Surgency bias: effects on teacher performance evaluation

Robert L. Newsum
Iowa State University

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

Part of the Educational Administration and Supervision Commons

Recommended Citation

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

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

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

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

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps. Each original is also photographed in one exposure and is included in reduced form at the back of the book.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.
Surgency bias: Effects on teacher performance evaluation

Newsum, Robert L., Ph.D.

Iowa State University, 1990
Surgency bias: Effects on teacher performance evaluation

by

Robert L. Newsum

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

1990
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>CHAPTER I. INTRODUCTION</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Statement of the Problem</td>
<td>1</td>
</tr>
<tr>
<td>Purposes of the Study</td>
<td>2</td>
</tr>
<tr>
<td>Objectives of the Study</td>
<td>4</td>
</tr>
<tr>
<td>Research Questions</td>
<td>4</td>
</tr>
<tr>
<td>Hypotheses</td>
<td>5</td>
</tr>
<tr>
<td>Basic Assumptions</td>
<td>6</td>
</tr>
<tr>
<td>Delimitations or Scope of the Study</td>
<td>8</td>
</tr>
<tr>
<td>Definition of Terms</td>
<td>10</td>
</tr>
</tbody>
</table>

| CHAPTER II. REVIEW OF LITERATURE         | 12   |
| Introduction                             | 12   |
| Origins of Performance Evaluation        | 12   |
| Technical Aspects of Teacher Performance Evaluation | 16   |
| Bias in Performance Evaluation          | 19   |
| Halo effect                              | 20   |
| Central tendency                         | 21   |
| Leniency/severity                        | 21   |
| Surgency bias                            | 23   |
| Summary                                  | 24   |

<p>| CHAPTER III. METHODS AND PROCEDURES      | 27   |
| Introduction                             | 27   |
| Materials                                | 28   |
| Subjects used                            | 32   |
| Conducting the experiment                | 33   |
| Data Analyses Procedures                 | 36   |</p>
<table>
<thead>
<tr>
<th>Chapter</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>IV</td>
<td>Analyses and Research Findings</td>
<td>38</td>
</tr>
<tr>
<td></td>
<td>Descriptive Data</td>
<td>38</td>
</tr>
<tr>
<td></td>
<td>Research Hypothesis Testing</td>
<td>47</td>
</tr>
<tr>
<td>V</td>
<td>Summary, Conclusions, and Recommendations</td>
<td>58</td>
</tr>
<tr>
<td></td>
<td>Summary</td>
<td>58</td>
</tr>
<tr>
<td></td>
<td>Conclusions</td>
<td>58</td>
</tr>
<tr>
<td></td>
<td>Discussion</td>
<td>61</td>
</tr>
<tr>
<td></td>
<td>Limitations</td>
<td>62</td>
</tr>
<tr>
<td></td>
<td>Recommendations for Further Research</td>
<td>63</td>
</tr>
<tr>
<td></td>
<td>Recommendations for Practitioners</td>
<td>64</td>
</tr>
<tr>
<td></td>
<td>Bibliography</td>
<td>66</td>
</tr>
<tr>
<td></td>
<td>Acknowledgments</td>
<td>76</td>
</tr>
<tr>
<td>A</td>
<td>Appendix A. Teacher Performance Rating Scale</td>
<td>77</td>
</tr>
<tr>
<td>B</td>
<td>Appendix B. Personal Data Form</td>
<td>79</td>
</tr>
<tr>
<td>C</td>
<td>Appendix C. Permission Slip</td>
<td>81</td>
</tr>
<tr>
<td>D</td>
<td>Appendix D. Workshop Schedule</td>
<td>83</td>
</tr>
<tr>
<td>E</td>
<td>Appendix E. Judy Oliver Surgency Study Packet</td>
<td>85</td>
</tr>
<tr>
<td>F</td>
<td>Appendix F. Procedures and Explanation</td>
<td>107</td>
</tr>
</tbody>
</table>
LIST OF TABLES

Table 1. Research studies investigating rater characteristics as sources of bias in the evaluation of performance

Table 2. ANOVA analysis and means of the differences in overall rating of performance by evaluators grouped into ten-year age brackets

Table 3. T-test analysis of overall ratings of performance between evaluators in Salina and San Antonio

Table 4. ANOVA and means of the differences of the Professional Improvement Commitment (PIC) selection between subjects in the surgent and nonsurgent teacher performance evaluation rater groups in Salina and San Antonio

Table 5. T-test analysis of difference in the average number of days of teacher performance evaluation training between subjects in the surgent and nonsurgent teacher performance evaluation rater groups

Table 6. T-test analysis of the difference in the average number of years of teacher performance evaluation experience between subjects in the surgent and nonsurgent teacher performance evaluation rater groups

Table 7. T-test analysis of the difference in the average prediction of teacher effectiveness between subjects in surgent and nonsurgent teacher performance evaluation rater groups

Table 8. T-test analysis of the difference in average overall ratings between subjects in the surgent and nonsurgent teacher performance evaluation rater groups

Table 9. T-test analysis of the difference in mean ratings on ten specific teacher behaviors between subjects in the surgent and nonsurgent teacher performance evaluation rater groups
Table 10. Pearson correlation coefficient between the amount of teacher performance evaluation training and mean overall ratings  
Table 11. Pearson correlation coefficient between number of years of teacher performance evaluation experience and mean overall ratings  
Table 12. ANOVA and means of the differences in average overall ratings between subjects in the surgent and nonsurgent teacher performance evaluation rater groups at specific levels of administrative assignment  
Table 13. T-test analysis for the difference in average Professional Improvement Commitments (PICs) selected between subjects in surgent and nonsurgent teacher performance evaluation rater groups  
Table 14. T-test analysis of the difference in mean overall ratings between male and female evaluators  
Table 15. Summary table showing findings related to surgency bias
LIST OF FIGURES

Figure 1. Percentages of subjects in this study 39
Figure 2. Percentages of subjects at San Antonio 40
Figure 3. Percentages of subjects at Salina 40
Figure 4. Percentages of subjects by age range 41
Figure 5. Percentages of subjects by position 43
Figure 6. Percentages of subjects by grade level assignment 43
CHAPTER I. INTRODUCTION

More than any other time in our history, mankind faces a crossroads. One path leads to utter despair and hopelessness. The other leads to total extinction. Let's all pray that we have the wisdom to choose correctly (Woody Allen, 1980).

For the performance evaluator, the use of current evaluation systems may represent the path toward "utter despair and hopelessness," while its abandonment represents the path toward "total extinction" (Bernardin and Beatty, 1984).

Of all the educational reforms that have been instituted since the publication of A Nation at Risk in 1983, that with the greatest potential impact on increased student achievement or outcomes is related to the improvement of teaching performance. Teaching performance can be improved by developing and implementing a comprehensive, valid, and reliable system of evaluation (Daniels, 1989). While performance evaluation can be the most powerful tool that supervisors have for improving productivity, it is also capable of stirring strong feelings and conflict in the workplace (Eichel and Bender, 1984). Because of this potential, evaluation responsibilities are often allowed to suffer from benign neglect. This finally results in the curtailing of critical supervisory steps or the permitting of supervisory processes to consistently fall short of stated objectives (Henderson, 1976).

Medley and Coker, recognized teacher performance evaluation researchers, apparently subscribe to the "total extinction" belief. They maintain that supervisors will never possess the abilities, skills, or objective attitude necessary to accurately observe and evaluate another's
teaching performance and have concluded from a recent study, involving 46 principals and 322 teachers, that no support can be made for the widely-held notion that the average principal can accurately judge teacher performance (Medley and Coker, 1987). A considerable amount of research time and attention at Iowa State University has been focused on the assumption that teacher performance should be evaluated and can be improved through implementation of a research-based system of performance evaluation (Manatt and Stow, 1986).

Statement of the Problem

Philip Hosford, retired professor at New Mexico State University, is credited with coining the term "surgency" to describe those physical and personality characteristics that a person possesses that tend to influence another's rating of their competence or performance (Hosford and Martin, 1980). It would appear that educators are becoming more scientific, systematic, and sophisticated in the observation and evaluation of teacher performance (Smith, Peterson, and Micceri, 1987). Unfortunately there is a serious flaw in the formal procedures that are currently used to apply the research base regarding effective instruction. While there may be extensive knowledge about which effective educational practices and behaviors are necessary to provide the most optimum learning experience for children, use of that knowledge in performance appraisal is being filtered through a variety of prejudices and biases. There is a spectrum of biases occurring at any particular time during the evaluation process. Examples range from age, gender, and race, to leniency or severity, halo
effect, and personal biases (Berman and Kenny, 1976; Borman, 1977; Cooper, 1981; Latham and Wexley, 1981; Plumlee, 1983). Unless awareness of these biases is maintained, they may directly affect how administrators react to what is observed, and ultimately, the quality of education provided.

Various forms of bias have been investigated extensively by those with personnel responsibility in private business and industry (Fiske and Cox, 1960; Mullins and Force, 1962; Thornton and Zorich, 1980; Bernardin and Beatty, 1984). In many ways the tasks of isolating, measuring, and providing training to control bias effects are easier in the private sector, because of the ability to conduct "time and motion" studies to count units of outcome (Christner, 1981). First impression bias, related to surgency bias, and its effect on the rating of performance, has been studied by business researchers (Latham, Wexley, and Purcell, 1975). More recent studies have been conducted to determine the impact of factors such as physical attractiveness of the appraisee on the rating of performance (Landy and Farr, 1983; and Henderson, 1984).

It is much more difficult to determine the sources of bias and potential effects in the educational setting, because of the imprecise nature of learning. One of the earliest studies of bias associated with teacher performance evaluation determined that social factors do influence perceptions (Sherif, 1935). In a study conducted in 1960, D. G. Ryans attempted to show a correlation between predicted teacher performance and summative effectiveness rating. Recent studies conducted at Iowa State University (Harrington, 1984; Bourisaw, 1988; and Peterson, 1988) have concluded that biases do influence the rating of teacher performance.
Purposes of the Study

Teachers' confidence and trust in the process of performance evaluation and some of the credence they have in their administrator or evaluator rests, to a great extent, on their perceptions of fair and unbiased treatment throughout the various activities associated with the evaluation process. To overcome the possible effects of surgency bias, if any, there were several steps to be accomplished. The purpose of this study was to determine the association of age, gender, amount of teacher performance evaluation training, amount of administrative experience, and grade level responsibility with surgency bias. Furthermore, the purpose was to attempt to determine when in the evaluation cycle this bias associated with surgency might occur.

Objectives of the Study

To accomplish the purposes of this study, it was necessary to complete the following objectives:

1. Conduct a thorough review of the literature as it relates to teacher performance evaluation and bias.

2. Develop a teacher performance evaluation rating scale to accurately measure classroom performance.

3. Design videotaped preobservation conferences with surgent and nonsurgent teachers portrayed.

4. Develop a simulated script tape of an appropriate lesson to use in conjunction with the rating scale.
5. Develop several possible Professional Improvement Commitments (PICs) to use in conjunction with the rating scale.

6. Administer the videotaped preobservation conferences, script tape, rating scale, and Professional Improvement Commitments (PICs) to a sample population.

7. Assess the extent of surgency bias on evaluator ratings and Professional Improvement Commitment (PIC) selection.

8. Determine the extent to which teacher performance evaluation training is related to surgency bias.

9. Determine the extent to which teacher performance evaluation experience is related to surgency bias.

10. Determine the extent to which gender is related to surgency bias.

11. Determine the extent to which grade level of administration is related to surgency bias.

12. Determine the extent to which district level of administration is related to surgency bias.

Research Questions

As this study was formulated, several research questions were postulated.

Question 1: At the conclusion of a videotaped preobservation conference, will observers predict that a surgent teacher's classroom performance will be more effective than that of a nonsurgent teacher?

Question 2: Will a surgent teacher's overall performance be rated as more effective than a nonsurgent teacher's overall performance?
Question 3: Will specific criteria used to measure the performance of a surgent teacher be rated as more effective than those criteria that are used to measure the performance of a nonsurgent teacher?

