is compared with established norms.

2. Standardized tests are administered to students to determine how much learning is increased over time. The amount of desired gain is established in advance by school personnel, teachers, and an independent evaluator.

3. Students in each grade or subject area are tested at the beginning and end of each semester or school year. Gain scores are computed to contrast class performance with classes of comparable ability. Teacher effectiveness is class performance with classes of comparable ability. Teacher effectiveness is measured by the portion of gainers to losers.

* 4. Informal observations and ratings of the teacher are conducted by the principal and/or other supervisory personnel. Comments by students, parents, and colleagues are incorporated in the final evaluation.

* 5. Systematic observation of the teacher is conducted by the principal and/or supervisor, using a rating form that lists characteristics of good teachers. The teacher's evaluation score is compared to a school or district standard.

* 6. The teacher is systematically observed and rated by peers on the extent to which he or she exhibits important characteristics of good teaching. A predetermined school or district standard is the criterion.

* 7. The teacher's students use a rating form to judge the extent to which the teacher exhibits important characteristics of good teaching. The teacher must meet a predetermined school or district standard of effectiveness.

8. Teachers are required to take the National Teacher Examination and achieve a predetermined standard composite score.

9. Periodically, the teacher is provided with an instructional objective, a sample test item measuring that objective, and information about the content it covers. Students are assigned to that teacher randomly and after instruction, students are tested on the objective. Teacher effectiveness is determined on the basis of how well the students achieved the objective.

10. The Teacher Perceiver Interview is administered to teachers. Teacher effectiveness is based on how well the teacher meets a predetermined criterion or norm-referenced score.

11. The teacher is given written descriptions and/or shown films of typical classroom problems. The teacher's effectiveness is judged on the basis of answer quality following questioning.

*12. The teacher together with the principal and/or curriculum supervisor establishes mutually agreed upon instructional goals and objectives for the year. Observation data and other sources of information are gathered at regular intervals during the year and are used to monitor and evaluate attainment of goals.

While the use of one approach to teacher evaluation does not preclude or exclude the use of other approaches, most districts tend to subscribe to a single approach sometimes with slight adaptations. The variety of evaluation approaches have focused upon a fairly universally accepted set of goals as depicted in a 1977 survey conducted in three hundred sixty-two school districts (ERS Report, 1978). In rank order, the most frequently identified goals of teacher evaluation were: (1) to help staff members improve their teaching performances, (2) to decide on renewed appointments of probationary teachers, (3) to recommend probationary teachers for tenure or continuing contract status, (4) to recommend dismissal of unsatisfactory tenured or continuing contract teachers, (5) to select teachers for promotion to supervisory or administrative positions, (6) to qualify teachers for regular salary increments, (7) to select teachers for special commendation, (8) to select teachers for layoff during reduction-in-force, (9) to qualify teachers for longevity pay increments, and (10) to qualify teachers for merit pay increments.

The survey further confirmed that 97.9% of the districts contacted engaged in formal evaluation of teaching performance. The majority of the individual teacher evaluations were performed by the principal who was then responsible for preparing a summative report. One-third of the surveyed school districts required teachers to evaluate themselves. An analysis of evaluation instruments confirmed

that all surveyed school districts required an evaluation of all certified staff members and, in the course of those evaluations, sought evidence of the quality of classroom teacher performance (ERS, 1978).

In summary, while there are many teacher evaluation approaches available for school districts and they vary considerably from district to district, the purposes of teacher evaluation are far more narrowly defined and performance-based than at any previous time in history. Since the majority of school districts do use instrumentation to record teacher performance, the importance of the development and format of the instrument becomes critical to the success of any evaluation approach in achieving desired outcomes. Furthermore, since the majority of evaluative approaches utilize definitive data and similar goals, the development of valid and reliable instruments to measure these goals is a need of nearly every school district in the United States.

Instrumentation

Research findings related to instrument development and format, both past and present, suggest the need for more precise teacher evaluation instruments. Prior to the 1970s, teacher performance evaluations could best be described as infrequent, subjective, formative observations in which information on general teacher behavior and professionalism

was sought and recorded. Today, performance evaluations are expected to be frequent, objective-based and summative, and provide information on level of performance and effective teaching practices. Teachers - whose effectiveness was once determined on the basis of performing such mundane tasks as heating water and firing the stove - are now expected to utilize precision teaching methods which, hopefully, lead to higher student achievement as assessed by standardized and criterion-referenced testing instruments. Expectations are linked to contemporary goals which undergrid teacher performance evaluation. These underlying goals form the foundation for the current development of teacher evaluation instruments. A brief discussion of recent studies and events provides a background for understanding how district goals, plus contract negotiations, and court decisions have influenced the development of refined, objective teacher evaluation instruments.

Research studies of thirty-two school districts (McLaughlin, 1982) identified four broad goals for teacher evaluation: personnel decisions, staff development, school improvement, and accountability. In addition to the research on evaluation, the public has their opinion. A Gallop Poll (1979) listed "improving teacher quality" as the most frequent response to the question, "What public schools could do to earn an 'A' grade?" (Wise, Darling-Hammond & Pease, 1982).

Even recently negotiated contract agreements reflect the

affirmation of the "new importance" of effective teacher evaluation. Time and again in recent years, teacher evaluation procedures have been tightened in collective bargaining agreements. In one year alone, the percentage of contracts dealing with teacher evaluation increased from forty-five to sixty-five percent (Wise, Darling-Hammond & Pease, 1982).

In due process hearings, courts are increasingly requiring formal dismissal procedures, documentation of teacher performance evaluation, and documentation detailing ways in which a teacher's performance may violate acceptable teaching standards (Beckham, 1981; Peterson & Kauchak, 1982; Strike & Bull, 1981). These rulings necessitated the inclusion of select criteria in school district evaluation policies (Beckham, 1981) such as:

1. A predetermined standard of teacher knowledge, competencies, and skills.
2. An evaluation system capable of detecting and preventing teacher incompetencies.
3. A system for informing teachers of their required standards and according them an opportunity to correct teaching deficiencies.

In light of these requirements, it appears essential that evaluation systems both stipulate predetermined criteria and minimum standards, and produce an evaluation recording instrument that is valid, objective, not overly time-

consuming, and feasible in the organizational context (Knapp, 1982).

Soar, Medley and Coker (1983), in examining currently used methods of teacher evaluation stated, "There is a need for structured analysis underlying any evaluation instrument that tests assumptions about teacher behavior." And, "obtaining a record of teacher behavior in a scoreable form is crucial if we are to be sure we are using identical procedures for evaluating all teachers, thus, minimizing bias, planning for remedial training, and subsequently, measuring changes that occur" (Soar, Medley & Coker, 1983).

Increasingly in recent years, then, the case appears to be made - legalistically and philosophically - for established and documented criteria in developing teacher evaluation instruments which would lead directly to valid ratings by supervisors.

From the 1960s to present day, there appears to have been two performance evaluation goals related to instrumentation:

1. Finding ways to link classroom behaviors of teachers to outcome variables according to specific theories on teacher behavior.
2. Finding ways of recording almost everything of major significance that might occur in the classroom (Walberg, 1974).

As important as these goals appear, the history of teacher evaluation instrument development points out that

there are a few studies which discuss, in-depth, any procedures used in the actual development of the instrument itself - a key component in either trend as stated. However, this gap in research has not prevented the use of a multiplicity of instruments. In 1972, 88.1% of schools surveyed by the Educational Research Service reported the use of some type of instrument to evaluate teaching based on the comparison of teacher performance against prescribed standards. And, in studies of thirty-two districts, McLaughlin (1982) found that instruments used to evaluate teacher performance varied substantively in format - ranging from narratives to checklists to ratings incorporating three, five, or seven point scales.

Most instruments used to record teacher performance do follow one of three types of reporting systems - narratives, checklists and rating scales (ERS, 1978). A brief discussion of each of these reporting systems follows.

Narrative

The narrative is a reporting system in which the evaluator observes general performance, takes notes, records and writes a detailed summary and, subsequently, holds a conference with the teacher regarding performance in the classroom. Essentially, the rater provides a written description of the employee's performance (Henderson, 1984). McGreal (1983) cited the following advantages in using

narrative instruments: the rater can report events that actually occurred; the system can allow for discussion, explanation, and feedback between supervisor and teacher; the procedure is less threatening than a "satisfactory" or "unsatisfactory" rating; the process can provide a holistic view of the performance; and the report can be written descriptively rather than judgmentally (Clements & Evertson, 1981).

Brandt (1973), however, cited several disadvantages in using the narrative instrument format: the recording of data is a time-consuming process for the supervisor; areas for improvement may not be easily targeted; and its use requires skilled and trained supervisors in order to guide the discussion of the written information to desired outcomes.

Brandt (1973) implied that, while narrative instrument usage may facilitate employee-supervisor discussion, the format does not lend itself to identification of specific areas requiring improvement - a key goal in performance evaluation.

Checklists

Checklists are frequently used to record teacher performance. These instruments normally provide a list or series of classroom performance behaviors that could be observed in a typical classroom setting (examples: uses a lesson plan, provides feedback to students, uses audio-visual

materials, etc.). During the observation, items are checked on the list by the supervisor. Based on information taken from the check-list, feedback and reinforcement are furnished in the summative conference.

Brandt (1973) identified advantages in using the checklist as an evaluation instrument: a quick and easy assessment of performance; requires little evaluator training; requires more frequent evaluations; and can be designed to fit specific needs (easily altered). Griffith (1973) suggested it can also provide for some degree of objectivity.

Some disadvantages of checklists, however, include: use of ill-defined and non-specific criteria; qualitative tendencies offering little indication of degree (Medley, 1979); inappropriate identification of the teaching behaviors under observation (providing few feedback opportunities leading to improvement); routinized and mechanical in delivery (Brandt, 1973); judgments may be inferred without careful reflection or analysis (Brandt, 1973); and primarily usable as a formative tool rather than a summative tool. In summary, Brandt suggests that checklists are easy to use but are not designed to give information which leads to specific improvement in job performance.

Rating scales

Rating scales are used more frequently in summative evaluations than any other recording format (McLaughlin,

1982). Traditionally, these scales are specifically designed to facilitate assignment of a number or a written statement to a teacher's performance - depicting the level of quality on a specific criterion or, generally. This number or label typically compares teacher's performance to an established standard.

Formats for rating scales include continuous or numerical ordering, graphic response modes, and descriptive statements (Rummel, 1958). Examples of the most frequently used rating scales include: Behaviorally Anchored Rating Scales (BARS), Behavioral Expectation Scales (BES), Behavioral Observation Scales (BOS), and Performance Distribution Assessments (Jacobs, Kafry, & Zedeck, 1980).

The format of rating scales shows wide variance in the number of points or categories on which to rate. Most scales range from two to nine points or categories. The number of points or categories on a rating scale has historically received much attention, but hundreds of studies on this issue have failed to conclude the optimal number of points to achieve acceptable degree of reliability. On this subject, Aiken (1983), after an exhaustive study, concluded that the number of response categories does make a difference in the "mean and variance of item responses and total scale scores, but efforts to increase the spread of responses by emphasizing greater numbers of response categories (beyond five) will not necessarily improve scale reliability."

Research studies from the 1950s on scale construction concluded that fewer than five points lead to "coarse and loose meeting" while more than five led to less reliability. On the other hand, when three points scales were used, raters tended to select the middle category - avoiding the tendency to give extreme ratings, and gave ratings in the direction of the mean of the group (Rummel, 1958).

Interestingly, the findings from a study conducted by Masters as cited in Aiken (1983) pointed out that the internal consistency of Likert-type rating scales was independent of the number of response categories if opinions about the content being rated was widely divided. If opinions were more harmonious, little rating variance was noted, and, the fewer the number of response categories, the greater the instrument's reliability. Implications for teacher evaluation instrumentation suggest that, if rater agreement exists on the criteria selected for inclusion in the instrument, reliability is enhanced by few rating categories (even as few as three).

The most popular rating scale used in this past decade was the Behaviorally Anchored Rating Scale (BARS) (Landy & Fan, 1983). This format described behavior in short phrases or sentences allowing the rater to make a judgment of performance level based on the description of the behavior being evaluated. Eight comparative studies of the BARS format concluded that inter-rater reliability was enhanced by using the BARS rating format. Landy found the BARS a superior format

when compared to alternative rating scales due to the former's ability to achieve agreement between raters about a ratee's level of job performance (Landy & Fan, 1983). However, Borman's study (1979) on the effects on errors of rater training and instrument format revealed no one format was consistently better or worse than other formats. He stated that, thus far, there has not been a rating instrument judged to be "superior" in minimizing rating errors. He concluded a far-reaching approach to performance evaluation was needed which would include scale construction and utilization as a starting point but also needing to incorporate performance feedback in instrument design. The goal of such efforts should lead directly to the training, selection, placement and promotion of employees (Jacobs, Kafry & Zedeck, 1980).

Advantages of the use of the rating scale have been identified by McGreal (1983). They are: 1) allows for some degree of objectivity; 2) provides for recording degrees of performance; 3) establishes criteria for judgment; 4) provides the rater with specific items to consider during observation; and 5) permits the evaluatee to obtain feedback on the performance criterion. Disadvantages of the rating scale as an instrument format include: the use of inappropriate or non-performance related criteria; ill-defined criteria; vague directions for improvement (McGreal, 1983; Brandt, 1973); lack of opportunity to identify extraneous influences on performance (such as scheduling, types of students);

performance variations based on individual differences; excessive demands on rater knowledge of scale intent (Henderson, 1984); and tendency for the scale instrument to be more useful in defining extremely effective or extremely poor teaching but providing little information between those points (Popham, 1974). Even as precise as a rating scale purports to be, supervisors may still succumb to the "halo effect" or to "central tendency error" (Rice, 1985a). The "halo effect" is a rating error in which the rated individual receives high scores because he/she is well-liked by the supervisor. This phenomenon can occur in reverse as well if the rated individual is not personally well-liked by the supervisor. The "central tendency error" occurs when the evaluating supervisor avoids rating at the extremes of a scale's criteria.

To recapitulate, the disadvantages of using rating scales for evaluative purposes have not reduced their popularity. By far, the popular choice among formats continues to be the traditional numericial or graphic rating scale (Rice, 1985a).

Another point of significance is that the design of the rating scale itself may effect how a rater interprets the information found in the rating scale. Testing this notion, hundreds of studies on rating scale format have generally established that certain rating forms either encourage or discourage certain predictable judgments by the rater. That is, if the form requests little information and assessment

from the rater, the results will provide minimal assistance in improving or reinforcing the performance in question. The rating scale is not designed to formulate judgments, only assist the rater in synthesizing individual judgments (Landy & Fan, 1983). Nonetheless, from his studies, Landy concluded that it was clear a rating scale format must be an integral part of any model purporting to explain performance because, if designed correctly, it assists in the identification of necessary improvements (Landy & Fan, 1983).

One final point relating to instrumentation is that central to the development and use of any instrument format is the need to arrive at a reliable and valid product - one capable of discriminating teacher behavior with enough refinement that judgments can be made about specified evaluative goals. Reliability concerns focus on two main points: the degree to which two or more individuals can observe a third individual at the same point in time and independently draw the same evaluative conclusions, and the degree to which this can be done consistently in varying contexts over time (Mazur, 1980). Similarly, few studies discuss in depth any procedures used in instrument validation (Walberg, 1974). Both reliability and validity have historically been lacking in instrument design (Borich, 1977).

In summary, instrumentation as applied to teacher evaluation, has spawned increasing concern and discussion during a time period when the goals of teacher performance

evaluation have become increasingly narrow and defined. These goals for evaluation outcomes appear to be fairly univeral, due, in part, to the continued research on effective teaching behaviors. In addition to the research on effective teaching, public concern, court proceedings requiring documentation of evaluation conclusions, and teacher bargaining agreements have all contributed to the careful scrutiny of evaluative procedures and instrumentation in the past decade. Reporting systems to record and document evaluation of teaching do vary as the most commonly used formats include narrative, checklists, and rating scales. The most frequently used format, however, is the rating scale. The advent of the Behaviorally Anchored System (BARS) thrust behavior-based rating formats into wide acceptance by many types of organizations in recent years - including education. This graphic representation of performance rating is the most used evaluation format of today. But due to the scarcity of research on instrument format alone, there is a need to develop more reliable instruments, improve instrument content, and develop instruments which meet the challenge of being able to validate ratings of teacher performance and to identify areas for growth and reinforcement (Henderson, 1984).

Criteria Selection

A discussion of criteria selection is pertinent to instrument development since this is one of the most time-consuming activities for any district or established committee seeking to review, refine, or develop a valid teacher performance evaluation system.

As efforts to assess existing evaluation procedures or develop new ones are under way, it soon becomes apparent that research on criteria or that which constitutes "good teaching" is readily available. But, it is also evident that research studies and educational theorists fail to agree on whether or not effective teaching behaviors can be identified and generalized across all teachers and systems. Centra and Potter (1980) observed, "student achievement is affected by a considerable number of variables, of which teacher behavior is but one." Additionally, teacher and pupil performance may also be affected by factors such as school size, programmatic issues, resources, administration and incentives (Joyce & Weil, 1972; McKenna, 1981), and specific contexts and situations requiring teachers to dispense a wide repertoire of diagnostic, instructional, managerial and therapeutic skills (Brophy and Evertson, 1976).

Studies of effective teaching criteria have - as often as not - failed to aid evaluators by presenting a range of conflicting conclusions. For example, Popham reported that

criteria are often ill-defined and vary from rater to rater, thus, creating inconsistency and confusion (ERS, 1978). Medley (1979) added that research of the early 1970s indicated little if any relationship between ratings of teacher effectiveness and measures of pupil gains. He implied that instrument designs themselves have failed to discriminate among teacher behaviors related to effective teaching practices.

Direct instruction, time spent on learning, goal setting, feedback, classroom climate and management, teacher authoritarianism (Rosenshine, 1979; Berliner, 1977; Bruner, 1976; Stallings, 1977; Hunter & Russell, 1977) have surfaced as central focuses in performance evaluation, but the degree of importance of each factor, even in controlled settings and studies, is unverified. The result of inconclusive research findings has caused local school districts to spend an inordinate amount of time on evaluation system development and criteria selection.

As districts examined the best available research data and practices, commonalities in procedures for arriving at criteria selection have emerged from a wide-range of school districts (McLaughlin, 1982). In a study of thirty-two districts in twenty-four states, McLaughlin found that with few exceptions, teacher evaluation systems resulted from well-organized committees of teachers, administrators, union representatives, principals and, occasionally, parents.

Development of process and instrument took from six to twelve months. While some districts relied on outside consultants such as Manatt, Redfern, and Hunter, most developed systems without outside assistance.

The results obtained in the study of instrument development revealed surprising consistency in the categories of criteria contained in the instruments (McLaughlin, 1982) - regardless of whether or not instruments were developed by internal or external means. Five broad criteria categories emerged:

- Teaching procedures
- Classroom management
- Knowledge of subject matter
- Personal characteristics
- Professional responsibility

In a subsequent examination of fifty evaluation forms from thirty states, the presence of the aforementioned categories of criteria was confirmed.

Equally important to criteria selection in instrument design is the matter of criteria reliability. Several research studies concentrated on the reliability of variables (criteria) found in teacher evaluation instruments. There appeared to be a distinction between what Borich (1977) called "high and low inference variables" within each of the main criteria categories of the instruments examined. Inference means the "amount of judgment the observer must apply to

determine the presence, absence, or quality of a phenomenon" (Borich, 1977). High inference criteria or variables such as "warmth" or "enthusiasm" lack reliability since rater judgment is based on personal perceptions of those abstract concepts. Low inference variables or criteria such as "presence of specified lesson objectives" or "allows for student participation" fall into an acceptable range of reliability because they are observable. Thus, the conclusion Borich reached was that criteria selection, if it is to be reliable, should focus heavily on low inference variables if the district stresses performance-based objectives in evaluation. Additionally, sub-variables with definitions (indicators) and examples must accompany general variables (criteria) to achieve adequate levels of reliability because those sub-variables further define what is meant by the selected criterion (Donovan & Kathryn Peterson, 1984). It has also been found that the more specific, defined, observable and objective-based are criteria, the more likely that the rating of criteria will be reliable across numerous raters (Franklin & Thrasher, 1976).

In addition to identification of major criteria categories and reliability of variables with these categories, the selection of valid criteria has been discussed and researched. Criteria selection becomes meaningless if instrument format does not produce ratings which illustrate qualitative performance differences between teachers in the

performance of those criteria (Menne, 1972). In order to differentiate performance, valid criteria need to be identified. Torgerson, as early as 1935, stated that the establishment of valid criteria is one of the most difficult tasks in validating a research instrument and is one of the most essential (Good & Barr, 1935).

It has also been found that instrument validation can be enhanced if the criteria selected included the following characteristics:

1. A definition of each performance criterion with explanatory behavior incidents written for each criterion.
2. A consensual agreement on the placement of the criterion into categories.
3. The inclusion of clarification statements succeeding the criteria.

These clarification statements referred to in the third characteristic assist the rater in determining the degree of presence or absence of the behavior associated with each criterion (Jacobs, Kafry, & Zedeck, 1980).

In summary, the issues surrounding criteria and criteria selection may complicate the instrument development process as districts look for research conclusions to assist in defining good teaching. Even though hundreds of studies have been conducted to identify effective teaching behaviors, experts still fail to agree on any one set of criteria. However,

major categories of criteria do appear to be universally accepted by districts developing evaluation instruments. The selection of criteria for teacher evaluation is still a difficult and time-consuming task, but one that greatly influences instrument design. If expert opinion is taken into account, then criteria need to be specific, detailed, and backed by descriptive statements further explaining behaviors associated with the specified criteria.

Formative and Summative Evaluation

As instruments are developed, evaluators need to be cognizant of whether or not the instruments are designed for formative or summative evaluation purposes. Formative evaluation is an on-going, descriptive, developmental, and non-judgmental evaluative process. The intended purpose of formative evaluation is to improve the performance of one person based on the process of instruction. It occurs from the bottom up in the supervisory hierarchy with a team approach oriented toward serving the individual (Manatt, 1981). Brock views the mission of formative evaluation as improving "subsequent educational practices allowing the image of a cycle of educational practice to take shape" (Millman, 1981). Scriven believes formative evaluation is used for faculty development with feedback from the evaluation process going directly to the teacher or to a designated consultant

(Millman, 1981). Basically, formative evaluation is designed to improve performance by aiding employees through the identification of areas where improvement is desired. Persons other than a direct supervisor (such as students, peers or parents) may take part in formative evaluation, aiding in identifying a teacher's strengths and weaknesses which will lead to improved performance (Howsam, 1973).