Question 4: Will those evaluators with more teacher performance evaluation training show less surgency bias in their rating of teacher performance?

Question 5: Will those evaluators with more teacher performance evaluator experience show less surgency bias in their rating of teacher performance?

Question 6: Will evaluators from any specific grade levels of responsibility rate the overall performance of a surgent teacher as more effective than that of a nonsurgent teacher?

Question 7: Will a Professional Improvement Commitment (PIC) chosen for a surgent teacher be less rigorous than that chosen for a nonsurgent teacher?

Question 8: Will male evaluators' overall ratings of a surgent and a nonsurgent teacher's performance differ from overall ratings of female evaluators?

Hypotheses

The research questions were then developed into hypotheses to be tested:

Hypothesis 1: There is no significant difference in the predicted teaching effectiveness of a surgent teacher and a nonsurgent teacher, after observing them in a videotaped preobservation conference.
Hypothesis 2: There is no significant difference in rating the overall performance of a surgent teacher and a nonsurgent teacher, after observing them in a videotaped preobservation conference and examining a script tape of their lesson.

Hypothesis 3: There is no significant difference in rating the performance of a surgent and a nonsurgent teacher, after observing them in a videotaped preobservation conference and examining a script tape of their lesson, on specific teaching behaviors.

Hypothesis 4: There is no significant positive correlation between evaluator ratings on the overall performance of a surgent and a nonsurgent teacher and the amount of teacher performance evaluation training.

Hypothesis 5: There is no significant positive correlation between evaluator ratings on the overall performance of a surgent and a nonsurgent teacher and the amount of teacher performance evaluation experience.

Hypothesis 6: There is no significant difference in the rating of the overall performance of a surgent and a nonsurgent teacher by elementary level, middle level, secondary level, and central office evaluators.

Hypothesis 7: There is no significant difference in the level of rigor of Professional Improvement Commitments (PICs) selected by evaluators for a surgent and a nonsurgent teacher.

Hypothesis 8: There is no significant difference in the overall rating of a surgent and a nonsurgent teacher's performance by male evaluators and female evaluators.
Basic Assumptions

This study was based on the following assumptions:

1. Surgency is a source of bias in teacher performance evaluation.
2. An evaluator's rating of a teacher's performance represents a valid measure of that teacher's performance in delivering the lesson depicted in the script tape.
3. Existing statistical measures can be used to accurately measure the effects of surgency on an evaluator's rating of a teacher's performance.
4. Surgency can be depicted with a videotaped simulation.

Delimitations or Scope of the Study

This study was intended to determine the effects of surgency bias, if any, on the overall rating of teacher performance and the impact of gender, performance evaluation training, and performance evaluation experience on these ratings. Since participants were not required to provide feedback or discuss ratings of performance with the surgent or nonsurgent teacher, ratings may have been affected. While it is understood that rater characteristics, biases, and attitudes will have an effect on teacher performance evaluation, only surgency bias, gender, performance evaluation training, performance evaluation experience, age, and grade level of administrative assignment were considered for analysis.

This study specifically addressed:

1. Surgency as one of many sources of potential bias in the process of performance evaluation.
2. Surgency bias as it relates to the rater of teacher performance.
3. Surgency bias as it relates to the age of the rater.
4. Surgency bias as it relates to the gender of the rater.
5. Surgency bias as it relates to the amount of teacher performance evaluation training of the rater.
6. Surgency bias as it relates to the amount of teacher performance evaluation experience of the rater.
7. Surgency bias as it relates to the grade level assignment of the rater.
8. Subjects for this study were practicing evaluators from Kansas who were attending a KanLead workshop, practitioners who were attending a session at the national A.S.C.D. convention in San Antonio, and graduate educational administration students who had voluntarily attended teacher performance evaluation training.
9. Actresses who portrayed the surgent and nonsurgent teachers were female, white, and approximately thirty-five years old.

This study did not specifically address:
1. Surgency bias associated with the gender of the appraisee.
2. Surgency bias associated to the age of the appraisee.
3. Surgency bias associated to the race of the appraisee.
4. Any effects of surgency bias related to increased student achievement.
5. Investigating the longitudinal effects of surgency bias.
6. Variation in amount of surgency bias associated with regions of the United States.

7. The effects of surgency bias on teacher performance evaluators who are not seeking additional training.

Definition of Terms

1. **Bias (statistical):** A tendency of an estimate to deviate in one direction from a true value (as by reason of nonrandom sampling).

2. **Central Tendency Bias:** The tendency to cluster rating scores in a narrow range, generally around the "mean" or average.

3. **Discrimination Power:** The ability of an evaluation criterion to identify high, medium, and low teacher performance.

4. **Graphic Response Mode:** An evaluation response that combines descriptive and numerical ratings of performance.

5. **Halo Bias:** The degree to which extraneous factors influence the rating of one's actual performance in a positive manner.

6. **Hunter Model:** An accepted educational model for systematically describing the effective teaching behaviors attributed to planning and delivering a lesson.

7. **Overall Rating:** A single score used to judge a person's total performance.

8. **Personal Bias:** The degree to which "like me/not like me" attitudes distort performance rating. Personal biases include gender, race, age, and social factors.
9. **Rater Bias**: Human error in rating performance which is traced not to actual performance, but to characteristics of the rater or of the situation in which the rating occurs.

10. **Rater Characteristics**: Those traits possessed by the rater that can influence the actual rating of another's performance.

11. **Reliability**: The extent to which ratings of one's performance are consistent among evaluators and across time.

12. **Script Tape**: A written narrative account of the sequence of events that occurred during a period of instruction.

13. **SurGENCY**: An attractive combination of personality factors characterized by quickness and cleverness.

14. **SurGENCY Bias**: A distortion, predilection, or influence in perception or rating due to physical attractiveness and charisma of the ratee.

15. **Validity**: The degree to which a performance rating accurately measures what it is intended to measure.
CHAPTER II. REVIEW OF LITERATURE

Introduction

An organization's formal system of evaluation to improve performance is only as good as the ability of the personnel within the organization to implement it accurately, consistently, and fairly (Bernardin and Beatty, 1984; Manatt, 1984; Daniels, 1989). When the review of literature was conducted regarding the possible relationship between surgency bias and performance evaluation, two major sources of information related to accurately rating performance were explored. One of the areas searched was the body of knowledge from performance appraisal in business and industry. Studies conducted within the past two decades involving the evaluation of educational personnel was the second major source.

The review of literature related to the appraisal of performance in business and industry and the evaluation of educational performance attempted to: (1) provide a brief historical summary of some attempts at evaluating performance, (2) provide a brief summary of the state of the art in teacher performance evaluation, (3) identify and describe those human factors that may affect the validity and reliability of performance evaluation systems.

Origins of Performance Evaluation

Attempts to evaluate the performance of another have occurred throughout the course of human existence. These efforts have consistently been met with suspicion and mistrust. During the Wei Dynasty, third century A.D., an "Imperial Rater" evaluated the performance of the members
of the official family. His methods of appraisal were subjective. In fact, according to the Chinese philosopher, Sin Yu, "the Imperial Rater of Nine Grades seldom rates men according to their merits, but always according to his likes and dislikes" (Eichel and Bender, 1984).

A performance appraisal method used in the early stages of the Industrial Revolution was a rectangular piece of wood approximately two inches long and one inch wide that was painted a different color on each side and hung over the work station of each employee. Each day the supervisor would turn the wood block to the color that he thought denoted the employee's performance for the preceding day--black for bad, blue for indifferent, yellow for good, and white for excellent (Landy and Trumbo, 1981).

In 1915, a delegate to a convention of the National Education Association stated that performance appraisal was demeaning, arbitrary, perfunctory, and superficial. If an organization is to operate a performance appraisal system that improves its productivity and avoids the 1915 impact, it must meet a variety of needs (Henderson, 1984).

A means for using a graphic response model to rate performance was introduced by Paterson in 1922. Growing concern for the arbitrary nature of the graphic system was studied extensively by Ryan in 1958 and alternative formats were introduced by Barrett, Taylor, Parker, and Martens that same year. One format consisted of a horizontal line with fifteen divisions labeled by trait names. No trait definitions were included in this format. Another format was similar, but traits were defined. This work served as the model for work completed in 1963 by
Smith and Kendall. They called their method of rating scale development "behavioral expectation scaling." This then evolved into the Behaviorally Anchored Rating Scale (BARS) (Landy, Zedeck, and Cleveland, 1983). The BARS system is comprised of rating criteria, or anchors, that appear at intervals on a scale and are examples of actual behaviors, rather than definitions or adjectives. There continues to be a problem with identifying anchors that can discriminate for the central portions of the scale (Barnes and Landy, 1979). The graphic response mode format has become widely accepted in the development of teacher performance evaluation instruments (Manatt and Stow, 1984).

The presence of a "halo effect" as a source of influence in rating performance has been considered for many decades and the supposition made that recollection of information needed to complete a rating would be affected by general impressions the rater formed of the ratee (Wells, 1907; Thorndike, 1920). For example, if a ratee were perceived as very friendly, the rater would generalize this perception and overestimate that person's other related characteristics such as sociability, kindness, or honesty. The halo effect would therefore result in a high correlation with those traits or behaviors sharing similar connotations (Berman and Kenny, 1976).

This concept of "halo effect" that is derived from immediate observations was extended by later studies. Researchers found that any correlations between behaviors referring to the same trait were consistently higher when they were derived from retrospective ratings than those derived from immediate ratings. It was reasoned that the higher
correlations in the retrospective (summative) ratings must have resulted from "logical presuppositions in the mind of the rater, rather than from actual behavior" (Newcomb, 1931). One of the latest replications of the Newcomb study, using similar comparisons between immediate and retrospective ratings, has further confirmed the possibilities of halo effect rating bias (D'Andrade, 1974).

The measurement of performance in the work place has continued to occupy the attention of industrial psychologists since the preliminary efforts of Taylor, Wells, Thorndike, Rugg, and Hawthorne near the beginning of the twentieth century. Unfortunately, recognizing the importance of performance measurement and actually measuring this performance accurately are two different matters. An ideal system of performance evaluation would include the combination of objective, personal, and judgmental indices (Landy and Trumbo, 1981).

Studies conducted at Ohio State University during World War II identified logical and recurring patterns in the rating process in the military. It was determined that these components included observation, storage, retrieval, and judgment (Wherry, 1952).

As reported in Personnel Psychology between 1950 and 1955, 81 percent of the published studies used judgmental indices of one type or another to rate performance (Guion, 1965). Similar findings for the period of time between 1965 and 1975 were reported in the Journal of Applied Psychology. During that period 72 percent of the published studies revealed judgmental criteria as the primary form of performance measurement (Landy and Trumbo, 1981).
In spite of the widespread use of judgmental criteria for measuring performance, there remains a constant dissatisfaction with these measures. The source of this dissatisfaction has been the vulnerability of these measures to both intentional and inadvertent bias (Landy and Farr, 1980).

While some employees feel that personnel decisions should be based primarily on seniority (this fundamental belief is borne out in typical educational unit collective bargaining practices), most believe that good performance should be recognized and rewarded. Despite the abundance of empirical research devoted to specific aspects of the evaluation process in the private sector, it has remained one of the most neglected areas in all of human resource management. While over 90 percent of organizations report some kind of appraisal system, the majority also report dissatisfaction with the process (Locher and Teel, 1977).

Technical Aspects of Teacher Performance Evaluation

The process of evaluating teacher performance has evolved dramatically since the middle of the 1960s. Before this time much of the investigation attempted to rate characteristics or describe certain traits or attitudes which appeared to be attributable to "good" or "poor" teachers (Ryans, 1960). Much of the research was predicated on a fairly high degree of inference. Nevertheless, these earliest studies were significant and heralded the beginning of an era of research that attempted to isolate teaching behaviors that could be associated with improved student achievement.
The new approach that was introduced later in the 1960s concentrated on identifying "low-inference" behaviors which were recorded as they occurred. This approach was based on quantifiable objective measurements of specific behaviors (Hosford and Martin, 1980). Researchers associated with the School Improvement Model (SIM) at the College of Education at Iowa State University, had begun creating generic job descriptions for teachers and were identifying many potential performance criteria. As these criteria were considered, they were screened for validity, reliability, and discrimination power. Research on recognized effective teaching was used to validate each criterion (Manatt and Stow, 1984). Reliability, or consistency of interrater and individual rating, was sought for each criterion (Manatt, Palmer, and Hidlebaugh, 1976). The ability to identify high, average, and low performance on each criterion, or discrimination power, was also deemed important.