In contrast, summative evaluation is the final, judgmental, and comparative evaluation founded on an organized body of previous knowledge and collected information. It relates to improvement in the school organization and involves both products and processes of instruction. According to those subscribing to summative evaluation philosophy, excellence is achieved by individuals only if supervised by others with focus from the top down in the supervisory hierarchy, serving all stake holders for mutual benefit (Manatt, 1981). As Brock points out, summative evaluation can be used to make personnel decisions (Millman, 1981). In addition, summative evaluation primarily exists because: "1) human careers are at stake, not single 'mere' improvement; 2) if it is not possible to tell when teaching is bad or good overall, it is not possible to tell when it has improved; and 3) if it is possible to tell when it is bad or good, personnel decisions can be made even though it is not known how to make improvements. In short, diagnosis is sometimes easier than healing, and an essential preliminary to it" (Scriven as cited

in Millman, 1981). Responsibility for summative evaluation is normally assigned to the building principal.

Employing current research, both formative and summative evaluation are integral parts of a continuing cycle to make teacher performance evaluation valid, reliable, and legally discriminating. It is essential for local districts to have developed the primary purpose or plan of the evaluation program prior to actually engaging in the process of evaluation with staff members.

Both formative and summative evaluation procedures substantially influence instrument design since each demands different outcomes. This study utilized summative evaluation procedures and the instruments selected for use in this study reflect the purpose of summative evaluation.

Further Research Topics

Instrument design and format are connected to an abundance of broader topics and are linked through discussion, research and dialogue to teacher and performance evaluation. Since the mid-1970s, volumes of research, propositions, recommendations, and opinions have been published and publicized on these topics. Teacher evaluation is the backbone of the Effective Schools Research (Edmonds, 1978), teacher competency testing, administrator evaluation and training, college and university reform involving teacher

education programs, individual State Task Force Reports (twenty-seven states to date have established State Task Force Committees), and legislation at both State and Federal levels. Merit pay, master teachers, career ladders and incentives - all issues tied to teacher performance evaluation - have produced a wealth of material for review. School improvement models, curriculum revision, and long-term staff development goals are also connected with and affect teacher performance evaluation and instrumentation.

Summary

Instrument format is not a highly researched or publicized topic when compared to the substantial amount of literature published on the broader topic of teacher performance evaluation. Factors which directly influence a school district's development of performance evaluation instruments are as follows:

1. The history and background of teacher evaluation from the early 1900s to the present.
2. Approaches to teacher evaluation which determine content and context of instrument development.
3. Instrumentation, both past and present, reflecting the influence of educational reforms, public demands, legal considerations, and commonly used reporting formats.

4. Criteria selection based on research and the subsequent implication for instrument design.
5. Type of evaluation (formative or summative) which determines the basic intent and purpose for teacher evaluation procedures and processes.

Since the majority of states require teacher performance evaluation, the need to examine instrument format seems critical to determine how format affects the ability of evaluators to make valid ratings of teacher performance and to identify areas for growth and reinforcement. If evaluative instruments cannot assist in discriminating between effective and ineffective teaching behaviors, then the evaluation process will become frustrating, ineffective, and unproductive for both the teacher and supervisor.

If for no other reason, the time investment in evaluation procedures should produce results which both inform and create changes in teacher behavior. Teachers themselves want qualitative feedback and, in a study of thirty-two school districts (McLaughlin, 1982), it was found that teachers report an increased sense of professionalism and motivation to improve classroom practices as a result of an evaluation program geared toward recognition of competence. However, other conclusions concerning teacher attitudes toward evaluation point to a need for development of more systematized, consistent and fair evaluation procedures and instruments. As an example, less than half of the districts

in McLaughlin's study reported full support by teachers of the evaluation program - due primarily, to lack of uniformity and consistency across the district in the use of the instrument itself (McLaughlin, 1982). Considering past and present research findings, it becomes even more apparent that evaluative instruments themselves can and do play a substantive role in the efficacy of the total evaluation program and process.

CHAPTER III. METHODS AND PROCEDURES

This study was designed to examine the efficacy of two teacher evaluation instrument formats to determine 1) the effect of format in influencing validated ratings of a teaching performance on a specified criterion, 2) the effect of instrument format on the agreement of performance ratings by evaluators (inter-rater, 3) the effect of format in assisting evaluators in the identification areas (teaching behaviors) to improve and those to reinforce in a given teaching segment using a specified criterion, and 4) the effect of using a continuous rating scale versus a point rating scale by evaluators rating performance on indicators and/or the criterion.

This chapter describes the methods and procedures used to collect and analyze the data to complete this study. The first section of this chapter is "Collection of Data" and includes several subsections: materials, sample, expert panel, and procedures. The second section, "Analysis of Data" describes how the data were analyzed.

Collection of Data

Materials

Two instrument formats were examined in this study: 1) Graphic Response Mode/Indicator (GRM Indicator), and 2) Double

Scale Response Mode/Forced Indicator Rating (DSRM/Forced Indicator Rating) using a point rating scale. A variation of the DSRM/Forced Indicator Rating format, the Double Scale Response Mode/Forced Indicator Rating using a continuous scale, was also examined but due to difficulties in scale design, data collected from this instrument could not be statistically analyzed. Descriptive data for this scale are presented, however. Problems encountered with this scale will be discussed later in this chapter and in Chapter IV. All instruments were specifically designed for this study and required an examination of related literature and the help of some members of the Department of Professional Studies at Iowa State University. Criteria and indicators (statements describing effective performance associated with a criterion), were obtained from evaluation instruments currently used in Iowa, Georgia, Florida, and North Carolina and from a list of validated criteria identified in conjunction with the research of the School Improvement Model, a project located at Iowa State University. Individuals who participated in this study received one of two packets of materials - one contained the GRM/Indicator format and explanatory information: the other contained both DSRM/Forced Indicator Rating formats and explanatory information. Both packets contained the Improvement and Strength Areas Reporting Form which was used by the evaluators after making performance ratings to identify the teacher performance areas to improve and those to

reinforce. A registration card on which the evaluators could supply demographic information was also provided in both packets.

The GRM/Indicator and DSRM/Forced Indicator Rating instrument formats were used to rate the same performance criterion, "Communicates Effectively with Students", and used the same rating categories - "Must Improve", "Needs Improvement", "Meets Standards", and "Exemplary." The explanation of the levels of performance indicated by each of the rating categories can be seen in Appendix D.

The Graphic Response Mode/Indicator (GRM/Indicator) instrument format was an adaptation of several summative instruments collected for research in conjunction with the School Improvement Model project. Forty-nine participants used this instrument to rate teaching performance on the specified performance criterion. The criterion and the four rating categories were stated on the instrument. Each rating category had one or two sentences explaining/describing teaching behavior on the criterion at that level. Appendix A shows the GRM/Indicator instrument format. Raters were to place a check mark on the line segment over the description which best represented their observation of the level of performance on the criterion. Eight indicators, performance descriptors, were provided on a separate sheet preceding the GRM/Indicator instrument. These were provided to assist evaluators in rating performance on the specified criterion.

The indicators were not to be rated, but used as a guide. The indicator sheet can be seen in Appendix E. The title, Graphic Response Mode/Indicator format was selected because the rating statements described levels of performance and indicators were provided to guide the rater.

The Double Scale Response Mode/Forced Indicator Rating using a four point point scale was included in the packet of materials used by fifty-six participants. It was developed by the researcher and Dr. James Sweeney. This instrument format required evaluators to rate performance on the specified criterion only after they had rated eight indicators which described effective teaching performance behaviors related to the criterion (hence, double scale). The eight performance indicators are described below.

1. Clarity of Directions (regarding assignments, procedures, or homework).
2. Logical Concepts (referring to the logical, sequential order of the teaching lesson).
3. Questioning Techniques (referring to eliciting and prompting of student responses).
4. Feedback (referring to meaningful information concerning correctness of student responses).
5. Speech Rate (rate or speed of the teacher's delivery).

6. Delivery Skills (referring to tone, pitch, and word patterns used by the teacher).
7. Body movement (referring to how the teacher positioned himself in relation to students, and to facial expression and gestures).
8. Vocabulary Level (referring to choice or selection of words used to present content).

These indicators were the same as those provided in the GRM/Indicator instrument format packet.

The Double Scale Response Mode/Forced Indicator Rating instrument using the four point scale was designed so that evaluators had to first rate the eight performance indicators (descriptors of effective performance on the criterion) by placing a check mark on the line segment under the rating category (point) best representing the performance on each indicator, and then, on the criterion. It was posited that the forced rating of the imbedded indicators would lead to more valid performance ratings. This instrument format can be seen in Appendix B.

The Double Scale Response Mode/Forced Indicator Rating format using the continuous scale was included in the same fifty-six participants' packets as was the DSRM/Forced Indicator Rating instrument which used the four point scale. It had the same design and purpose as that of the DSRM/Forced Indicator Rating format discussed previously but allowed evaluators to record the rating of teaching performance on

each of the indicators and the criterion at any point along a continuous line. Thus, it was a four point Likert-scale (as used in the parallel format) but the evaluator could rate performance in any one of the four category points or at any point between the two rating limits for each category on the scale. This format was designed to ascertain if evaluators would opt to use the continuous scale and, if so, in what direction the ratings would be drawn.[1] This instrument format can be seen in Appendix C.

To summarize, there were major differences in the usage of two types of instrument formats provided in this study. Evaluators who used the DSRM/Forced Indicator Rating formats were asked to rate performance on eight embedded indicators (included on the same page as the specified criterion) before they made an overall teaching performance rating on the criterion. Those who used the GRM/Indicator format had access to the indicators on a separate sheet but were not required to rate them.

---

[1] The format of the DSRM/Forced Indicator Rating instrument (point scale) was piloted in the Mason City, Iowa, Community School District in 1984-85 and was adopted as the summative teacher evaluation instrument for 1985-86. This decision was based on extensive committee study, administrator and teacher input, and the conclusions from the piloting which made this instrument format gain wide acceptance in the district due to its design, specificity, comprehensive nature and ease of analysis (Rice, 1985b).

The Improvement and Strength Areas Reporting Form was included in all packets of materials. It was used by evaluators to record three teaching behaviors related to the criterion "Communicates Effectively with Students" to reinforce or strengthen, and two communication areas to target for growth or improvement (in priority order). This reporting form may be seen in Appendix F.

A cover sheet for both packets, Information/Directions, was included to provide specific information about the use of packet materials. This sheet may be seen in Appendix H. A final document, a Registration Card, was also included. It was used to obtain demographic information from participants. This is shown in Appendix G.

In summary

In summary, two packets of materials were used for this study. Each packet included:

- An information/direction page
- A rating category explanation page
- A separate sheet listing indicators of the criterion included in the packet containing the GRM instrument
- Either the GRM instrument or both forms of the DSRM format
- The Improvement and Strength Areas Reporting Form
- A Registration Card

A twenty minute videotape of an eighth grade social studies lesson was also used in this study. The tape was selected for the following reasons: teacher performance on the specified criterion could be assessed and was at a level at which variance in evaluator rating may occur; the lesson allowed enough time for evaluators to rate performance on the criterion and indicators; visual and auditory quality of the tape were high, and the eighth grade classroom provided a "middle ground" for elementary and secondary supervisors.

All materials used in this study were field tested with seventy-five administrators in Chesterfield County, Virginia, on January 8, 1985. As a result of the field test, the information/direction sheet was altered to better define the intent and purpose of the study. Since a number of the participants (about 1/5) did not rate performance on the specified criterion even though they did rate performance on the indicators when using the DSRM formats, the directions provided at the beginning of each instrument format were revised to be more explicit highlighting the importance of rating the criterion. These changes were helpful since only three of the sample participants did not rate the criterion.

Sample

The sample consisted of one hundred five administrators who attended a three-day multi-district teacher performance evaluation workshop sponsored by the Butler County Educational

Service Region in Canton, Illinois, from June 10th through June 12th, 1985. Dr. Richard Manatt conducted the workshop sessions. These data were collected on the final day of the workshop after the following had occurred: training of participants in data gathering and background, discussion of formats of evaluation, viewing of videotaped teaching sessions, explanation of formative and summative evaluation purposes, discussion of prior observations and data collected on the teacher to be observed on videotape, and guided practice in summative evaluation.

The participants were, on the average, male elementary principals with five or more years of experience in supervision and responsible for supervising twenty to forty teachers. Most of the participants had attended at least one workshop on teacher evaluation of their own volition prior to this training. Tables 1, 2, and 3 present a more complete description of the sample. Little variation in experience, background and training between groups surfaced.

The packets of materials necessary to conduct this study were randomly distributed to all participants - so that each received one or the other packet as they entered for the workshop that day. A total of forty-nine participants received the packet containing the GRM/Indicator instrument: fifty-six participants received the packet containing the DSRM/Forced Indicator Rating instruments.

Table 1. Number and percent of participants by job title and job level

Job Title	GRM Evaluator Group (N=49)	%	DSRM Evaluator Group (N=56)	%	Total Number (N=105)
Superintendent	6	12.2	11	19.6	17
Assistant Superintendent	3	6.1	3	5.3	6
Principal	35	71.4	32	57.1	67
Assistant Principal	2	4.0	3	5.3	5
Supervisor	3	6.1	1	1.7	4
Department Head	1	2.0	0	0.0	1
Teacher	4	8.1	4	7.1	8
Other	0	0.0	3	5.3	3
Job Level					
Elementary	20	40.8	24	42.8	44
Middle School	8	16.3	10	17.8	18
Senior High	4	8.1	4	7.1	7
High School	10	20.4	7	12.5	17
Other	5	10.2	8	14.2	13

Table 2. Number and percent of participants by district size and years in supervision

District Size	GRM Evaluator Group (N=49)	%	DSRM Evaluator Group (N=56)	%	Total Number (N=105)
0 - 1000	20	40.8	23	41.0	43
1000 - 2000	13	26.5	11	19.6	24
2000 - 3000	2	4.1	4	7.1	6
3000 - 4000	11	22.4	13	23.2	24
4000 - 5000	1	2.0	2	3.5	3
5000 - 6000	1	2.0	0	0.0	1
6000 - 7000	0	0.0	0	0.0	0
Over 8000	0	0.0	0	0.0	0
Missing	1	2.0	3	5.3	4
Years Supervising					
0 - 1	7	14.2	4	7.1	11
2 - 3	4	8.2	6	10.7	10
4 - 5	10	20.4	9	16.7	19
6 - 7	5	10.2	3	5.3	8
8 - 9	2	4.0	5	8.9	7
10 - 11	4	8.1	5	8.9	9
12 - 13	5	10.2	3	5.3	8
14 - 15	8	16.3	5	8.9	13
Over 15	0	0.0	12	21.4	12
Missing	4	8.1	4	7.1	8

Table 3. Number and percent of participants by sex, number of teachers supervised, and previous experience

Sex	GRM Evaluator Group (N=49)	%	DSRM Evaluator Group (N=56)	%	Total Number (N=105)
Male	39	79.5	44	78.5	83
Female	5	10.2	5	8.9	10
Missing	5	10.2	7	12.5	12
Number of Teachers Responsible for Supervising					
0	3	6.1	6	10.7	9
1 - 10	4	8.0	5	8.9	9
11 - 20	19	38.7	12	21.4	31
21 - 30	13	26.5	15	26.7	28
31 - 40	5	10.2	8	14.2	13
41 - 50	2	4.1	3	5.3	5
51 - 60	0	0.0	0	0.0	0
61 - 70	0	0.0	1	1.7	1
Over 70	0	0.0	1	1.7	1
Missing	3	6.1	0	0.0	3
Previous Experience					
Workshop on own	29	59.1	36	64.2	65
Workshop required	7	14.3	9	16.1	16
District inservice	13	26.5	18	32.1	31
Coursework	21	42.8	27	48.2	48
Mentor	9	18.4	10	17.8	19
Other	1	2.0	6	10.7	7

Expert panel

The expert panel for this study included Dr. Richard Manatt, Dr. Shirley Stow, and Dr. James Sweeney, all Iowa State University professors in the Department of Professional Studies. They were selected because of their expertise in teacher evaluation and effective teaching.

The panel members met on September 16, 1985 at Iowa State University. Each had observed (more than fifty times) the videotaped lesson to be used in the evaluation. The packets of materials were provided to them and they were to attempt to reach consensus and:

1. Rate the teaching performance on the criterion specified on the GRM/Indicator instrument.
2. Rate the teaching performance on each of the eight indicators found on the DSRM/Forced Indicator Rating format (point scale).
3. Rate the teaching performance on the criterion specified on the DSRM/Forced Indicator Rating format (point scale) after rating the indicators.
4. Identify two performance areas (teaching behaviors) to target for improvement - in priority order and using the communication criterion.

5. Identify three performance areas (teaching behaviors) to reinforce using the communication criterion.
6. Use the DSRM/Forced Indicator Rating instrument format (continuous scale) to rate performance on indicators and criterion.

The expert panel was able to reach consensus in all areas. Their ratings and identification of teacher behaviors to improve and reinforce became the standard against which the data collected from the sample were compared.

Procedures

Participants had received two days of training in the following areas prior to obtaining the packets of material needed to complete this study: data gathering and background, formats of evaluation, videotaped lesson to observe and rate, information on formative and summative evaluation purposes, discussion of prior observations and data collected on the teacher to be observed on videotape, and guided practice in summative evaluation.

The two packets of material used in this study were randomly distributed to participants as they entered the final day of the workshop session. Participants first read the information/direction sheet, then the videotaped lesson was shown to all participants at the same time. Participants were asked to assume that they would only make one observation and

were to make a summative rating of the teaching performance using the instruments found in their packets. Following the rating of performance, participants identified the performance areas (specific teaching behaviors) to target for improvement and those to reinforce. Finally, the registration card was completed and all packets were returned to Dr. Manatt.

The data obtained from the completed packets were coded and key punched at the Iowa State University Computer Science Center. Results were analyzed using SAS (Statistical Analysis Systems) techniques. To summarize, the following procedures occurred:

1. Training and practice were provided to evaluators prior to collecting the data.
2. All participants were asked to assume that they would only make one observation and were to make a summative rating based on that observation and other information provided prior to the viewing of the videotape.
3. Packet materials were randomly distributed to participants.
4. A videotaped teaching lesson was observed by all participants.
5. The task was clarified and participants then rated the teaching performance using the required format.

6. All participants identified two teacher behaviors to target for improvement and three areas to reinforce.
7. Data were collected.
8. Data were coded, tabulated, and analyzed.

The time-frame for this study was:

January, 1985	-	Field test
May, 1985	-	Sample selected
June, 1985	-	Data collected
July, 1985	-	Data coded, tabulated
October, 1985	-	Analysis completed

Analysis of Data

Data was collected, coded, and prepared for transfer to key punch cards for computer analysis at the Iowa State University Computer Center using Statistical Analysis Systems. Descriptive statistics (frequencies, means, standard deviations, chi-square) were used to analyze the criterion ratings and areas of performance to improve and reinforce. These data from both groups were compared with expert panel data to determine significant statistical differences. The expert panel rating was considered to be the "correct rating." Descriptive statistics were also used to analyze differences in indicator ratings and to analyze the effect of the

continuous scale of the DSRM/Forced Indicator Rating format.

The first hypothesis was tested using analysis of variance, ANOVA. The evaluator's ratings of performance on the criterion using the GRM instrument and the DSRM instruments were compared to the panel's ratings.

Hypothesis two was analyzed using analysis of variance and the Levene Test of Equality of Variance to test variance in evaluation ratings, by format used, when compared to the panel's ratings.

Hypotheses three and four were analyzed using chi-square and the calculation of Z to determine significant differences of paired variables. The expert panel's identification of targets for growth and reinforcement were considered to be correct. Data collected from participant responses, by format used, were aggregated, combining the first and second identified performance areas to improve and those to reinforce.

In preparation for data analysis, it was found that statistical testing involving the DSRM/Forced Indicator Rating format using the continuous scale could not be adequately addressed. Problems with scale design prohibited statistical tests of differences between this instrument and a point scale. Numerical ratings on the DSRM/Forced Indicator Rating instrument using a continuous scale could not be translated into scores for comparison without conducting a separate psychometric analysis allowing for such testing. The breadth

of such a conversion would lead to a separate voluminous study. Discussion of the descriptive data which sheds light on the use of the continuous scale will be included in Chapter IV and in the Recommendations section of Chapter V.

CHAPTER IV. FINDINGS

The purpose of this chapter is to report the results of the investigation examining the relationship between instrument format and 1) evaluator ability to make valid teacher performance ratings, 2) the reliability of ratings among evaluators using one or the other format (inter-rater reliability), 3) the ability of evaluators to identify performance areas related to the criterion "Communicates Effectively with Students" to target for improvement and reinforcement, and 4) the effect of a continuous scale on evaluator ratings. Following training and the viewing of a videotaped teaching lesson, evaluators used one of two instrument formats to record their ratings - the Graphic Response Mode Indicator Rating or the Double Scale Response Mode/Forced Indicator Rating using a point scale. A third instrument, the Double Scale Response Mode/Forced Indicator Rating format using a continuous scale was also used but data collected from this instrument were not statistically analyzed due to difficulties with the scale. Problems with this analysis are discussed later in this chapter. Suggestions for testing the effects of this format in future research are discussed in Chapter V. The Improvement and Strength Areas Report Form was used by the evaluators to record behaviors in the teacher's performance to improve those to reinforce.

This chapter is divided into two sections: 1) Descriptive

Data, which includes frequencies, means, and standard deviations of the data collected, and 2) Inferential Statistics, which reports the data analysis using analysis of variance, chi-square and standard normal (Z) approximation to the binominal test (the same as the chi-square test with one degree of freedom).

Four issues of concern in this study were to determine if instrument format affected 1) the evaluators' ability to make teaching performance ratings using a specified criterion, 2) the agreement of performance ratings by evaluators using a specified criterion (inter-rater reliability), 3) the ability of evaluators to identify teacher performance areas to improve or to reinforce relating to the specified criterion, and 4) the ratings by evaluators using a continuous scale as opposed to a point scale.

A total of one hundred five administrators participated in this study. Forty-nine of the participants used the Graphic Response Mode/Indicator instrument format to record ratings and fifty-six participants used two forms of Double Scale Response Mode/Forced Indicator Rating instrument format. Eight indicators which characterized effective performance on the criterion were provided to all evaluators. Evaluators who used the DSRM/Forced Indicator Rating formats were required to rate performance on these indicators - imbedded in the instrument format. All evaluators were asked to identify teacher performance areas to target for improvement and those

to target for reinforcement which related to the specified criterion.

Descriptive Data

Evaluators were asked to rate the teacher's performance on the criterion "Communicates Effectively with Students." Table 4 presents the frequency, mean, mode, and standard deviation for all evaluators' ratings. The expert panel rating is also provided.

Table 4. Ratings of all evaluators on the criterion "Communicates Effectively with Students"

Rating Category	Value	Frequency	Percent
Must Improve	1	40	38.1
Needs Improvement*	2	60	57.1
Meets Standards	3	2	1.9
Exemplary	4	0	0

Mean	Mode	Standard Deviation
1.627	2.0	.525

N=102 cases, 3 missing
* Expert panel rating

Of the 102 evaluators who rated performance, 60 rated the teacher's performance on the criterion as "needs improvement" - the same as did the expert panel - and 40 rated it "must improve." Only two evaluators reported that the teacher met district standards on the criterion. Fifty-seven percent of

the evaluators agreed with the expert panel's rating - needs improvement.