This period of educational research associated with the evaluation of teaching performance was clearly interested in the relationship of classroom behaviors and verbal interactions. Verbal interactions observation systems of various types that were developed within this ten-year period included an Observation and Record (OScAR), created by Medley and Mitzel in 1959, Flanders Interaction Analysis system, introduced in 1965, and a variety of observational techniques to use to collect data that were contained in the anthology Mirrors for the Classroom (Simon and Boyer, 1970).

The 1970s could generally be classified as the Process-Product time of investigation. Correlational studies primarily examined the relationship
between a teacher behavior criterion or variable and student achievement scores (Hosford and Martin, 1980).

The focus of recent research has centered on what components should make up a comprehensive and successful teacher performance evaluation system and what is necessary to support teacher and teaching growth (Manatt and Stow, 1986; Conley, 1987; Duke and Stiggins, 1986; McGreal, 1984). According to Stanley and Popham, 1988, it appears that an evaluation system is more likely to encourage teaching growth if it:

1. Includes clear criteria, established with significant stakeholder involvement, that reflect the school organization’s framework for looking at and talking about teaching.
2. Provides opportunities for increased teacher involvement within the actual functioning of the system.
3. Provides opportunities to use multiple sources of data to ensure the fullest possible picture of teaching.
4. Allows and encourages feedback activities that have been shown to encourage professional growth (p. 13).

A major premise that is clearly indicated and reinforced in reviews of teacher performance evaluation literature is that differences between effective and ineffective teachers are often associated with their interpersonal communication skills. Judgments of these skills have always been included within systems of performance evaluations, with the major problem being that of objectivity—that is, in clearly demonstrating that these judgments are free from subjective influence or bias (Hosford and Martin, 1980).
Bias in Performance Evaluation

While industrial psychologists have been investigating major components of systems for performance appraisal since the beginning of the twentieth century (Landy, Zedeck, and Cleveland, 1983) performance evaluation in the educational setting has become a concentrated focus for research and study much more recently. As the body of knowledge regarding the evaluation of performance in the field of education grows, those studies that focus on the technical aspects of systems of performance evaluation serve to reinforce the conclusions drawn from earlier studies conducted by investigators in the private sector.

Any system used to evaluate performance must rely, to some degree, on subjective interpretation. Richard Henderson (1984) summarized ratee concerns about this subjectivity:

Possibly the most critical barriers to accurate and valid measurement of employee performance lie deep within the genetic and learned makeup of all people. A wide variety of emotional, psychological, intellectual, and physical problems that, at first glance, may appear to be separate and irrelevant factors, may combine in any number of different ways during the performance appraisal process to completely neutralize or lay waste to any program designed to measure employee performance (p. 2).

Henderson goes on to explain how employees fear this subjective interpretation might be manifested during the process of evaluation:

Almost all employees are extremely wary of performance ratings.... This fear of performance ratings is reinforced by many actions and failures to act by immediate supervisors, administrators, and managers at all levels. Possibly the most common fear expressed by a ratee is that of rater subjectivity. What worries the ratee is that the rater will not measure his or her performance on the actual behaviors demonstrated and results achieved during the rating period, but will instead use a variety of subjective biases to rate performance (p. 3).
Henderson concludes with this concern about the organizational "stamp of approval" that is associated with a formal system of performance evaluation:

The final fear that many ratees have of any formal performance appraisal process is that the well written, precise operating policies, and the well designed measurement instrument provide an aura of objectivity, while actually permitting raters and the organization to operate in a subjective, highly biased manner (p. 6).

Biases that may influence one's evaluation of another's performance, and are the source of much of the concern for ratees, threaten the validity of performance evaluation instrumentation that the organization so carefully designs (Bernardin and Beatty, 1984). Kingsbury (1922) was apparently the first researcher to list the three sources of rater error that are most frequently discussed in recent literature (viz., halo effect, central tendency, and leniency).

**Halo effect**

Halo effect has been defined as a tendency to attend to a global impression from one aspect of a person's performance rather than to distinguish carefully among performance levels for specific criteria. As a consequence, the evaluator may fail to discriminate among conceptually different and possibly independent components of an employee's behavior. Halo effect tends to be used to describe the influence a general positive impression for one aspect of performance has on specific performance criteria, hence the reference to "halo." "After the rater has cast his halo around his subject, he is so dazzled by its radiance that he cannot
differentiate the subject’s separate qualities" (Johnson and Vidulich, 1956). Closely related is the horn effect (the tendency of an unsatisfactory rating on one aspect to influence the rater to give a person a similar rating or lower-than-deserved rating on other specific criteria) (Landy and Farr, 1983; Henderson, 1984).

Central tendency

Central tendency is characterized as the avoidance of extreme (favorable or unfavorable) ratings or a tendency to cluster ratings at or near the scale midpoint. This is the most common and serious kind of error. Since many employees do perform within the "average" range, it is an easily rationalized alternative to making a valid evaluation of performance. Central tendency error lowers the ratings of the above-average employee and raises the ratings of poorer performers, which can lead to lower organizational performance (Baker, 1989; Henderson, 1984).

Leniency/severity

Leniency bias is evident when the rater assigns all employees high performance ratings and all ratings cluster at top levels on the measurement instrument. The conceptual opposite of leniency bias is severity. Severity bias is characterized by rating consistently lower than the normal or average or when all ratings cluster at the lower levels on the measurement instrument (Eichel and Bender, 1984; Henderson, 1984).

Since halo effect, central tendency, and leniency were identified as potential sources of bias, several additional biases have been studied and
their effects on the evaluation of performance analyzed. Demographic factors that have been investigated include gender, race, age, and level of education. Psychological factors include a variety of personality and cognitive variables. Job-related factors include experience, level of performance of the rater, leadership style, and proximity of the rater to the ratee (Landy and Farr, 1980; Bernardin and Beatty, 1984).

Results of studies that have been published since 1970 have not shown any consistent effect of gender on performance evaluation (Centra and Linn, 1973; Elmore and LaPointe, 1974, 1975; Schmitt and Lappin, 1980). In more recent studies of the relationship between gender and performance evaluation in the educational setting, gender was concluded to be a main effect. Females were found to be tougher raters than males (Harrington, 1984; Peterson, 1988).

Possible effect of racial bias was the subject of investigations conducted by the U.S. Civil Service Commission. While no clear relationship between racial bias and performance evaluation was determined, in a majority of cases, raters did give higher ratings to ratees of their own race than to ratees of a different race (Crooks, 1972).

Investigations conducted to examine the effect of the age of the rater on performance evaluation have been inconclusive. While some found that younger supervisors were less lenient in their ratings (Mandell, 1956), others found no effect on performance ratings (Klores, 1966). Rater education has been associated with an effect on supervisory ratings of the job performance of police officers. The effect accounted for such a small
percentage of the total rating variance, however, that they concluded that level of education was of no practical significance (Cascio and Valenzi, 1977).

Surgency Bias

Surgency bias is defined as a distortion, predilection, or influence in perception or rating due to physical attractiveness and charisma of the ratee. Studies of physical attractiveness conducted by industrial and social psychologists have consistently shown that people rated as attractive are generally treated better than unattractive people (Dion, Berscheid, and Walster, 1972). Additionally, attractive people are assigned more positive interpersonal attributes, such as intelligence, happiness, and extroversion (Miller, 1970; Dion, 1972) and are predicted to be more successful in their personal and professional lives (Berscheid and Walster, 1974). The employment potential for both attractive males and females was rated higher than that of unattractive applicants, and attractive applicants were rated more qualified than unattractive applicants for jobs stereotyped as consistent with applicants' gender (Cash, Gillen, and Burns, 1977).

There is, however, an increasing amount of more recent evidence that people discriminate against the most attractive members of their own gender (Krebs and Adinolfi, 1975). When Dermer and Thiel (1975) had female subjects assign characteristics to other females who varied in physical attractiveness, they found that women attributed a variety of socially undesirable characteristics to more attractive members of their
own gender. These results suggest that an attractiveness horn effect operates if members of the same gender judge the work of one another.

In summarizing their study of the effects of facial attractiveness and gender, completed in 1986, Spencer and Taylor concluded:

While it may be true that people have higher expectations of success for attractive individuals, they may base their expectations on false premises. Attractive women of high ability may be assumed to have gained external assistance, and attractive men who perform well may be viewed as exerting little effort. When attractive people fail to meet performance standards, that failure is attributed to internal factors and is treated more harshly than that of others. Hidden prejudices may greatly complicate the work life of attractive people, who may appear to be performing at less than their capabilities, when in fact they are working at the limit of their talents and endowments (p. 282).

Heilman and Stopec (1985) concluded their study of physical attractiveness and gendertyping of female jobs with this advice, "women should strive to appear as unattractive as possible if they are to succeed in advancing their careers."

Summary

Teacher performance evaluation remains one of the most sensitive and emotional endeavors that is shared by teachers and administrators. In this new age of responsibility and accountability, it is evident that instrumentation and methodology that are sound and well-grounded in research are available. The aspect of the entire performance evaluation process that must be examined and improved is related to the subtle and not-so-subtle elements that influence evaluators as they make judgments. Bias has always existed. It is critical that the various forms of bias that interfere with accurate assessment of performance be identified and addressed through training.
Surgency as a source of bias has been studied extensively by both personnel and social psychologists. In many of the studies conducted in the early to middle 1970s, surgent people were found to be generally evaluated and treated better than their nonsurgent colleagues (Berscheid and Walster, 1972; Spence and Helmreich, 1972; Dion, 1972; Clifford and Walster, 1973; Landy and Sigall, 1974). Findings in more recent studies, conducted since 1975, are beginning to suggest that in some situations, surgency may not be significantly associated with any bias related to evaluation (Dipboye, 1977), or may be associated negatively with performance evaluation (Krebs and Aldofini, 1975; Kaplan, 1978; Heilman and Stopeć, 1985; Spencer and Taylor, 1986). A summary of previous research studies investigating rater characteristics as sources of bias in the evaluation of performance is shown in Table 1.
Table 1. Research studies investigating rater characteristics as sources of bias in the evaluation of performance