Table 5 presents the evaluators' ratings of performance on the criterion when ratings were separated by the instrument format they used.

Table 5. Ratings of the overall criterion, "Communicates Effectively with Students", for GRM/Indicator format and DSRM/Forced Indicator Rating format

Format	Frequency Rating	Number of Participants	Mean	Standard Deviation
GRM/ Indicator	1-Must Improve	30	1.388	0.492
	2-Needs Improvement	19		
	3-Meets Standards	0		
	4-Exemplary	0		
DSRM/ Forced Rating	1-Must Improve	10	1.849	0.456
	2-Needs Improvement	41		
	3-Meets Standards			
	4-Exemplary			
Expert Panel	2-Needs Improvement		2.000	0.000

N=102 cases, 3 missing

Of the 49 participants who used the GRM/Indicator instrument format, 19 rated the performance the same as the expert panel - "needs improvement" while 30 rated performance "must improve" resulting in a mean rating of 1.388. Of the 56 evaluators who used the DSRM/Forced Indicator Rating format, 41 rated performance as "needs improvement" resulting in a mean rating of 1.849. Only 12 evaluators who used the DSRM

format rated the performance different from the expert panel while 30 evaluators who used the GRM format rated the performance different from the panel. More of the DSRM evaluators' ratings were closer to the expert panel rating than were those using the GRM format. The standard deviation (.456 for the DSRM ratings compared to .492 for the GRM ratings) shows less variance in ratings by the DSRM evaluators than in ratings by the GRM evaluators.

All evaluators were asked to identify and rank two performance areas relating to the criterion of communication to target for improvement - areas judged to be important to improve for growth in teaching performance to occur. These data were taken from the Improvement and Strength Areas Reporting form (found in all participant's packets of materials) and are reported by format used. The first and second choices of both groups were aggregated and compared to the expert panel choices. Table 6 shows the results.

The expert panel identified and ranked two teaching behaviors for improvement in Communication with Students: 1) Questioning Techniques and, 2) Feedback. Evaluators who used the GRM/Indicator Instrument format identified Vocabulary Level and Delivery Skills as the two most important areas (behaviors) to target for improvement. Their third choice was Questioning Techniques, the first choice of the expert panel. Feedback also identified by the expert panel, ranked fifth. Evaluators who used the DSRM/Forced Indicator Rating

TABLE 6. Summary of communication areas identified to improve by the expert panel and by evaluators using the two instrument formats

Expert Panel Priority Areas to Improve	GRM (N = 48) No. Identifying Same Area as Expert Panel
Questioning Techniques	14
Feedback	4

Rank Order of All Identified Areas Targeted for Improvement by GRM Evaluators	Number of Evaluators (N = 48) - first or second choices
1 Vocabulary Level	27
2 Delivery Skills	19
3 Questioning Techniques	14
4 Logical Concepts	7
5 Feedback	4
5 Speech Rate	4
5 Encourages Student Participation	4
5 Difficulty of Material	4
6 Clarity of Directions	3
6 Body Movement	3
7 Other	2

Rank (using all areas identified to improve)	DSRM (N = 54) No. Identifying Same Area as Expert Panel	Rank (using all areas identified to improve)
3rd	15	2nd
5th	5	6th

Rank Order of All Identified Areas Targeted for Improvement by DSRM Evaluators	Number of Evaluators (N = 48) - first and second choices
1 Vocabulary Level	20
2 Delivery Skills	15
2 Questioning Techniques	15
3 Encourages Student Participation	12
4 Logical Concepts	8
5 Varies Teaching Methods	7
6 Speech Rate	5
6 Enthusiasm	5
6 Feedback	
7 Clarity of Directions	4
8 Body Movement	3
8 Effective Teaching Strategies	3
9 Difficulty of Material	1
9 Use of Objectives/Other	1

instrument format identified Vocabulary Level as the most important behavior to improve upon. Delivery Skills and Questioning Techniques (the expert panel's first choice) were identified equally as a second choice while Feedback, a second choice by the expert panel, ranked sixth. There was little difference between the two groups in the areas (behaviors) identified for improvement in the teachers' performance; both were relatively close in ranking of the panel's selections of Questioning Techniques and Feedback. Table 6 also shows the other areas targeted for improvement by both groups of evaluators - in rank order by frequency of selection.

All evaluators were also asked to identify for reinforcement, three areas related to the communication criterion. Reinforcement areas represent the teaching behaviors which should be continued or expanded upon to maintain or surpass acceptable standards. The data related to performance areas to reinforce are reported in Table 7 which shows the top three choices of the expert panel compared to evaluators who used the GRM and those who used the DSRM format. The expert panel identified 1) Logical Concepts, 2) Speech Rate, and 3) Appearance as the teaching behaviors to reinforce. Only one of those areas, logical concepts, was also identified by a substantial number of evaluators using the GRM/Indicator format (21 of 48) and those using the DSRM/Forced Indicator Rating format (13 of 54). Twenty-one evaluators who used the GRM/Indicator format identified

TABLE 7. Summary of the communication areas identified to strengthen or reinforce by evaluators using the GRM, DSRM, and by expert panel (first, second, and third choices combined)

Expert Panel Areas to Strengthen/ Reinforce	GRM (N=48) No. Identifying Same Area As Expert Panel	GRM (N=54) No. Identifying Same Area As Expert Panel
Logical Concepts	21	13
Speech Rate	3	8
Appearance	2	2

GRM - ranking of identified areas to strengthen/ reinforce by frequency of selection		DSRM - ranking of identified areas to strengthen/ reinforce by frequency of selection	
Logical Concepts	21	Delivery Skills	21
Questioning Techniques	19	Use of Chalkboard	14
Delivery Skills	15	Logical Concepts	13
Feedback (to students)	11	Body Movement	13
Clarity of Directions	10	Knowledge of Content	10
Body Movement	9	Speech Rate	8
Knowledge of Content	7	Questioning Techniques	8
Vocabulary Level	5	Clarity of Directions	6
Use of Chalkboard	4	Vocabulary Level	5
Speech Rate	3	Appearance	2
Appearance	2	Reviews	2
Encourages Student Participation	1	Use of Objectives	2
		Feedback to Students	1

Logical Concepts, 19 selected Questioning Techniques, and 15 chose Delivery Skills as the top three performance areas to reinforce. Of those who used the DSRM/Forced Indicator Rating format, 21 identified Delivery Skills, 14 selected Use of Blackboard, and 13 chose Logical Concepts and Body Movement equally as the third choice. Only three evaluators who used the GRM/Indicator format selected the expert panel choice, Speech Rate, while eight of the evaluators who used the DSRM/Forced Indicator Rating format. Appearance, the third expert panel area to reinforce, was identified by only two evaluators from both groups. Table 7 also shows the rank order by frequency of all areas chosen for reinforcement by both groups.

Participants who used the DSRM/Forced Indicator Rating instrument format were required to rate performance on eight indicators, descriptors of performance on the criterion, before making the rating of performance on the specified criterion. The rating of performance on the eight indicators was the major difference between the two types of instrument formats used in this study (the GRM and DSRM). Table 8 presents the evaluator and expert panel ratings on these indicators imbedded in the DSRM/Forced Indicator Rating instrument format. Although these data were not used in data analysis, it is instructive to briefly discuss the findings.

The expert panel rated Questioning Techniques and Feedback "must improve." A majority of evaluators (over 50%)

TABLE 8. A comparison of performance ratings on the criterion and on eight indicators by evaluators who used both forms of the DSRM/Forced Indicator Rating instrument format

	Number of evaluators who moved two or more spaces using the 1-22 grid (N=56)	+ = higher rating - = lower rating
Criterion	7	-
Clarity of Directions	13	-
Logical Concepts	12	-
Questioning Techniques	10	-
Feedback	8	-
Speech Rate	13	-
Delivery Skills	6	-
Body Movement	21	-
Vocabulary Level	8	-

N = 56

did not rate any indicator "must improve"; the largest percentage of evaluators rating any indicator "must improve" was 42% on the indicator, Feedback. Neither the expert panel nor the evaluators rated performance on any indicator as "exemplary." The expert panel and the evaluators were in relative agreement on the ratings of the other indicators with no other noteworthy findings. These findings, though not part of the study hypotheses, show that many evaluators avoided giving extreme ratings on any of the performance indicators.

As was previously mentioned, the DSRM/Forced Indicator Rating format using a continuous scale was provided to fifty-six evaluators in addition to the same format which used a point scale. The DSRM/Forced Indicator Rating format, continuous scale, was developed to ascertain if participants would rate the indicators and the criterion differently when using a continuum rather than the four point rating scale used on the other DSRM/Forced Indicator Rating format and, if so, in what direction. The criterion, indicators, and rating categories were the same as those on the parallel format using a point scale but evaluators were instructed to record the performance ratings anywhere on the continuum - and did so closest to their assessment of the teacher's performance level.

In comparing the performance ratings on indicators and the specified criterion by evaluators who used the DSRM/Forced Indicator Rating format using a continuous scale to their

ratings using the point scale, several interesting patterns emerged. Table 8a presents the results. A majority of evaluators did use the continuum and did, therefore, rate performance differently than when they used the point scale. Twenty-one changed the rating on the indicator Body Movements to a lower rating than when using the point scale. The least number of evaluators changing ratings on any indicator was 4 on Feedback. Most evaluators who changed their rating rated performance lower when using the DSRM/Forced Indicator Rating format, continuous scale; ratings increased in the "must improve" and "needs improvement" range while ratings dropped in the "meets standard" range. As Table 8a presents, the "needs improvement" range of ratings by evaluators who used the continuous scale was the most frequently used on all indicators except Feedback.

In summary, many evaluators who used the continuous scale did rate performance on the criterion and indicators differently and in the direction of lower ratings.

The difficulties associated with statistical analysis of the ratings using the continuous scale were due to the lack of definitive points on each scale that the evaluators could use as a reference. A grid, which divided the continuum into twenty-two equal parts, was developed to be placed over the continuous lines, for the purpose of coding the ratings on each line. However, no referent point (number) was provided on the point scale so that evaluators could mark a rating to

TABLE 8a. Rating differences of evaluators using the DSRM point scale compared to using the DSRM continuous scale (based on frequencies) [a] N = 56

Criterion/Indicators	Must Improve Point/Continuous		Needs Improvement Point/Continuous	
Communicates Effectively with Students	10	15	41	30
Clarity of Directions	5	11	34	41
Logical Concepts	11	15	26	34
Questioning Techniques	13	23	36	31
Feedback	24	28	27	27
Speech Rate	9	13	21	30
Delivery Skills	14	16	31	35
Body Movements	4	16	23	32
Vocabulary Level	14	16	22	28

[a] The 1-22 point grid was used to tabulate frequencies and the following scale was used to determine rating category limits:

1 - 5	Must Improve
6 - 11	Needs Improvement
12 - 17	Meets Standards
18 - 22	Exemplary

Meets Standards Point	Meets Standards Continuous	Exemplary Point	Exemplary Continuous
2	0	0	0
17	0	0	0
18	7	0	0
7	2	0	0
5	1	0	0
26	13	0	0
11	5	0	0
29	8	0	0
20	12	0	0

either side of the referent point when using the continuous scale.

Also, it was difficult to know where one category ended and the other began on the continuous scale making statistical comparisons of ratings with the point scale quite arbitrary.

Finally, the directions on the format which included the continuous scale did not specifically state (as on the point scale format) to place the rating mark on the continuous line at a point where the evaluator assessed the performance level to be on the criterion or indicators. This would have been helpful to the evaluators in clarifying the purpose of the continuum for rating purposes.

Thus, the descriptive findings presented indicated that many evaluators used the continuous scale to rate performance differently (lower) but many questions relating to scale design remain unanswered and should be pursued in further research.

Inferential Statistics

Four hypotheses provided focus for the study. These are provided in operational form below and in the null form later for testing. Significance was set at the .05 level and reported at that level and beyond.

Hypotheses

1. The mean score rating on the performance criterion of evaluators who used the Double Scale Response Mode/Forced Indicator Rating format using a point scale will be significantly closer to those of the expert panel mean score rating than the mean score rating of those who used the Graphic Response Mode/Indicator format.
2. There will be significantly less variance among ratings of evaluators who used the Double Scale Response Mode/Forced Indicator Rating format using a point scale than those who used the Graphic Response Mode/Indicator format.
3. The identified job improvement targets by evaluators who used the Double Scale Response Mode/Forced Indicator Rating formats will be significantly closer to those of the expert panel than those who used the Graphic Response Mode/Indicator format.
4. The identified performance areas to reinforce by evaluators who used the Double Scale Response Mode/Forced Indicator Rating formats will be significantly closer to those of the expert panel than those who used the Graphic Response Mode/Indicator format.

Hypotheses testing

In this sub-section, the results of the hypotheses testing are reported. Four hypotheses were stated in the null form and tested using analysis of variance, Levene Test of Equality of Variance, and chi-square. The first two hypotheses compared evaluator's ratings of performance on a specified criterion, using different recording instrument formats, to those of an expert panel (validity) and to evaluators using identical instrument formats (reliability). The results showed which format led to more valid performance ratings and how much variance in ratings occurred. The last two hypotheses examined the effect of the use of either the DSRM/Forced Indicator Rating formats or the GRM/Indicator format in evaluator's ability to identify performance areas for improvement or reinforcement.

The first hypothesis was designed to compare the ratings of performance on a specified criterion by evaluators using different instrument formats to determine which format led to a more valid rating of the performance.

Ho_1: The mean score rating on the specified criterion by evaluators who used the DSRM/Forced Indicator Rating (point scale) format will be the same as or further from those of an expert panel than the mean score rating of evaluators who used the GRM/Indicator format.

Table 9 presents the data for the first hypothesis. As the table shows, there was a highly significant difference

($p < .001$) between the mean score ratings of each group of evaluators. Those who used the DSRM/Forced Indicator Rating (point scale) format, requiring the rating of eight indicators before making the rating of performance on the criterion, had a mean score rating of 1.8491 compared to 1.3878 for those who used the the GRM/Indicator format. The DSRM/Forced Indicator Rating instrument format ratings were closer to those of the expert panel (2.0). Since the mean score ratings of the evaluators who used the DSRM/Forced Indicator Rating format were significantly closer to the expert panel at the .001 level, hypothesis one was rejected.

TABLE 9. Analysis of variance using criterion ratings, by group

Group	N	Mean	Absolute Value (mean difference from the expert panel)	F Value
1 (GRM)	49	1.3878	0.47481224	10.23***
2 (DSRM)	53	1.8491	0.32037170	

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

The second hypothesis was formulated to examine which rating format resulted in the most within group variance in ratings of performance on the specified criterion (inter-rater reliability).

Ho_2: There is no difference in rating variance among evaluators who used the DSRM/Forced Indicator Rating (point scale) format than among evaluators who used the GRM/Indicator format.

Table 10 presents the data for testing the second hypothesis. There was a highly significant difference in the variance of ratings between the two groups. Evaluators who used the DSRM/Forced Indicator Rating (point scale) format had less variance in ratings than those who used the GRM/Indicator format. Since the difference in variance was highly significant at the .0001 level, the hypothesis was rejected.

TABLE 10. Analysis of variance of mean differences by group

Group	N	Absolute value (Evaluators' score minus the mean)	Absolute value	F Value
1 (GRM)	49	-0.61224490	0.61224490	18.12****
2 (DSRM)	53	-0.15094340	0.22641509	

**** Significant at $p < .0001$ level.

Hypothesis three was formulated to examine if the use of the DSRM/Forced Indicator Rating formats assisted evaluators in identifying performance areas to target for improvement

relating to the criterion "Communicates Effectively with Students."

Ho_3: The identified job improvement targets by evaluators who used the DSRM/Forced Indicator Rating format will be equal to or further from those identified by the expert panel than those who used the GRM/Indicator format.

Table 11 presents the results for the third hypothesis. To determine statistical significance, it was necessary to calculate the standard normal (Z) approximation to the binomial test with significance set at the .05 level. As the table shows, there was no significant difference between evaluators, using either format, in identifying performance areas to target for improvement. Since none were significant at the .05 level, hypothesis three was not rejected.

TABLE 11. Summary of (Z) calculations, by format, of areas identified for improvement

Expert Panel Identified Areas	Ratios (first choice) GRM	DSRM	(Z)	Ratios (second choice) GRM	DSRM	(Z)
Questioning Techniques	6:48	7:54	.069*	8:47	8:51	.227*
Feedback	2:48	3:54	1.035*	2:47	2:51	.083*

*None significant for ± 1.96, $p < .05$.

Hypothesis four was designed to examine if instrument format made a difference in evaluators' ability to identify

performance areas to reinforce relating to the specified criterion.

> Ho_4: The identified reinforceable areas of performance by evaluators who used the DSRM/Forced Indicator Rating format will be equal to or farther from the areas identified by the expert panel than those identified by evaluators who used the GRM/Indicator format.

For this hypothesis to be rejected, both the first and second choices by the expert panel had to be selected by significantly more of evaluators who used the DSRM/Forced Indicator Rating format. Significance was set at the .05 level.

Table 12 presents the results for hypothesis four. The table shows that the evaluators who used the GRM/Indicator format were significantly closer to the expert panel than those who used the DSRM/Forced Indicator Rating format in the selection (as a first or second choice) of one performance area to reinforce (Logical Concepts). However, no significant differences between groups was found in the second performance area (Speech Rate) identified by the expert panel. Thus, hypothesis four was not rejected.

TABLE 12. Summary of (Z) calculations, by format used, of areas identified for reinforcement

Expert Panel Identified Areas	Ratio (first choice) GRM	DSRM	(Z)	Ratio (second choice) GRM	DSRM	(Z)
Logical Concepts	11:47	5:52	3.549*	9:41	2:45	2.446*
Speech Rate	1:47	3:52	.944	2:41	4:45	.740

* Significant for $^{+}1.96$, $p < .05$.

CHAPTER V. SUMMARY, CONCLUSIONS, AND RECOMMENDATIONS

The purposes of this study were to 1) compare the efficacy of two teacher evaluation formats to determine which would assist evaluators in making valid rating of teacher performance on a specified criterion, 2) determine if instrument format affected agreement among evaluators in their rating of performance on a specified criterion (inter-rater reliability), 3) determine if instrument format influenced evaluators' ability to identify teaching behaviors to improve and reinforce, and 4) examine how a continuous point scale affected ratings. In this chapter, conclusions of the study based on the analysis of data are reported and recommendations for further research are presented. The chapter has been organized into three sections: 1) summary and conclusions from the data, 2) limitations, and 3) recommendations for further research.

Summary and Conclusions from the Data

The data gathered for this study were collected in a workshop from trained administrators responsible for evaluating teachers. These data were used to examine the effects of a continuous scale and to test four hypotheses related to instrument format. The findings are presented in summary form followed by discussion.

Results from hypotheses testing

Evaluators were asked to rate a videotaped lesson on the criterion "communicates effectively with students." Two types of instrument formats were provided to evaluators to record their ratings of performance; forty-nine evaluators used the Graphic Response Mode/Indicator (GRM) format while fifty-six used two forms of the Double Scale Response Mode/Forced Indicator Rating (DSRM) format which required evaluators to rate eight performance indicators prior to making a rating of teacher performance on the specified criterion. Study findings indicate the following:

1. Significantly more of the evaluators who used the DSRM/Forced Indicator Rating instrument format (point scale) agreed with the expert panel rating of the teacher's performance on the criterion "communicates effectively with students" than did those who used the GRM/Indicator format. The expert panel rating was "needs improvement."
2. The mean score ratings of evaluators who used the DSRM/Forced Indicator Rating instrument format (point scale) were significantly closer to those of the expert panel than were those who used the GRM/Indicator instrument format.
3. There was less variance in the ratings of evaluators who used the DSRM/Forced Indicator

Rating instrument format than those who used the GRM/Indicator format. The evaluators who used the DSRM/Forced Indicator Rating format were in greater agreement on the rating of the teaching performance on the criterion than were the evaluators who used the GRM/Indicator format.

4. Instrument format did not significantly influence evaluators' ability to identify teaching behaviors (related to the criterion) to improve upon or reinforce. There was no significant difference in the ability of evaluators to identify improvement or reinforcement areas using either the DSRM/Forced Indicator Rating format or the GRM/Indicator format.

Discussion

Ratings of evaluators using the DSRM/Forced Indicator Rating instrument format were closer to the expert panel rating than were those of evaluators using the GRM/Indicator instrument format. It seems that instrument format may help influence teacher performance ratings and enhance the validity of ratings. Evaluators who used the DSRM/Forced Indicator Rating instrument were required to rate performance on indicators which characterized effective communication in the classroom prior to rating the teaching performance on the

criterion. Evaluators who used the GRM/Indicator instrument format had these same indicators but they were provided on a separate sheet in the materials packet and were for reference and clarification purposes only. It may be that the forced rating of indicators "caused" the evaluators to consider each facet of performance more clearly and reduced the tendency to make a global and less precise rating.

Because the evaluators who used the GRM/Indicator format did not have the indicators (descriptors) on the same page, and because they were not forced to rate those indicators, perhaps they were less inclined to assess the important aspects of communication in the classroom. Teacher evaluation has been criticized for being too subjective. Perhaps rating indicators helps to reduce subjectivity. Given the need to develop instruments which can assist evaluators in making valid, discriminating performance ratings, this may help to spur further research.

It was surprising that little difference in evaluators' ability to target performance areas to improve or reinforce was found. Evaluators who used the DSRM formats did have the indicator ratings to assist in the identification of either the "weak" or the "strong" teaching behaviors while those who did not have those ratings available used the GRM/Indicator format. However, most evaluators who used the DSRM formats did not rate any indicator as "must improve." In contrast, the expert panel rated two areas as "must improve" -

Questioning Techniques and Feedback - and then targeted these for improvement. It would be interesting to know that if the evaluators had rated performance on any indicator as "must improve", would they then have used that rating as their guide for identifying teaching behaviors to improve? Similarly, most evaluators rated only two indicators as "meets standards" - Speech Rate and Body Movement. Body movement tied for third choice of evaluators who used the DSRM/Forced Indicator Rating format as a teaching behavior to reinforce but, Speech Rate ranked 6th in their identified reinforceable behaviors. Had the evaluators used their indicator ratings to assist in identifying performance areas to reinforce, Speech Rate could have ranked higher and also would have been one of the same areas identified by the expert panel.

It appears that being able to make a more valid rating of teacher performance on a specified criterion and choosing target areas for growth and areas for reinforcement require different processes. It's back to the drawing board on this one.