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Study</th>
<th>Direction of bias</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender</td>
<td>Cash et al., 1977</td>
<td>+</td>
</tr>
<tr>
<td></td>
<td>Centra and Linn, 1973</td>
<td>+</td>
</tr>
<tr>
<td></td>
<td>Crooks, 1972</td>
<td>+</td>
</tr>
<tr>
<td></td>
<td>Dipboye et al., 1977</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>Elmore and LaPointe, 1974</td>
<td>M</td>
</tr>
<tr>
<td></td>
<td>Elmore and LaPointe, 1975</td>
<td>M</td>
</tr>
<tr>
<td></td>
<td>Peck, 1978</td>
<td>+</td>
</tr>
<tr>
<td></td>
<td>Peterson, 1988</td>
<td>+</td>
</tr>
<tr>
<td></td>
<td>Schmitt and Lappin, 1980</td>
<td>M</td>
</tr>
<tr>
<td></td>
<td>Spence and Helmreich, 1972</td>
<td>+</td>
</tr>
<tr>
<td></td>
<td>Terborg and Shingledecker, 1983</td>
<td>M</td>
</tr>
<tr>
<td></td>
<td>Wexley and Palukos, 1983</td>
<td>+</td>
</tr>
<tr>
<td>Race</td>
<td>Cox and Glick, 1986</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>Crooks, 1972</td>
<td>M</td>
</tr>
<tr>
<td></td>
<td>Peterson, 1988</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>Schmitt and Lappin, 1980</td>
<td>M</td>
</tr>
<tr>
<td>Physical appearance</td>
<td>Alicke et al., 1968</td>
<td>M</td>
</tr>
<tr>
<td></td>
<td>Beehr and Gilmore, 1982</td>
<td>M</td>
</tr>
<tr>
<td></td>
<td>Berman et al., 1981</td>
<td>M</td>
</tr>
<tr>
<td></td>
<td>Berscheid and Walster, 1972</td>
<td>+</td>
</tr>
<tr>
<td></td>
<td>Berscheid and Walster, 1974</td>
<td>+</td>
</tr>
<tr>
<td></td>
<td>Cash et al., 1977</td>
<td>+</td>
</tr>
<tr>
<td></td>
<td>Clifford and Walster, 1973</td>
<td>+</td>
</tr>
<tr>
<td></td>
<td>Cox and Glick, 1986</td>
<td>M</td>
</tr>
<tr>
<td></td>
<td>Dermer and Theil, 1975</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>Dickey-Bryant et al., 1986</td>
<td>+</td>
</tr>
<tr>
<td></td>
<td>Dion, 1972</td>
<td>+</td>
</tr>
<tr>
<td></td>
<td>Dipboye et al., 1977</td>
<td>O</td>
</tr>
<tr>
<td></td>
<td>Heilman and Stopec, 1985</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>Hosford and Martin, 1980</td>
<td>O</td>
</tr>
<tr>
<td></td>
<td>Janda et al., 1981</td>
<td>+</td>
</tr>
<tr>
<td></td>
<td>Kaplan, 1978</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>Krebs and Aldofini, 1975</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>Landy and Sigall, 1974</td>
<td>+</td>
</tr>
<tr>
<td></td>
<td>Michalini et al., 1981</td>
<td>+</td>
</tr>
<tr>
<td></td>
<td>Miller, 1970</td>
<td>M</td>
</tr>
<tr>
<td></td>
<td>Peck, 1978</td>
<td>M</td>
</tr>
<tr>
<td></td>
<td>Spence and Helmreich, 1972</td>
<td>+</td>
</tr>
<tr>
<td></td>
<td>Wexley and Palukos, 1983</td>
<td>+</td>
</tr>
<tr>
<td></td>
<td>Wood and Mitchell, 1981</td>
<td>M</td>
</tr>
</tbody>
</table>

*aKey:  + = Helps rating;  - = Hurts rating;  O = No significant difference;  M = Mixed results.*
CHAPTER III. METHODS AND PROCEDURES

Introduction

The purpose of this study was to determine whether surgency bias affects an evaluator's rating of another's performance. To determine surgency bias effects, a demographic data collection instrument, teacher performance rating instrument, preobservation conference videotapes, and a script tape of a lesson and additional supplemental materials were developed.

As this study was conceptualized and the procedures and methodology for the investigation were developed, several previous research projects were reviewed. The methodology that served as a prototype came from a study of anticipatory set bias and its relationship to teacher performance evaluation, conducted by Bourisaw, in 1988. Once preliminary investigation procedures were created, a pilot test was administered to a group of graduate level educational administration students. The pilot test group completed the entire series of activities and critiqued all instruments, instructions, and supplemental materials. After debriefing them and collecting their ideas, suggestions, and recommendations, final documents and supplemental materials were created.

Experimental sites were selected and subjects attending workshops at these sites were assigned to experimental groups. Data were collected from participants in San Antonio, Texas, and Salina, Kansas.

This surgency study research design was a quasi-experimental post-test only two-group design, depicted by the following diagram (Borg and Gall, 1983):
Subjects were assigned to one of two experimental groups (surgent and nonsurgent teacher performance evaluation rater groups A and B). Following a treatment (observation of one of two videotaped preobservation conferences), all subjects completed a written evaluation of the teacher's performance.

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

**Materials**

The Teacher Performance Rating Scale (Appendix A) used in this study was developed to determine if surgency bias exists in the rating of a teacher's performance. It was used by subjects to evaluate a lesson. The behaviors included on the instrument describe effective teaching behaviors based on the Hunter Model (Hunter, 1984). Additional evaluation items provided for the teacher evaluator to rate the teacher's overall performance on the lesson and to select a Professional Improvement Commitment (PIC) for the teacher to improve performance.

A five-point scale was used for this study:

1 - Level 1    Performance is highly unacceptable. The teacher does virtually nothing of educational value related to the criterion.
2 - Level 2  Performance is not at an acceptable level. The teacher is deficient enough to be ineffective for this criterion.

3 - Level 3  Performance is acceptable. The teacher demonstrates adequate skill for this criterion.

4 - Level 4  Performance is of high quality. The teacher is above average for this criterion, but not good enough to serve as a model for others.

5 - Level 5  Performance could serve as a model for other teachers. The teacher demonstrates a high proficiency for this criterion.

Each of the specific teacher behaviors was a validated criterion on the *Teacher Performance Rating Scale* (Manatt, 1985). All criteria were intended to be generic and were designed to be used for all subjects, grade levels, and genders. This investigation postulated that the difference in surgency portrayed in the videotaped preobservation conferences would result in significantly different ratings on predicted teacher effectiveness, specific teacher behaviors, overall performance, and PIC selection.

Item 11 on the *Teacher Performance Rating Scale* was "Overall Rating." This item asked evaluators to rate the teacher's overall performance in the lesson.

The final section of the *Teacher Performance Rating Scale* was a menu of Professional Improvement Commitments (PICs). PICs were carefully designed so that each succeeding one would be more challenging and difficult for the teacher observed during the videotaped preobservation conference to complete.
The ten specific teacher behaviors that comprise the Teacher Performance Rating Scale were originally developed for utilization in evaluating a videotaped physical science lesson, taught by Judy Oliver, and created for the Association for Supervision and Curriculum Development in 1987. A panel of four experts, skilled in teacher evaluation, helped to develop and validate the Teacher Performance Rating Scale. They were asked to provide feedback on the following elements of the instrument: clarity of directions, clarity and specificity of criteria, and the validity of the criteria. The teacher behaviors were strategies or techniques commonly accepted as reflective of effective teaching and were derived from the research on effective teaching (Hunter, 1984). One additional item was provided for the evaluators to make an overall rating of the lesson.

The Personal Data Form (Appendix B) used in this study was developed to collect demographic information from subjects that could be included with ratings on the Teacher Performance Rating Scale for data analyses. Since many of the original research questions posed at the onset of this study centered around possible surgency biases associated with gender, age, amount of teacher performance evaluation training, and actual administrative experience, these characteristics are included on the Personal Data Form. Because of the probability that many of the subjects in this study might hold more than one specific administrative position, more than one position could be marked. After subjects had completed the demographic portion of the Personal Data Form, they observed one of the preobservation conference videotapes. Immediately following this viewing,
they were asked to make a prediction about the perceived teaching effectiveness of the teacher they observed. The last item on the Personal Data Form includes definitions of the levels of performance and a scale for subjects to use to make this prediction.

To conduct this study it was necessary to design a lesson and select actresses to portray surgent and nonsurgent teachers for videotaping. Since subjects for the study would hold a wide grade level range of administrative positions, a ninth grade physical science lesson was selected. The verbatim written script tape of the Judy Oliver physical science lesson, created originally for the Association for Supervision and Curriculum Development (Manatt, 1985), became the document for subjects to refer to as they completed evaluations on the Teacher Performance Rating Scale.

Two actresses were chosen to portray Judy Oliver, the physical science teacher. The actress selected to assume the role of the surgent Judy Oliver is a professional actress/performer who is employed by a major television station to serve as a hostess/moderator. She has had considerable acting experience and is comfortable being videotaped. As the surgent teacher, she was required to be more physically attractive, use clear and assertive speaking skills, be poised, and display expressive nonverbal communications mannerisms. The actress selected to portray the nonsurgent Judy Oliver was required to be less physically attractive and exude much less poise and expressiveness in her communications efforts. Both actresses were approximately the same age. Permission was secured from the actresses to participate in this study (Appendix C).
Several discussions with members of the local district science department, science department chairperson, the actresses who would portray Judy Oliver, and experts in teacher performance evaluation were held to create the preobservation conference script and to review the plan for videotaping. The revised preobservation conference script was memorized by both actresses, so that their responses to preobservation conference questions were identical. Both videotaped preobservation conferences were of the same approximate length.

The location for the actual videotaping was in the local district superintendent's office and local district videotaping equipment was used. The district audio-visual director videotaped the preobservation conferences.

Subjects used

Ames, Iowa: This graduate level class, which served as the pilot test group, contained ten students, five female and five male. Years of experience as teacher performance evaluators ranged from one to 12, with an average of 5.8 years. Days spent in teacher performance evaluation training ranged from six to 18, with an average of 9.7 days.

Salina, Kansas: Eighty-nine teachers and administrators participated in a professional conference on "Effective Teacher Evaluation" November 29-30, 1989, February 27-28, 1990, and March 28, 1990, conducted by Dick Manatt, at this site; 25 were female and 64 were male. Forty-five of the participants were assigned to the nonsurgent teacher performance rater Group A and 44 were assigned to surrogate teacher performance rater Group B.
Years of experience as teacher performance evaluators ranged from zero to 30, with an average of 7.08 years. Days of training in teacher performance evaluation ranged from zero to 37 days, with an average of 10.42 days.

San Antonio, Texas: Thirty-two subjects attended this portion of a "Supervising the Marginal Teacher" session, conducted by Dick Manatt, at the Association for Supervision and Curriculum Development national convention in early February 1990. Fourteen of the participants were assigned to the nonsurgent teacher performance rater Group A and 18 subjects were assigned to surgent teacher performance rater Group B. Years of experience as teacher performance evaluators ranged from zero to 27, with an average of 9.3 years. Days of training in teacher performance evaluation ranged from zero to 37, with an average of 14.59 days.

Conducting the experiment

During February 1990 a field test was completed. The purpose of this field test was to further develop and validate the Personal Data Form and Teacher Performance Rating Scale, and confirm the level of effectiveness of the videotaped preobservation conferences and physical science lesson script tape. Ten members attended a graduate level educational administration teacher performance evaluation course, taught by Manatt. The ten students had varying levels of education and had received considerable training and experience in teacher performance evaluation, including workshops, staff development activities, and course work. During previous periods of instruction, they had received instruction on
the process of teacher performance evaluation, including a discussion on the criteria for effective teaching.

After some introductory comments by Professor Manatt, the ten students were assigned to one of two groups. Nonsurgent teacher performance rater Group A would observe one preobservation conference and predict the effectiveness of the teacher observed in this conference. Surgent teacher performance rater Group B would move to a different location, observe a different preobservation conference, and predict the effectiveness of the teacher observed in this conference. Prior to the students' viewing the videotaped preobservation conference, the Personal Data Form and the Teacher Performance Rating Scale were distributed to each student (Appendices A and B). Groups A and B then reconvened in the classroom, were given script tapes of the Judy Oliver physical science lesson and a Teacher Performance Rating Scale, and were asked to rate her performance.

Following the completion of the planned activities used to administer the surgency pilot test, participants were debriefed and asked to indicate errors in associated documents and procedures that needed to be modified or clarified. These suggestions and recommendations were considered and incorporated in the final documents and administration procedures.

Data were collected at two teacher performance evaluation workshops conducted by Professor Manatt during the months of February and March 1990. The workshop sessions ranged in length from one to three days lasting six and one-half to seven hours per day. For a complete workshop schedule refer to Appendix D. After completing a minimum of three hours of training, Manatt spent approximately 30 minutes reviewing the
components of effective instruction using the Hunter Model (Hunter, 1984). This review was followed by brief comments related to factors that might influence rating of performance. Next, participants were told the group was to be divided into two smaller groups in order to do a training exercise which would provide feedback and insights regarding their own behavior as teacher evaluators.

Participants were instructed to complete the demographic portion of the color-coded (yellow or pink) **Personal Data Form** (see Appendix B). The color-coded forms had been inserted alternately into workshop packets of materials that were distributed at the beginning of each workshop session. **Personal Data Forms** had been sequentially numbered prior to the workshops, and participants matched their **Personal Data Forms** and **Teacher Performance Rating Scales** by recording the number printed in the appropriate space on one to the other. Participants were assigned to one of the two experimental groups on the basis of the color of their **Personal Data Forms** (yellow or pink). Nonsurgent teacher performance rater Group A participants viewed the videotaped preobservation conference in a room separate from those in surgent teacher performance rater Group B.