In summary, the data confirmed the assumption that instrument format may affect evaluators' ability to make valid performance ratings. Requiring evaluators to rate performance indicators before rating the teacher performance on the specified criterion led to ratings closer to the expert panel than did the format which included the indicators in the packet of materials but did not require rating. Focusing on

indicators, descriptors characterizing effective performance on a given criterion, and rating these indicators may have helped the evaluators. Instrument format, however, did not seem to influence evaluators' ability to identify areas to improve or reinforce.

If instrument format can influence valid and reliable teacher ratings in one criterion, then it seems like a fertile area for further study.

Limitations

It is instructive to delineate the limitations of this study.

1. The use of one criterion and the indicators specifically characterizing performance on that criterion limited the scope and perhaps how generalizable the findings are to other criteria.
2. The fact that the administrators who participated in this study were from a ten district area in a midwestern state may have affected performance ratings due to potential similarities in background, experience, philosophy, and training.
3. Because this was a pioneer study - one new to the research comparing instrument formats - and

because the scope of the study was limited, the findings may not be substantiated in similar or broader contexts.

Recommendations for Further Research

Study results do suggest some areas for further research. In addition, it should be noted that this was a pioneer study into unexplored territory. I make no apologies for the research effort. I do feel a need to provide those who would pursue a similar study with suggestions for improving upon study design and procedures.

1. The study should be replicated. While improvements in design and procedures should be made, the basic design should be followed to further support findings of this study.
2. It might be interesting to devise a method to monitor the process that evaluators use in rating a criterion. It is possible that in this study they rated the criterion first and then rated the indicators. Three of the evaluators did not rate the criterion but did rate the indicators leading one to believe that they did follow the directions but without a monitoring system, no one can be certain.

3. It was assessed that instrument format may have affected the making of ratings closer to those of the expert panel, but that instrument format had no significant affect on raters' ability to identify areas to improve or reinforce. It is recommended that further study be conducted to assess how raters arrive at target areas for growth and reinforcement. In other words, devise a method to examine the procedure or process that evaluators use in selecting the behaviors that need to be improved and those that can be strengthened.
4. Before any firm conclusions can be drawn, it is suggested that this study be conducted using multiple examinations of a number of criteria and ratings. Rating other criteria using the two types of instrument formats designed for this study may help us to understand if findings can be generalized across criteria.
5. Additional research should be conducted to ascertain if findings are generalizable to other grade levels and content areas.
6. Rather than the expert panel, a panel of trained educators (field-based) could provide the source for data analysis. These trained educators would discuss teaching performance on criteria

and then use instruments designed for study purposes to rate performance.

7. The DSRM/Forced Indicator Rating instrument format using a continuous scale was developed to examine differences in participant ratings when using a continuum as opposed to using a point rating scale. Due to problems stemming from scale design, the statistical analysis of data collected on this instrument format was not possible. To determine the actual differences in rating, further research could be conducted using numerical points on both the point scale and continuous scale formats. Then, statistical comparisons could be made to determine if ratings would significantly change and, if so, in which direction and how much when using a continuous scale.

BIBLIOGRAPHY

Aiken, L. R. (1983). Number of response categories. Educational and Psychological Measurement, 43, 397-401.

Alkin, M. C., Kosecoff, J., Fitz-Gibbon, C., & Seligman, R. (1974). Evaluation and decision-making - Title VII experiment. Los Angeles, CA: University of California

Beckham, J. C. (1981). Legal aspects of teacher evaluation. Topeka, KA: National Organization on Legal Problems of Education.

Berliner, D. C. (1977). Instructional time in research on teaching. San Francisco, CA: Far West Laboratory for Educational Research and Development.

Borich, G. D. (1977). The appraisal of teaching: Concepts and process. Reading, MA: Addison-Wesley.

Borman, W. C. (1979). Performance evaluation ratings as cited in F. J. Landy & J. L. Fan (1983). The measurement of work performance, methods, theory, applications. New York, NY: Academic Press.

Boyer, E. & Levine A. (1981). A quest for common learning: The aims of general education. Washington, D.C.: Carnegie Foundation for the Advancement of Teaching.

Brandt, R. (1973). Observation in supervisory practice and school research. In observational methods in the classroom (pp. 79-83). Washington, D. C.: Association for Supervision and Curriculum Development.

Brophy, J. E. (1978). Context variables in teaching. Educational Psychologist, 12, 310-16.

Brophy, J. E. & Evertson, C. (1974). Process-product correlations in the Texas teacher effectiveness study: Final report. Austin, TX: Research and Development. Center for Teacher Education.

Brophy, J. & Evertson, C. (1976). Learning from teaching: A developmental perspective. Boston, MA: Allyn and Bacon.

Browder, L. H., Atkins, W. A., & Kaya, E. (1973). Developing an educationally accountable program. Berkeley, CA: McCutchan.

Bruner, J. S. (1976). Forward. In N. Bennett, J. Jordan, G. Long, & B. Wade (Eds.). Teaching styles and pupil progress. Cambridge, MA: Harvard University Press.

Carfield, R. D. & Walter, J. K. (1984). Teacher Evaluation and RIF - Can there be peaceful coexistence? NASSP, Bulletin (168), 475, November.

Centra, J. A. & Potter, D. A. (1980). School and teacher effects: An interrelational model. Review of Educational Research, 50(2), 273-291.

Clements, A. & Evertson, C. M. (1981). In J. Millman (Ed.) Handbook of teacher evaluation. Beverly Hills, CA: Sage Publications.

Coleman, J. S. (1983). The report of the president's commission on the excellence in education. Washington, D. C.: U. S. Office of Education.

Denham, C. & Lieberman, A. (1981). Part one: Policy making in education. Chicago, IL: National Society for the Study of Education.

DiRocco, A. & Igoe, J. (1977). Teacher evaluation. Albany, N. Y.: Thealan Associates Incorporated.

Dunkin, M. J. & Biddle, B. J. (1974). The study of teaching. New York, NY: Holt, Rinehart and Winston.

Dunkleberger, G. (1982). Classroom observations, "What should principals look for?" NASSP Bulletin, (66), 458.

Edmonds, R. (1978). A discussion of the literature and issues related to effective schooling. Presented at the National Conference of Urban Education, St. Louis, Missouri (July).

Educational Research Service Report. (1978). Evaluating teacher performance. Arlington, VA: ERS Incorporated.

Educational Research Service Report: ERS. (1972). Teacher Evaluation circular No. 2 at NEA Convention, February, National Education Association Research Division and American Association of School Administrators. Washington, D. C.: ERS Incorporated.

Floden, R. E. & Weiner, S. S. (1978). Rationality to ritual: The multiple roles of evaluation in governmental process. Policy Science, 9, 9-18.

Franklin, J. & Thrasher, J. (1976). An introduction to program evaluation. New York, NY: Wiley.

Gallop, G. H. (1979). The eleventh annual Gallup Poll of the public's attitudes toward the public schools. Phi Delta Kappan, 60, 33-45.

Garfield, R. D. and Walter, J. K. (1984). Teacher Evaluation and RIF - Can there be peaceful coexistence? NASSP, Bulletin (168), 475, November.

Glass, G. (1977). Teacher indirectedness and student achievement. Denver, CO: Laboratory of Educational Research.

Good, C. & Barr, D. (1935). The methodology of educational research. New York, NY: Appleton-Century-Crofts.

Good, T. L. & Power, C. N. (1976). Designing successful classroom environments for different types of students. Journal of Curriculum Studies, 8(1), 45-60.

Griffith, F. (1973). Handbook for the observation of teaching and learning. Midland, MI: Pendall.

Haefele, D. L. (1980). How to evaluate thee, teacher - let me count the ways. Phi Delta Kappan, 5, 349-352.

Henderson, R .I. (1984). Performance Appraisal. Reston, VA: Reston.

Howsam, R. B. (1973). Current issues in evaluation. National Elementary Principal, 52, 12-17, February.

Hunter, M. & Russell, D. (1977). Improved instruction. El Segundo, CA: TIP.

Iwanicki, E. F. (1981). Contract plans: A professional growth-oriented approach to evaluating teacher performance. In J. Millman (Ed.) Handbook of Teacher Evaluation. Beverly Hills, CA: Sage Publications.

Jacobs, R., Kafry, D. & Zedeck. S. (1980). Expectations of behaviorally anchored rating scales. Personnel Psychology, 33.

Joyce, B. R. & Weil. M. (1972). Models of teaching. Englewood Cliffs, N. J.: Prentice-Hall.

Knapp, M. S. (1982). Toward the study of teacher evaluation as an organizational process: A review of current research and practice. Menlo Park, CA: SRI.

Lamb, L. & Swick, K. (1975). A historical overview of classroom and teacher observation. The Educational Digest, 40, 39-42.

Landy, F. J. & Fan, T. L. (1983). The measurement of work performance, theory, applications. New York, NY: Academic Press.

Lucio, W. & McNeil, J. (1979). Supervision in thought and action. New York, NY: McGraw-Hill.

Manatt, R. (1981). Teacher performance criteria and student gains. National Science Foundation, Grant No. GV3373. Ames, Ia: Iowa State University.

Manatt, R. P., Palmer, K. L. & Hidlebaugh, E. (1976). Evaluating teacher performance with improved rating scales. NASSP Bulletin, 60(401), 21-23.

Mazur, J. (1980). Issues related to measurement of teaching performance. Due process in teacher evaluation. Washington, D. C.: University Press of America.

McGreal, T. L. (1983). Successful teacher evaluation. Alexandria, VA: Association for Supervision and Curriculum Development.

McKenna, B. H. (1981). Context/environment effects in teacher evaluation. In J. Millman (Ed.) Handbook on teacher evaluation. Beverly Hills, CA: Sage Publications.

McLaughlin, M. W. (1982). A preliminary investigation of teacher evaluation practices. Santa Monica, CA: National Institute of Education.

McNeil, J. D. (1981). The politics of teacher evaluation. In J. Millman (Ed.) Handbook of teacher evaluation. Beverly Hills, CA: Sage Publication.

Medley, D. (1979). The effectiveness of teachers. In P. L. Peterson & H. J. Walberg (Eds.). Research on teaching. Berkeley, CA: McCutchan.

Menne, J. W. (1972). Teacher evaluation. Unpublished paper at Teacher Evaluation Conference, Ames, Iowa. November 26-27.

Millman, J. (1981). Handbook on Teacher evaluation. Beverly Hills, CA: Sage Publications.

Peterson, D. & Peterson, K. (1984). A research based approach to teacher evaluation. NASSP Bulletin, (68) 469.

Peterson, K. & Kauchak, D. (1982). Teacher evaluation: perspectives, practices and promises. Salt Lake City, UT: Center for Educational Practice, University of Utah.

Popham, W. J. (1974). Pitfalls and pratfalls of teacher evaluation. Educational Leadership, 32, 141-146.

Popham, J. (1975). Educational evaluation. Englewood Cliffs, NJ: Prentice Hall.

Redfern, G. (1980). Evaluating teachers and administrators: A performance objectives approach. Boulder, CO: Westview Press.

Redfern, G. (1972). How to evaluate teaching: A performance objectives approach. Worthington, OH: School Management Institute.

Rice, B. (1985a). Performance review: The job nobody likes. Psychology Today, September, 19, 30-36.

Rice, R. (1985b). Summary of teacher evaluation instruments using DSRM format. Unpublished. Assistant Superintendent, Mason City, IA.

Rosenshine, B. (1970). The stability of teacher effects upon student achievement. Review of Educational Research, 40, 647-662.

Rosenshine, B. (1979). Content, time, and direct instruction. In P. L. Peterson and H. J. Walberg (Eds.) Research on teaching. Berkeley, CA: McCutchan.

Rosenshine, B. (1972). Review of teaching variables and student achievement. In G.D. Borich (Ed.). The appraisal of teaching: concepts and process. Reading, MA: Addison-Wesley.

Rummel, F. (1958). Introduction to research procedures in education. New York, NY: Harper and Brothers.

Sax, G. (1974). Principles of educational measurement and evaluation. Belmont, CA: Wadsworth Publications.

Scriven, M. (1973). School evaluation. Berkeley, CA: McCutchan.

Scriven, M. (1981). Summative teacher evaluation. In J. Millman (Ed.). Handbook of teacher evaluation (pp 244-271). Beverly Hills, CA: Sage Publications.

Shepherd, G. D. & Ragan, W. B. (1982). Perspective: society. modern elementary curriculum. New York, NY: Holt, Rinehart, and Winston.

Snedecor, G. W. & Cochran, W. G. (1981). Statistical methods. Ames, IA: Iowa State University Press.

Soar, R. S., Medley, D. M. & Coker, H. (1983). Teacher evaluation: A critique of currently used methods. Phi Delta Kappan, 65, 239-246.

Stallings, J. A. (1977). How instructional processes relate to child outcomes. In G. D. Borich (Ed.) The appraisal of teaching: concepts and process. Reading, MA: Addison-Wesley.

Stephens, J. M. (1976). The process of schooling. New York NY: Holt, Rinehart and Winston.

Stow, S. B. & Sweeney, J. (1981). Developing a teacher performance evaluation system. Educational Leadership 38, 538-541.

Twentieth Century Fund Task Force. (1981). Report on the federal elementary and secondary education policy. Washington, D. C.: Education Committee.

Strike, K. & Bull, B. (1981). Fairness and the legal context of teacher evaluation. In J. Millman (Ed.) Handbook of teacher evaluation (pp. 303-343). Beverly Hills, CA: Sage Publications.

Walberg, H. (1974). Evaluating educational performance. Berkeley, CA: McCutchan.

Wise, A. E., Darling-Hammond, L. McLaughlin, M. W. & Bernstein, H. T. (1984). Teacher evaluation: A study of effective practices. Santa Monica, CA: The Rand Corporation.

Wise, A., Darling-Hammond, L. & Pease, S. (1982). Teacher evaluation in the organizational context: A review of the literature. Santa Monica, CA: Rand Corporation.

ACKNOWLEDGEMENTS

Sincere appreciation is expressed to members of the Program of Studies Committee for their individual and collective assistance in the preparing, conducting, and writing of this research study. A very special thank you to Dr. Richard Manatt for conducting the field test and study yielding the date for analysis.

Also, appreciation is expressed to Dr. Shirley Stow for her expertise, words of encouragement, and support.

Dr. Robert Strahan has my utmost respect and regard for the many hours spent with me in the process of preparation, analysis, and interpretation of data.

For the unequivocal support, assistance and patience while I devoted time to this study away from my position as Director of Curriculum and Staff Development, I extend deep and heartfelt appreciation to Dr. Roger Worner, Superintendent of Schools, Mason City, Iowa.

Judy Stokes, my secretary, has my most deserving and sincere appreciation for her dedication throughout this study to read my incoherent writing, to patiently make numerous revisions, and to maintain her cheerful nature throughout these past months.

My greatest indebtedness goes to Dr. James Sweeney, my major professor, for his initial and continued support, encouragement, and unfaltering belief in my abilities.

I dedicate this study to my children, Ryan, Kenneth, and Shawn for their tolerance of my need to work on this research study taking precious time away from them. And I dedicate this completed dissertation to my parents, Hal and Grace Cooper, for their love and belief that their daughter can accomplish any goal in the world.

APPENDIX A:

GRAPHIC RESPONSE MODE (GRM)/INDICATOR INSTRUMENT FORMAT

GRAPHIC RESPONSE MODE (GRM)

(Evaluator's I.D.#)

DIRECTIONS: Please use the indicators listed on the following page to assist you in making your rating. After viewing the videotape, please check the line above the statements which best describe the evaluatee's performance on that item. The final page of the GRM allows space for you to write your identified strengths and weaknesses for this teacher on this given criterion.

CRITERIA	LEVELS OF PERFORMANCE			
	Must Improve	Needs Improvement	Meets Standard	Exemplary
The teacher...				
A. Communicates Effectively with Students	Communications from the teacher are frequently unclear; students often appear confused	Communications from the teacher are usually clear but student input is not encouraged	Communications from the teacher are clear; relevant dialogue is encouraged	In addition, the teacher is extremely skillful in using a variety of verbal and nonverbal communications

APPENDIX B:

DOUBLE SCALE RESPONSE MODE/FORCED INDICATOR RATING INSTRUMENT FORMAT USING A POINT SCALE

Evaluator's I.D.#

DOUBLE SCALE RESPONSE MODE

Directions: After viewing the video tape, please place a check on one of the blanks provided beside each indicator showing what you believe to be the most appropriate level of performance for that indicator. After all indicators have been rated, please make an overall rating of the criterion by placing a check on one of the level of performance blanks by the stated criterion.

CRITERIA

I. Communicates Effectively	Must Improve	Needs Improvement	Meets Standard	Exemplary

INDICATORS	Must Improve	Needs Improvement	Meets Standard	Exemplary
Clarity of Directions				
Presents Concepts/Ideas Logically				
Questioning Techniques				
Feedback to Students				
Rate of Speech				
Delivery Skill (pitch, volume, speech patterns)				
Body Movements, Gestures				
Vocabulary				

APPENDIX C:

DOUBLE SCALE RESPONSE MODE/FORCED INDICATOR RATING INSTRUMENT FORMAT USING A CONTINUOUS SCALE

Evaluator's I.D.#

DOUBLE SCALE RESPONSE MODE

Directions: Now, please rate each indicator again by placing a check on the continuous line next to each indicator. The object is for you to determine if you would rate the indicator any differently along a continuous scale rather than in distinct categories. After completing the ratings of each indicator, please make an overall rating along the continuous line for the stated criterion.

CRITERIA

I. Communicates Effectively	Must Improve	Needs Improvement	Meets Standard	Exemplary

INDICATORS	Must Improve	Needs Improvement	Meets Standard	Exemplary
Clarity of Directions				
Presents Concepts/Ideas Logically				
Questioning Techniques				
Feedback to Students				
Rate of Speech				
Delivery Skill (pitch, volume, speech patterns)				
Body Movements, Gestures				
Vocabulary				

APPENDIX D:

EXPLANATION OF THE RATING SCALE CATEGORIES USED IN THE GRM AND DSRM FORMATS

EXPLANATION OF THE SCALE USED IN THE GRM AND DSRM

Must Improve:	Performance jeopardizes continued employment in the district.
Needs Improvement:	Performance is below the district expectations.
Meets Standard:	Performance meets the expectations set by the district.
Exemplary:	Performance exceeds district expectations.

APPENDIX E:

INDICATOR EXPLANATION SHEET USED FOR THE GRM INSTRUMENT

GRM - INDICATOR EXPLANATION

CRITERIA	INDICATORS
I. Communicates Effectively with Students	The teacher....
	1. gives clear, concise and reasonable directions
	2. presents concepts/ideas logically
	3. uses questioning techniques
	4. provides feedback to students
	5. varies rate of speech to coincide with verbal content
	6. appears aware of delivery skills: pitch (high, low) volume (loud, soft) word patterns or repetitions
	7. uses body movements and gestures which enhance the message
	8. uses vocabulary at age level of students

APPENDIX F:

IMPROVEMENT AND STRENGTH AREAS REPORTING FORM

IMPROVEMENT AND STRENGTH AREAS REPORTING FORM

Please identify the areas you would consider to be strengths for the teacher under the criterion "communication." Strengths are areas you should reinforce and that the teacher can build upon to become even more effective. Please list from one to three of the most important reinforceable areas in the spaces provided below.

1.

2.

3.

Targets for Growth refer to teaching behaviors you would choose to focus upon in a conferencing situation that are vital for that teacher to improve. Please identify the two major target growth areas in the communication area that you would bring to the teacher's attention for improvement. Please prioritize by number.

1.

2.

APPENDIX G:

REGISTRATION CARD FOR DEMOGRAPHIC INFORMATION

REGISTRATION CARD
FORM A

I.D. #_____

DATE_____

CITY & STATE_____

(check all that apply)

JOB TITLE:
- Superintendent _____
- Asst. Superintendent _____
- Principal _____
- Asst. Principal _____
- Supervisor _____
- Department Head _____
- Teacher _____
- Other _____

JOB LEVEL:
- Preschool _____
- Elementary _____
- Middle School _____
- Junior High _____
- High School _____
- Other _____

SIZE OF YOUR SCHOOL DISTRICT:

0-1000 _____	3000-4000 _____	6000-7000 _____
1000-2000 _____	4000-5000 _____	7000-8000 _____
2000-3000 _____	5000-6000 _____	Over 8000 _____

NUMBER OF TEACHERS YOU ARE RESPONSIBLE FOR EVALUATION(TOTAL):

0 _____	21-30 _____	51-60 _____
1-10 _____	31-40 _____	61-70 _____
11-20 _____	41-50 _____	Over 70 _____

LENGTH OF TIME YOU HAVE BEEN RESPONSIBLE FOR TEACHER EVALUATION (IN YEARS):

0-1 _____	6-7 _____	12-13 _____
2-3 _____	8-9 _____	14-15 _____
4-5 _____	10-11 _____	Over 15 _____

PREVIOUS TRAINING IN TEACHER EVALUATION: (DO NOT COUNT THIS WORKSHOP)

Workshop (on your own)_____	Coursework _____
Workshop (required) _____	Previous or Present Administrator _____
District Inservice _____	Other _____

GENERAL COMMENTS RELATING TO TEACHER EVALUATION:

The above identification number is assigned to you and you only. Record this number and use it on all forms throughout this workshop. Information on this card shall be used for research only and will not be released in any form that will be identifiable to you. Thank you.

APPENDIX H:

INFORMATION/DIRECTION SHEET

INFORMATION/DIRECTION SHEET

Because teacher evaluation is mandated by nearly every state in the nation, it has become a vital component in improving instruction in the classroom. However, summative evaluation instruments may and do vary in depth, coverage, and format indicating a lack of consensus as to what type of instrument is most effective in discriminating among various levels of teacher performance. It is the purpose of this activity to examine instrument format, one component of summative evaluation, to determine if it alone affects evaluator rating of teacher performance on a given criterion. Also, this data, once collected, is part of a doctoral dissertation regarding instrument format in teacher evaluation. Your participation is not mandatory but would help to facilitate data collection leading to substantive conclusions regarding the influence of instrument format in rating teacher performance. If you do choose to participate you also have the opportunity to receive the final conclusions regarding the data analysis. The last sheet in this packet is a registration sheet asking for various types of information. At no time will you be identified in this study; the ID number is for record keeping purposes of how many individuals in the country have participated. If you should want a copy of the final results please write your name and address on a separate sheet of paper and turn it into the workshop facilitator.

Thank you for your cooperation and time!

Directions:
After receiving explanation, training, and guided practice you will:

1) Receive a packet of materials.
2) View the videotape.
3) Rate the performance of the teacher on "Communicates Effectively with Students" following the directions you receive. You should be using the format provided in your packet. (Remember to concentrate on "communication" rather than his/her teaching in general.)
4) Identify in writing, in the space provided on the format, one to three major strength areas you would reinforce to the teacher in "Communicating Effectively with Students."
5) Identify in writing, in the space provided on the format, two areas you would target as needing improvement in "Communicating Effectively with Students." Please prioritize by number.
6) Complete the last sheet of the packet.
7) Return the packet to the workshop coordinator.