Following the viewing of one of the videotaped preobservation conferences and marking the prediction of the teacher's classroom effectiveness, all participants returned to the original presentation room. There they were given a color-matched **Teacher Performance Rating Scale** and a Judy Oliver Surgency Study packet (Appendix E) containing a physical science instructional plan, student group description, preobservation conference data form (with teacher responses to specific
questions summarized), a lesson timeline (script tape), a graphic response mode for each of the ten teacher behaviors found on the Teacher Performance Rating Scale, and a menu of five Professional Improvement Commitments (PICs).

Each participant was asked to rate the teacher's performance independently, using the script tape of the physical science lesson and associated supplemental information and responding to the graphic response mode evaluation instrument. In addition to rating each of the ten specific teacher behaviors, participants made an overall rating of the teacher's performance and selected a Professional Improvement Commitment (PIC) for the teacher to consider to improve her teaching performance. There was no discussion during or after viewing the videotape and participants were given as much time as was necessary to complete the instrument. After a short break participants were debriefed on the occurrence of surgency bias during the evaluation process and were permitted to observe portions of both the surgency and nonsurgency preobservation conference videotapes.

Data Analysis Procedures

Data for statistical analysis were obtained from the Teacher Performance Rating Scale and the Personal Data Form. The data were analyzed using programs written specifically for this experiment by Alice Cheng, statistical programmer at Iowa State University, in Ames, Iowa. Descriptive statistics, t-tests for unmatched pairs, one-way analysis of variance, and two-way analysis of variance were used. The t-test for
unmatched pairs was used to determine if differences existed in predictions of surgent and nonsurgent teacher performance, overall ratings, and Professional Improvement Commitment (PIC) selections. T-tests were also used to measure the significance of the difference in the amount of teacher performance evaluation training and experience existed between members of the surgent and nonsurgent teacher performance rater groups. In addition, t-tests were calculated to determine whether differences existed between ratings made by men and women subjects. One-way analysis of variance (ANOVA) and two-way analysis of variance tests were used to determine differences in rating associated with age, grade level of administrative assignment, and the administrative position held.
CHAPTER IV. ANALYSES AND RESEARCH FINDINGS

The primary purpose of this study was to analyze predictions of teacher effectiveness, teacher performance evaluation ratings, and Professional Improvement Commitment (PIC) selection, from two groups participating in teacher performance evaluation training to determine whether surgency bias influences performance rating. Additional purposes of the study included determining the effects of teacher performance evaluation training and teacher performance evaluation experience on surgency bias. Section one of this chapter provides description data and section two summarizes the results of hypotheses testing.

Descriptive Data

Descriptive data, presented in Figures 1 through 6, depict important evaluator characteristics relevant to this study: gender of subjects assigned to surgent and nonsurgent teacher performance evaluation rater groups, the range of ages of subjects in the study, administrative positions held by subjects, and their general grade level administrative assignments. Figure 1 graphically illustrates the assignment of approximately equal numbers of females to the surgent (18) and the nonsurgent (21) teacher performance evaluation rater groups and males to the surgent (45) and the nonsurgent (37) teacher performance evaluation rater groups. The total study sample of 121 was separated into the two experiment sites, San Antonio and Salina, to show the assignment to surgent and nonsurgent teacher performance evaluation rater groups by gender.
As shown in Figure 2, in San Antonio, eight females were assigned to the surgent, six females to the nonsurgent, ten males to the surgent, and eight males were assigned to the nonsurgent teacher performance evaluation rater groups. Figure 3 shows that, in Salina, ten females were assigned to the surgent, 15 females to the nonsurgent, 35 males to the surgent, and 29 males to the nonsurgent teacher performance evaluation rater groups.

Subjects were asked to indicate their age on the Personal Data Form by circling the five-year bracket that contained their age. Eighty-eight participants, or nearly three-fourths of the total number, were between the ages of 36 and 50. A representation of the age categories is depicted in Figure 4. More than half of the participants were of the same approximate age as the actresses who portrayed the teachers.

Studies conducted by industrial psychologists have shown that younger evaluators tend to rate performance less leniently than do their older
Figure 2. Percentages of subjects at San Antonio

Figure 3. Percentages of subjects at Salina
counterparts. As shown by the F-value of .46 in Table 2, that tendency was not borne out in this study. The associated probability of .63 clearly indicates that the age of an evaluator in this study makes no significant difference in the rating of performance. While the average overall rating by subjects in the youngest age bracket, 31-40, is the lowest, 3.31, of the three brackets, it is clearly not different enough to be significant.

As shown in Figure 5, 83 subjects, or more than two-thirds, held the position of building administrator at the time of this study. The remaining one-third of the participants held positions of superintendent, central office administrator, teacher, or some other combination of these positions.

Of the 83 subjects who currently held building level administrative positions, more than half, 47, served in an elementary or middle grade
Table 2. ANOVA analysis and means of the differences in overall rating of performance by evaluators grouped into ten-year age brackets

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>S.S.</th>
<th>M.S.</th>
<th>F-value</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>2</td>
<td>.5692</td>
<td>.2846</td>
<td>0.46</td>
<td>0.6321</td>
</tr>
<tr>
<td>Error</td>
<td>118</td>
<td>72.9394</td>
<td>.6181</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected total</td>
<td>120</td>
<td>73.5041</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Mean ratings by age brackets

<table>
<thead>
<tr>
<th>Age</th>
<th>N</th>
<th>Mean overall rating</th>
</tr>
</thead>
<tbody>
<tr>
<td>31-40</td>
<td>35</td>
<td>3.3143</td>
</tr>
<tr>
<td>41-50</td>
<td>64</td>
<td>3.4688</td>
</tr>
<tr>
<td>51-60</td>
<td>22</td>
<td>3.4545</td>
</tr>
</tbody>
</table>

level assignment. Twenty-five subjects held senior high grade level positions and 11 had building administrator responsibilities that included kindergarten through grade 12 or some other combination of grade levels of responsibility. Grade level assignments are illustrated in Figure 6.

Since subjects for this study had attended one of two different workshop sites, a t-test was conducted to determine whether any significant differences could be shown to relate to location. As shown in Table 3, a t-test analysis for unmatched pairs on the average overall rating by subjects at both sites revealed that no significant difference
Figure 5. Percentages of subjects by position

- Building Administrators: 68.6%
- District Administrators: 10.7%
- Teachers: 5.0%
- Bldg. Admin. and Perform. Evaluators: 5.0%
- Other Admin. Assignment: 10.7%

Figure 6. Percentages of subjects by grade level assignment

- Elementary: 41.0%
- Middle Level: 15.7%
- Senior High: 30.1%
- Other Grade Level Assignments: 13.3%
Table 3. T-test analysis of overall ratings of performance between evaluators in Salina and San Antonio

<table>
<thead>
<tr>
<th>Location</th>
<th>N</th>
<th>Mean</th>
<th>S.D.</th>
<th>DF</th>
<th>t-value</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Salina</td>
<td>89</td>
<td>3.3933</td>
<td>.8342</td>
<td>119.0</td>
<td>-.6601</td>
<td>.5105</td>
</tr>
<tr>
<td>San Antonio</td>
<td>32</td>
<td>3.5000</td>
<td>.6222</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

occurred. The t-value of -.66 and the associated probability of .51 indicate that location of training accounted for no significant difference in average ratings of overall performance.

Table 4 depicts the results of ANOVA on the average Professional Improvement Commitment (PIC) selected when subjects were categorized by group, location, and the interaction of these two factors. This shows no significant difference on PICs between groups (F=.24, associated probability=.6248), between locations (F=1.15, associated probability=.2853), and the interaction of these two factors (F=.61, associated probability=.4362).

In addition to a concern that location might have some effect on this study of surgency bias, other extraneous factors with the potential to have an impact on ratings of performance were considered. T-tests for unmatched pairs were conducted to determine whether the amount of teacher performance evaluation training (Table 5) and teacher performance evaluation experience (Table 6) for subjects in the surgent group and nonsurgent group were significantly different. When subjects were assigned to one of two experimental groups at the study sites, measures
Table 4. ANOVA and means of the differences of the Professional Improvement Commitment (PIC) selection between subjects in the surgent and nonsurgent teacher performance evaluation rater groups in Salina and San Antonio

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>S.S.</th>
<th>M.S.</th>
<th>F-value</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>3</td>
<td>1.9287</td>
<td>.6429</td>
<td>.67</td>
<td>.5763</td>
</tr>
<tr>
<td>Group</td>
<td>1</td>
<td>.2316</td>
<td>.24</td>
<td>.6248</td>
<td></td>
</tr>
<tr>
<td>Location</td>
<td>1</td>
<td>1.1093</td>
<td>1.15</td>
<td>.2853</td>
<td></td>
</tr>
<tr>
<td>Group*Location</td>
<td>1</td>
<td>.5877</td>
<td>.61</td>
<td>.4362</td>
<td></td>
</tr>
<tr>
<td>Error</td>
<td>117</td>
<td>112.6664</td>
<td>.9630</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected total</td>
<td>120</td>
<td>114.5950</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Mean PIC selections by group and by location

<table>
<thead>
<tr>
<th>Group</th>
<th>Location</th>
<th>N</th>
<th>Mean PIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Surgent</td>
<td>Salina</td>
<td>49</td>
<td>3.1020</td>
</tr>
<tr>
<td>Surgent</td>
<td>San Antonio</td>
<td>14</td>
<td>2.7143</td>
</tr>
<tr>
<td>Nonsurgent</td>
<td>Salina</td>
<td>40</td>
<td>3.1250</td>
</tr>
<tr>
<td>Nonsurgent</td>
<td>San Antonio</td>
<td>18</td>
<td>3.0556</td>
</tr>
<tr>
<td></td>
<td></td>
<td>121</td>
<td>3.0579</td>
</tr>
</tbody>
</table>

were taken to balance gender between groups. As shown in Table 5, a t-value of 2.69 and the associated probability of .008 indicates that the average number of days of teacher performance evaluation training for members of the surgent group was significantly less than the average number of days of training for members of the nonsurgent group. While the t-test analysis indicated that the difference in amount of training is significantly different, the actual difference in amount of training is equal to less than one-half day.
Table 5. T-test analysis of difference in the average number of days of teacher performance evaluation training between subjects in the surgent and nonsurgent teacher performance evaluation rater groups

<table>
<thead>
<tr>
<th>Group</th>
<th>N</th>
<th>Mean</th>
<th>S.D.</th>
<th>DF</th>
<th>t-value</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Surgent</td>
<td>63</td>
<td>2.3651</td>
<td>.9722</td>
<td>119</td>
<td>2.6911</td>
<td>.008**</td>
</tr>
<tr>
<td>Nonsurgent</td>
<td>58</td>
<td>2.8448</td>
<td>.9877</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

** Equals <.01.

A t-test for unmatched pairs conducted to determine the difference in the average number of years of teacher performance evaluation experience for members of the surgent group and the nonsurgent group, shown in Table 6, disclosed no significant difference. This is indicated by the t-value of .24 and the associated probability of .81.

Table 6. T-test analysis of the difference in the average number of years of teacher performance evaluation experience between subjects in the surgent and nonsurgent teacher performance evaluation rater groups

<table>
<thead>
<tr>
<th>Group</th>
<th>N</th>
<th>Mean</th>
<th>S.D.</th>
<th>DF</th>
<th>t-value</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Surgent</td>
<td>63</td>
<td>2.4444</td>
<td>1.0892</td>
<td>119</td>
<td>.2381</td>
<td>.8122</td>
</tr>
<tr>
<td>Nonsurgent</td>
<td>58</td>
<td>2.3966</td>
<td>1.1230</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Research Hypothesis Testing

Research questions postulated as this study was developed each resulted in a specific research hypothesis. All hypotheses were tested for significance at the .05 level, with probabilities at the .01 level also reported.

Hypothesis 1: There is no significant difference in the average prediction of teaching effectiveness of a surgent teacher and a nonsurgent teacher, after observing these teachers in a videotaped preobservation conference.

The purpose of this hypothesis was to determine whether an evaluator's immediate reaction, following the observation of a videotaped preobservation conference, would reflect surgency bias. The average prediction of teacher effectiveness on the Personal Data Form was used to test this hypothesis. As shown in Table 7, using the t-test for unmatched pairs, a highly significant t-value of 6.66 and an associated probability of .0001 was obtained.

Table 7. T-test analysis of the difference in the average prediction of teacher effectiveness between subjects in surgent and nonsurgent teacher performance evaluation rater groups

<table>
<thead>
<tr>
<th>Group</th>
<th>N</th>
<th>Mean</th>
<th>S.D.</th>
<th>DF</th>
<th>t-value</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Surgent</td>
<td>63</td>
<td>4.3810</td>
<td>.6822</td>
<td>119</td>
<td>6.6611</td>
<td>.0001**</td>
</tr>
<tr>
<td>Nonsurgent</td>
<td>58</td>
<td>3.4310</td>
<td>.8808</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Equals <.01.
On the basis of this test, the null hypothesis was rejected. Evaluators who observed the videotaped preobservation conference of a surgent teacher predicted that she would be more effective as a teacher than those subjects who observed the videotaped preobservation conference of a nonsurgent teacher.

**Hypothesis 2:** There is no significant difference in rating the overall average performance of a surgent teacher and a nonsurgent teacher, after observing them in a videotaped preobservation conference and examining script tapes of their lesson.

The purpose of this hypothesis was to determine whether surgency bias affects subjects' ratings of the teacher's overall teaching performance. The average overall rating item on the *Teacher Performance Rating Scale* was used to test this hypothesis. Using a *t*-test for unmatched pairs, a highly significant *t*-value of 4.94 and associated probability of .0001 was obtained (see Table 8).

### Table 8. *T*-test analysis of the difference in average overall ratings between subjects in the surgent and nonsurgent teacher performance evaluation rater groups

<table>
<thead>
<tr>
<th>Group</th>
<th>N</th>
<th>Mean</th>
<th>S.D.</th>
<th>DF</th>
<th><em>t</em>-value</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Surgent</td>
<td>63</td>
<td>3.7301</td>
<td>.7664</td>
<td>119</td>
<td>4.9433</td>
<td>.0001**</td>
</tr>
<tr>
<td>Nonsurgent</td>
<td>58</td>
<td>3.0862</td>
<td>.6565</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

****Equals <.01.
On the basis of this test, the null hypothesis was rejected. Subjects who observed the surgent teacher's videotaped preobservation conference rated her overall teaching performance, based on examination of identical script tapes, significantly higher than those subjects who observed the nonsurgent teacher's videotaped preobservation conference.

**Hypothesis 3:** There is no significant difference in average rating of the performance of a surgent and a nonsurgent teacher, after observing them in a videotaped preobservation conference and examining a script tape of their lesson, for specific teaching behaviors.

This hypothesis was postulated to determine whether surgency bias affects average subjects' ratings of specific teacher behaviors that occur during a lesson. On the basis of these t-tests for unmatched pairs, subjects in the surgent teacher performance evaluation rater group rated the surgent teacher as being significantly more effective in her efforts to get all students to participate during the lesson. This is demonstrated by a t-value of 2.57 and accompanying probability of .01. She was also seen as much more effective in the methods and appropriate techniques used to question students. This resulted in a t-value of 2.31 and associated probability of .02. The surgent teacher's efforts to ensure student time-on-task, shown by the t-value of 3.39 and probability of .0009, and her sensitivity to the needs of all students in her class, depicted by the t-value of 2.89 and associated probability of .004, were rated as much more effective than those of the nonsurgent teacher. As shown in Table 9, while the average ratings of the surgent teacher's efforts to organize the classroom effectively for instruction,
Table 9. T-test analysis of the difference in mean ratings on ten specific teacher behaviors between subjects in the surgent and nonsurgent teacher performance evaluation rater groups

<table>
<thead>
<tr>
<th>Variable</th>
<th>Group</th>
<th>N</th>
<th>Mean</th>
<th>S.D.</th>
<th>DF</th>
<th>t-value</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Organizes students</td>
<td>Surgent</td>
<td>57</td>
<td>3.4211</td>
<td>1.1011</td>
<td>103</td>
<td>0.3262</td>
<td>0.7450</td>
</tr>
<tr>
<td></td>
<td>Nonsurgent</td>
<td>48</td>
<td>3.3542</td>
<td>0.9783</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>States objectives</td>
<td>Surgent</td>
<td>63</td>
<td>3.4921</td>
<td>1.0453</td>
<td>118</td>
<td>1.2364</td>
<td>0.2188</td>
</tr>
<tr>
<td></td>
<td>Nonsurgent</td>
<td>57</td>
<td>3.2456</td>
<td>1.1383</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Provides for participation</td>
<td>Surgent</td>
<td>63</td>
<td>3.9841</td>
<td>1.0079</td>
<td>118</td>
<td>2.5655</td>
<td>0.0116*</td>
</tr>
<tr>
<td></td>
<td>Nonsurgent</td>
<td>57</td>
<td>3.5088</td>
<td>1.0199</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Questioning techniques</td>
<td>Surgent</td>
<td>63</td>
<td>3.3016</td>
<td>1.0570</td>
<td>118</td>
<td>2.3078</td>
<td>0.0227*</td>
</tr>
<tr>
<td></td>
<td>Nonsurgent</td>
<td>57</td>
<td>2.8771</td>
<td>0.9462</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Checks for understanding</td>
<td>Surgent</td>
<td>63</td>
<td>3.4762</td>
<td>0.9648</td>
<td>117</td>
<td>1.9218</td>
<td>0.0571</td>
</tr>
<tr>
<td></td>
<td>Nonsurgent</td>
<td>56</td>
<td>3.1250</td>
<td>1.0280</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gives feedback</td>
<td>Surgent</td>
<td>62</td>
<td>3.3871</td>
<td>0.9296</td>
<td>116</td>
<td>0.9403</td>
<td>0.3490</td>
</tr>
<tr>
<td></td>
<td>Nonsurgent</td>
<td>56</td>
<td>3.2321</td>
<td>0.8526</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Models communications</td>
<td>Surgent</td>
<td>59</td>
<td>3.3729</td>
<td>0.8690</td>
<td>112</td>
<td>1.7163</td>
<td>0.0889</td>
</tr>
<tr>
<td></td>
<td>Nonsurgent</td>
<td>55</td>
<td>3.0727</td>
<td>0.9973</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Knowledge of subject matter</td>
<td>Surgent</td>
<td>62</td>
<td>3.7258</td>
<td>1.0270</td>
<td>117</td>
<td>1.3787</td>
<td>0.1706</td>
</tr>
<tr>
<td></td>
<td>Nonsurgent</td>
<td>57</td>
<td>3.4912</td>
<td>0.8045</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Equals <.05.
Table 9. Continued

<table>
<thead>
<tr>
<th>Variable</th>
<th>Group</th>
<th>N</th>
<th>Mean</th>
<th>S.D.</th>
<th>DF</th>
<th>t-value</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ensures time-on-task</td>
<td>Surgent</td>
<td>60</td>
<td>3.8833</td>
<td>1.0010</td>
<td>112</td>
<td>3.3993</td>
<td>.0009**</td>
</tr>
<tr>
<td>Nonsurgent</td>
<td>54</td>
<td>3.2778</td>
<td>.8777</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Demonstrates sensitivity</td>
<td>Surgent</td>
<td>60</td>
<td>3.5833</td>
<td>.9793</td>
<td>112</td>
<td>2.8981</td>
<td>.0045**</td>
</tr>
<tr>
<td>Nonsurgent</td>
<td>54</td>
<td>3.0741</td>
<td>.8871</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Equals <.01.
her efforts to state the purpose for the lesson, her attempts to check for student understanding of concepts introduced, her feedback to students on their responses, the communications opportunities that she included during the lesson, and her knowledge of the physical science subject matter were all higher than the average ratings for the nonsurgent teacher, they were not statistically higher.

**Hypothesis 4:** There is no significant positive correlation between mean evaluator ratings of the overall performance of a teacher and the amount of teacher performance evaluation training of the evaluator.

One of the research questions that was posed at the beginning of this study was whether training in teacher performance evaluation would result in presence of surgency bias. The correlation coefficient was used to compare the number of days of teacher performance evaluation training held by subjects and their overall ratings of the surgent or nonsurgent teacher's performance.

On the basis of this test, the null hypothesis was retained. There was no significant positive relationship between the amount of training in teacher performance evaluation and overall ratings of surgent or nonsurgent teacher's performance.

**Hypothesis 5:** There is no significant positive correlation between mean evaluator ratings of the overall performance of a teacher and the amount of teacher performance evaluation experience of the evaluator.

Another one of the research questions considered at the outset of this study was whether the number of years of experience one had as a teacher
Table 10. Pearson correlation coefficient between the amount of teacher performance evaluation training and mean overall ratings

<table>
<thead>
<tr>
<th>N</th>
<th>Correlation coefficient</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>120</td>
<td>-.1013</td>
<td>.2688</td>
</tr>
</tbody>
</table>

Mean number of days of training and overall rating

<table>
<thead>
<tr>
<th>Variable</th>
<th>N</th>
<th>Mean</th>
<th>Standard deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall rating</td>
<td>120</td>
<td>3.4298</td>
<td>.7727</td>
</tr>
<tr>
<td>Number of days</td>
<td>120</td>
<td>2.5950</td>
<td>1.0048</td>
</tr>
</tbody>
</table>

performance evaluator would result in presence of surgency bias. To test this hypothesis, a correlation coefficient was calculated comparing the number of years of experience of teacher performance evaluation held by subjects and their overall ratings of the teacher's performance.

On the basis of this test, the null hypothesis was retained. There was no positive relationship between the number of years of experience one had as a teacher performance evaluator and overall ratings of a teacher's performance.

Hypothesis 6: There is no significant difference in the average rating of the overall performance of a teacher when categorized by group, level of administrative assignment, and the interaction of these two factors.

The purpose of this hypothesis was to determine whether different grade levels of responsibility would affect evaluators' awareness of surgency bias as they rated the overall performance of a teacher.
Table 11. Pearson correlation coefficient between number of years of teacher performance evaluation experience and mean overall ratings

<table>
<thead>
<tr>
<th>N</th>
<th>Correlation coefficient</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>120</td>
<td>-.0250</td>
<td>.7867</td>
</tr>
</tbody>
</table>

Mean number of years of experience and overall ratings

<table>
<thead>
<tr>
<th>Variable</th>
<th>N</th>
<th>Mean</th>
<th>Standard deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall rating</td>
<td>120</td>
<td>3.4298</td>
<td>.7727</td>
</tr>
<tr>
<td>Experience</td>
<td>120</td>
<td>2.4125</td>
<td>1.1012</td>
</tr>
</tbody>
</table>

A two-way analysis of variance was used to determine differences in average ratings. On the basis of this test, the hypothesis was retained. Average overall ratings by evaluators at various grade levels of administrative assignment were not significantly different, as shown by the probability of .9424 in Table 12. When trying to account for the factors that contributed to any significant difference, the group variable had an overwhelming impact. This is depicted by the F-value of 24.77 and associated probability of .0001. The interaction of the group and level variables, with an F-value of 2.21 and associated probability of .0725, are worth noting.
Table 12. ANOVA and means of the differences in average overall ratings between subjects in the surgeon and nonsurgeon teacher performance evaluation rater groups at specific levels of administrative assignment

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>S.S.</th>
<th>M.S.</th>
<th>F-value</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>9</td>
<td>17.3793</td>
<td>1.9310</td>
<td>3.82</td>
<td>.0003**</td>
</tr>
<tr>
<td>Group</td>
<td>1</td>
<td>12.5225</td>
<td>12.5225</td>
<td>24.77</td>
<td>.0001**</td>
</tr>
<tr>
<td>Group*Level</td>
<td>4</td>
<td>4.4696</td>
<td>1.1174</td>
<td>2.21</td>
<td>.0725</td>
</tr>
<tr>
<td>Error</td>
<td>111</td>
<td>56.1248</td>
<td>.5056</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected totals</td>
<td>120</td>
<td>73.5041</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Mean overall ratings by group and by level

<table>
<thead>
<tr>
<th>Group</th>
<th>Level</th>
<th>N</th>
<th>Mean overall rating</th>
</tr>
</thead>
<tbody>
<tr>
<td>Surgeant</td>
<td>Elementary</td>
<td>25</td>
<td>3.7600</td>
</tr>
<tr>
<td>Surgeant</td>
<td>Middle level</td>
<td>4</td>
<td>3.2500</td>
</tr>
<tr>
<td>Surgeant</td>
<td>Senior high</td>
<td>14</td>
<td>3.7143</td>
</tr>
<tr>
<td>Surgeant</td>
<td>Kindergarten-Grade 12</td>
<td>3</td>
<td>3.3333</td>
</tr>
<tr>
<td>Surgeant</td>
<td>Central office</td>
<td>17</td>
<td>3.8823</td>
</tr>
<tr>
<td>Nonsurgeon</td>
<td>Elementary</td>
<td>19</td>
<td>3.1579</td>
</tr>
<tr>
<td>Nonsurgeon</td>
<td>Middle level</td>
<td>10</td>
<td>3.4000</td>
</tr>
<tr>
<td>Nonsurgeon</td>
<td>Senior high</td>
<td>17</td>
<td>3.0000</td>
</tr>
<tr>
<td>Nonsurgeon</td>
<td>Kindergarten-Grade 12</td>
<td>2</td>
<td>3.5000</td>
</tr>
<tr>
<td>Nonsurgeon</td>
<td>Central office</td>
<td>10</td>
<td>2.7000</td>
</tr>
<tr>
<td></td>
<td></td>
<td>121</td>
<td>3.4215</td>
</tr>
</tbody>
</table>

**Equals <.01.
Hypothesis 7: There is no significant difference in the average level of rigor of Professional Improvement Commitments (PICs) selected by evaluators for a surgent and a nonsurgent teacher.

Another research question formulated at the beginning of this study was whether surgency bias would affect PICs selected for a surgent teacher and a nonsurgent teacher. Would those subjects who observed the nonsurgent teacher select a more rigorous or challenging PIC for her to accomplish than those subjects who observed the surgent teacher?

A t-test analysis for unmatched pairs revealed that the average Professional Improvement Commitment (PIC) selected for the surgent teacher was not significantly different than the average PIC chosen for the nonsurgent teacher, resulting in a t-value of .49 and associated probability of .62.

On the basis of this test, the null hypothesis was retained. The average Professional Improvement Commitment (PIC) selected for the nonsurgent teacher was more rigorous than the PIC selected for the surgent teacher, but not significantly so.

Table 13. T-test analysis for the difference in average Professional Improvement Commitments (PICs) selected between subjects in surgent and nonsurgent teacher performance evaluation rater groups

<table>
<thead>
<tr>
<th>Group</th>
<th>N</th>
<th>Mean</th>
<th>S.D.</th>
<th>DF</th>
<th>t-value</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Surgent</td>
<td>63</td>
<td>3.0159</td>
<td>.9587</td>
<td>119</td>
<td>.4909</td>
<td>.6244</td>
</tr>
<tr>
<td>Nonsurgent</td>
<td>58</td>
<td>3.1033</td>
<td>1.0032</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Hypothesis 8: There is no significant difference in the mean overall rating of a teacher's performance when categorized by the gender of the evaluator.

There have been mixed findings in recent studies conducted by industrial psychologists and educational researchers regarding the effect gender has on performance evaluation. As stated in this hypothesis, gender of the evaluators had little contribution to the average overall ratings of performance, regardless of their gender. As indicated by the t-value of .3561 and the associated probability of .72 for the t-test for unmatched pairs, the average ratings of overall performance were not significantly different. Said another way, in this investigation women evaluators were just as biased by surgency of a woman teacher as were men.

Table 14. T-test analysis of the difference in mean overall ratings between male and female evaluators

<table>
<thead>
<tr>
<th>Gender</th>
<th>N</th>
<th>Mean</th>
<th>S.D.</th>
<th>DF</th>
<th>t-value</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>82</td>
<td>3.4390</td>
<td>.7552</td>
<td>119</td>
<td>.3561</td>
<td>.7224</td>
</tr>
<tr>
<td>Female</td>
<td>39</td>
<td>3.3846</td>
<td>.8465</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
The primary purposes of this study were to (1) examine predictions of teacher effectiveness, overall ratings of teacher performance, ratings of ten specific teaching behaviors, and Professional Improvement Commitment (PIC) selections from two groups participating in teacher performance evaluation training, to determine the effects of surgency bias, if any, on various aspects of performance evaluation; (2) to determine the effects of surgency bias as it relates to gender, administrative assignment, and grade levels of responsibility; and (3) to determine the effects of teacher performance evaluation training and teacher performance evaluation experience on surgency bias.

Summary

One hundred twenty-one subjects involved in teacher performance evaluation training in San Antonio, Texas; Salina, Kansas; and Ames, Iowa, provided data for this study to explore surgency bias. A summary of the findings, based on data gathered in the early spring of 1990, follows.

Conclusions

This study has significant implications for educators, performance evaluators, those who train evaluators and, ultimately, for students. If the findings of this study are supported by further research, the conclusion could be drawn that without remediation and awareness training, any evaluation that includes conferencing and classroom observation may be affected and invalid because of the influence of surgency bias.
As stated in Chapter 2 and summarized in Table 1, previous research studies, conducted in the private sector, have shown mixed findings and reached inconsistent conclusions regarding any possible effects of surgency bias as it relates to performance evaluation. Earliest research study findings tended to indicate that physical attractiveness was associated with higher performance ratings, while, in later studies this potential source of bias was apparently diminishing. Table 15 indicates that surgency bias was apparently present in the minds of subjects in this study during much of the evaluation process.

Table 15. Summary table showing findings related to surgency bias

<table>
<thead>
<tr>
<th>Variable</th>
<th>Surgent group</th>
<th>Nonsurgent group</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>S.D.</td>
<td>Mean</td>
</tr>
<tr>
<td>Prediction</td>
<td>4.38</td>
<td>.6822</td>
<td>3.43</td>
</tr>
<tr>
<td>Overall rating</td>
<td>3.73</td>
<td>.7664</td>
<td>3.09</td>
</tr>
<tr>
<td>Teaching behavior</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Provides for participation</td>
<td>3.98</td>
<td>1.0079</td>
<td>3.51</td>
</tr>
<tr>
<td>Questioning techniques</td>
<td>3.30</td>
<td>1.0570</td>
<td>2.88</td>
</tr>
<tr>
<td>Time-on-task</td>
<td>3.88</td>
<td>1.0010</td>
<td>3.28</td>
</tr>
<tr>
<td>Demonstrates sensitivity</td>
<td>3.58</td>
<td>.9793</td>
<td>3.07</td>
</tr>
<tr>
<td>PIC</td>
<td>3.02</td>
<td>.9587</td>
<td>3.10</td>
</tr>
</tbody>
</table>

*Equals <.05.

**Equals <.01.
Obviously, this could have a tremendous impact on performance-based pay, on tenure, on selection of Professional Improvement Commitments (PICs), and on a broad spectrum of related instructional evaluation issues. The conclusions follow:

1. Surgency bias significantly influenced evaluators' predictions of the teacher’s effectiveness.

2. Surgency bias significantly influenced evaluators’ ratings of the teacher’s overall performance.

3. Surgency bias significantly influenced evaluators’ ratings of four specific teaching behaviors. These included providing for student participation, questioning techniques, ensuring time-on-task, and demonstrating sensitivity to student needs. The six other teaching behaviors rated in this study were not influenced to a significant degree.

4. Surgency bias did not significantly influence evaluators' Professional Improvement Commitment (PIC) selection. The mean PIC selected for the nonsurgent teacher was only slightly more rigorous.

5. The amount of teacher performance evaluation training was not related to evaluators' tendency to be influenced by surgency.

6. The number of years of teacher performance evaluation experience did not influence the evaluator's tendency to be affected by surgency.

7. Surgency bias did not significantly vary by subjects' levels of administrative assignment.

8. Surgency bias of overall ratings of performance did not vary by gender of evaluators.
Discussion

The implications of some of these findings have powerful ramifications on performance evaluation, while others are more subtle. Three overarching conclusions can be drawn from this study: (1) Surgency bias influences the first impression of a teacher's potential for effective classroom performance and preliminary confidence in his/her capabilities, (2) surgency bias influences the overall rating of a teacher's performance, and (3) surgency bias influences evaluator ratings of some independent teaching behaviors. Evaluators who observed the surgent teacher preobservation conference videotape were much more likely to predict higher teacher effectiveness, rate overall performance higher, and rate performance for specific teaching behaviors higher than those who viewed the nonsurgent teacher preobservation conference videotape. To their credit, subjects from both groups selected similar PICs, based on an analysis of a script tape. Evaluators who are influenced by the initial impression made by the teacher during a preobservation conference may generalize that bias to judgments made throughout the process of evaluation.

A logical extension of that assumption then follows that rating of specific teaching behaviors and overall performance and monitoring the accomplishment of Professional Improvement Commitments may also be influenced. The result of surgency bias may be that a teacher's performance is not accurately and objectively observed and recorded and any decisions made may be distorted. Teachers who are seen as nonsurgent may get low ratings on specific teaching criteria, low overall ratings,
and be required to complete more challenging or rigorous PICs. On the other hand, surgent teachers may receive higher ratings and be required to complete easier PICs for similar classroom performance.

If surgency bias has such a significant effect on an evaluator, is it possible that the impact on students may be just as powerful? If the teacher is viewed as surgent, does that attraction, charisma, and enthusiasm that is sensed by students translate into higher student achievement? If the answer is yes, recruiting and retaining surgent educators would be an excellent goal.

The results of this study failed to show a significant relationship between teacher performance evaluation training and experience and any influence associated with surgency bias. This is not too surprising, since very little attention and related training has been given. Until evaluators are made aware of this bias, influence of surgency on performance evaluation will continue to occur. Experience apparently does not nullify instances of surgency bias. Evaluators tend to repeat past practices unless they receive awareness training and inservice. Nonsurgent teachers stand the most to gain from this study, if the findings are disseminated and adjustments are made.

Limitations

The generalizability of the findings of this study is limited by the inability to control several variables.

1. This study was conducted at three locations--Ames, Iowa; San Antonio, Texas; and Salina, Kansas.
2. This study involved a limited sample of subjects who were voluntarily attending professional growth activities.

3. The actresses who portrayed the surgent and nonsurgent teachers were both female and white.

4. Subjects did not observe actual teaching performance, but had to rely on script tapes of a lesson to make their ratings and PIC selections.

5. The subject area depicted in this study was science.

6. The grade level selected for this study was ninth grade.

7. The actress who portrayed the nonsurgent teacher may not have captured a neutral position, but may have created a negative role.

Recommendations for Further Research

This study was an initial attempt into a previously unexplored area of the teacher performance evaluation process. Replication and additional study are needed to confirm findings and analyze this phenomenon more thoroughly.

1. This study should be replicated in other settings and with other teachers and evaluators. For example, two male actors could be used in the same situation.

2. The effects of surgency bias on the evaluator and the teacher have been addressed in this study. Additional research on the effects of surgency bias on students and achievement is suggested.

3. Possible surgency bias effects may be reduced through awareness and training. A study of the effect of evaluator training on ratings of performance is suggested.
4. Surgency bias has a strong positive impact on performance ratings. Studying evaluator ratings and student perceptions before and after training may aid in determining the extent of this bias.

5. Surgency bias is not the only bias affecting performance ratings. Similar studies should be conducted assessing the effects of age, gender, race, and other possible biases and any interaction of these that might affect performance ratings and student achievement.

Recommendations for Practitioners

Below are several recommendations which could make a difference in the evaluation of performance and education of students.

1. If surgency bias affects evaluator ratings of performance, it may also influence students. Teachers should be aware of its potential impact.

2. Evaluators should be made aware of the effects of surgency bias on interviewing, observing performance, and evaluating performance. Discussion and awareness training related to surgency bias and its effects on ratings, PIC selection, and student achievement should be included in evaluator training.

3. Evaluators of evaluators should be made aware of the presence of surgency bias throughout the process of evaluation and should include a bias awareness module within evaluation training.

4. Effects of surgency bias may be reduced through the use of multiple appraisers to evaluate performance.
5. Utilizing multiple data sets, including student achievement data, self-ratings, student ratings, and longitudinal evaluation data is another appropriate method for minimizing effects of surgency bias.

6. A similar experiment should be tried using multiple raters to determine if averaging ratings after rater discussion would reduce surgency bias.
BIBLIOGRAPHY


Heilman, M. E., and Stopec, M. H. "Is Being Attractive an Advantage or Disadvantage? Performance-Based Evaluations and Recommended Personnel Actions as a Function of Appearance, Sex, and Job Type." 


ACKNOWLEDGMENTS

Several people were instrumental in helping me formulate, implement, and conclude this study. I want to especially express my thanks and gratitude to Dick Manatt for his inspiration, advice, and counsel throughout this project. Appreciation is also extended to POS committee members Shirley Stow, Tony Netusil, Sande McNabb, and Mark Redmond for their insights and technical assistance.

Morgan Halgren and Michelle Larche did an outstanding job of portraying the essence of surgency and nonsurgency and Jerry Kelley and Judy Dietrich did an equally exceptional job of capturing their efforts on videotape.

Alice Cheng, statistical programmer, and Bonnie Trede, typist, were patient, dependable, and accurate in transforming my drafts into a finished product.

Finally, my special thanks, appreciation, and expression of love to my wife, Linda, and our children for hanging in there with me and giving up much time and companionship this past year. This really was a family effort.
APPENDIX A. TEACHER PERFORMANCE RATING SCALE
### TEACHER PERFORMANCE RATING SCALE

**NAME:**

**GROUP:** A or B, I.D. #

**DIRECTIONS:** After viewing the videotape, please mark the level of performance for each criterion that indicates your rating of the teacher's performance.

### Teacher Behaviors

**Teacher Behaviors**

(The teacher . . . )

<table>
<thead>
<tr>
<th>Levels of Performance</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
</tr>
</tbody>
</table>

1. Organizes students for effective instruction.

2. States instructional objective(s)

3. Provides opportunities for student participation.

4. Incorporates effective questioning techniques.

5. Checks for student understanding.

6. Gives supportive and immediate feedback to students.

7. Models effective communications skills.

8. Displays a thorough knowledge of curriculum and subject matter.

9. Ensures student time-on-task.

10. Demonstrates sensitivity in relating to students.

**OVERALL RATING OF PERFORMANCE**

**DIRECTIONS:** Please rate the teacher's overall performance.

<table>
<thead>
<tr>
<th>Levels of Performance</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
</tr>
</tbody>
</table>

11. Overall rating.

**PROFESSIONAL IMPROVEMENT COMMITMENT** (PIC)

Circle One: 1 2 3 4 5
APPENDIX B. PERSONAL DATA FORM
PERSONAL DATA FORM

NAME: ____________________________________________________ GROUP: A or B I.D.#____________________________

Mark one: Male: ____ Female: ____


Mark Your Administrator Position: 
- Building Administrator
- elementary
- middle level
- secondary
- Central Office Staff
- Performance Evaluator
- District Administrator
- Teacher

Years experience as teacher performance evaluator (circle one):

0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37+

Number days teacher performance evaluation training (circle one):

0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37+

Based on the videotaped preobservation conference with Judy Oliver you just observed, predict the overall effectiveness of her performance:

1 - Level 1 Performance is highly unacceptable. The teacher does virtually nothing of educational value related to this criterion.

2 - Level 2 Performance is not at an acceptable level. The teacher is deficient enough to be ineffective for this criterion.

3 - Level 3 Performance is acceptable. The teacher demonstrates adequate skill for this criterion.

4 - Level 4 Performance is of high quality. The teacher is above average for this criterion, but not good enough to serve as a model for others.

5 - Level 5 Performance could serve as a model for other teachers. The teacher demonstrates a high proficiency for this criterion.

Level of Performance

1 2 3 4 5

[ ] [ ] [ ] [ ] [ ]
APPENDIX C. PERMISSION SLIP
January 26, 1990

Dear Student and Parent:

I am completing a research study as part of my Ph. D. program at Iowa State University. A component of this research study is a six minute video tape of a simulated (staged) preobservation conference. As a lead-in to this conference, I need some shots of students in a science classroom at their seats and in the lab. These shots will be background for the introductory comments I make that will be added later, so there is no need for any speaking parts for students.

You child has indicated that he/she would like to be an actor for this project, but I need your permission before I can allow his/her participation. If it is all right with you for him/her to participate, please sign the appropriate space on this release and have your child return it to me.

Please call me at school (961-7475) or at home (961-4932) if you have any questions about this project. All actors are volunteers and will receive no reimbursement for participation. Actual filming will occur on Tuesday, January 30, 1990.

Thank you for your cooperation and permission,

Bob Newsum

Yes, you have my permission to include my child, ______________________________, in this video project.

Student Signature ______________________________ (Date)

Parent Signature ______________________________ (Date)
APPENDIX D. WORKSHOP SCHEDULE
### INSTRUCTIONAL PLAN

**Title:** Evaluating, Conferencing, Motivating Teachers  
**Group or School:** USA KanLEAD  
**Date(s):** Wednesday, March 28, 1990  
**Presenting Consultant(s):** Dick Manatt  
**Attending:** Administrators

"Skills Booster Shot" plus  
The Rest of the Story on Marginal Teachers

<table>
<thead>
<tr>
<th>TIME</th>
<th>TOPIC</th>
<th>PRESENTER</th>
<th>MODE</th>
<th>VISUALS</th>
<th>HANDOUTS</th>
<th>REMARKS</th>
</tr>
</thead>
<tbody>
<tr>
<td>9:00</td>
<td>Review and Preview</td>
<td>Manatt</td>
<td>LGI</td>
<td>O/H</td>
<td>Recipe Cards</td>
<td>--</td>
</tr>
<tr>
<td>9:30</td>
<td>The Complete TPE Cycle</td>
<td>Manatt</td>
<td>IS</td>
<td>O/H, Video</td>
<td>&quot;Judy Oliver&quot; Kit</td>
<td>Progress Check</td>
</tr>
<tr>
<td>10:15</td>
<td>--Break--</td>
<td>OYO</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>10:30</td>
<td>Bias: Human Error in Rating</td>
<td>Manatt</td>
<td>LGI</td>
<td>O/H</td>
<td>Dallas Findings</td>
<td>--</td>
</tr>
<tr>
<td>11:00</td>
<td>Overcoming Bias</td>
<td>Manatt</td>
<td>LGI/IS</td>
<td>O/H, Video</td>
<td>&quot;Kelly Mitchell&quot;</td>
<td>Working Solo</td>
</tr>
<tr>
<td>12:00</td>
<td>--Lunch--</td>
<td>OYO</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>12:45</td>
<td>Due Process Supervision</td>
<td>Manatt</td>
<td>LGI</td>
<td>O/H</td>
<td>SMT Workbook</td>
<td>--</td>
</tr>
<tr>
<td>1:45</td>
<td>--Break--</td>
<td>OYO</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>2:00</td>
<td>The Dismissal Team</td>
<td>Manatt</td>
<td>LGI</td>
<td>O/H</td>
<td>SMT Workbook</td>
<td>--</td>
</tr>
<tr>
<td>2:45</td>
<td>Dismissal</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
The Teacher Performance Evaluation Cycle:
NINTH GRADE PHYSICAL SCIENCE -
JUDY OLIVER

INSTRUCTION MATERIALS PACKET (abstracted for training)

A Videotape Series
for Staff Development

Richard P. Manatt
Bob Newsum
GROUP DESCRIPTION

Class: Physical Science

Grade: 9

Teacher: Judy Oliver

Number of Students: 20 (10 boys and 10 girls)

Age Range: 14-16 years

Ability Range: Below average to above average

Typical Interest and Involvement: Students in the class are generally friendly and cooperative. Students are usually creative in their approach to problem solving and are comfortable in working in cooperative teams.

Prior Relevant Information: This class has been studying solutions and are designing science fair experiments. In the previous period of instruction they:

a. discussed factors influencing the rate of solution.

b. discussed how the solution process occurs.

c. wrote a lab report in preparation for this lesson.
INSTRUCTIONAL PLANS

Teacher: Judy Oliver

Grade: 9

Subject: Physical Science

Period: 10:30-11:15 a.m.

Date: March 27, 19--

Unit topic: Solutions

Topic for this lesson: How temperature affects rate of solution

Instructional objective: After performing an experiment, students will construct a bar graph to show the differences in the rate of solution of Alka-Seltzer tablets at four different temperatures, and write a conclusion from the data.

Materials/media: Student-prepared lab report, a chalkboard, a portable laboratory, and graph paper and drawing tools.

Instructional procedures:
1. Review the factors affecting the rate of solution.
2. Discuss the experiment to be conducted.
3. Perform the experiment, involving the students.
4. Explain how to construct the bar graphs from the data.
5. Discuss possible conclusions from the data.
6. Assign students to write their conclusions.

Student activities:
1. Participate in discussion.
2. Assist with experiment.
3. Record data and draw bar graph.
4. Write conclusions based on data.

Evaluation of learner outcomes:
1. Assess class participation.
2. Check graphs and written conclusions.
PREOBSERVATION DATA FORM

For: Judy Oliver

Date: March 27, 19--  Class: Physical Science  Time: 10:30-11:15

Please complete the following information and return to the evaluator one day before the scheduled date of the observation.

1. Where are you in the course?
   Chemistry unit-solids, liquids, gases
   Science Fair Projects

2. What outcomes do you expect?
   After performing a group experiment, students will construct a bar graph to show differences in the rate of solution of an Alka Seltzer tablet at different temperatures, and write a conclusion from the data.

3. What teaching methods do you plan to use?
   Review
   Explanation-modeling
   Checking for understanding
   Guided and Independent practice

4. What learning activities will be observed?
   Responses to questions
   Guided practice-construction of bar graphs
   Independent practice-written conclusions

5. Are there any particular teaching behaviors that you especially want monitored?
   No

6. How are you going to know if the students have learned?
   Accuracy of bar graphs (construction and data represented)
   Accuracy of written conclusions

7. What special characteristics of the students should be noted?
   None
Lesson Timeline

Ninth Grade Physical Science - Judy Oliver

Date: March 27, 19--

<table>
<thead>
<tr>
<th>Time</th>
<th>Teacher narrative/behaviors</th>
<th>Student behaviors</th>
</tr>
</thead>
<tbody>
<tr>
<td>0:00</td>
<td>(Teacher at front of room) We've been talking... solutions ... things that change the rate of solutions.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Your book lists three things ... and we've discussed these before today.</td>
<td></td>
</tr>
<tr>
<td>0:15</td>
<td>Can someone remind me ... first thing ... change the rate of solution? Amy?</td>
<td>Amy responds.</td>
</tr>
<tr>
<td></td>
<td>Raise the temperature of what?</td>
<td>Amy repeats.</td>
</tr>
<tr>
<td></td>
<td>Of the solvent. Okay.</td>
<td></td>
</tr>
<tr>
<td>0:30</td>
<td>What's another ... Melissa?</td>
<td>Melissa responds.</td>
</tr>
<tr>
<td></td>
<td>By stirring it.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>David, what's a third one?</td>
<td>David responds.</td>
</tr>
<tr>
<td></td>
<td>Breaking it ... crushing it up into little pieces.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Okay, so those three things ... cause a solution to happen faster.</td>
<td></td>
</tr>
<tr>
<td>0:48</td>
<td>Today, we want to study one ... effect of temperature on that rate of solution.</td>
<td></td>
</tr>
<tr>
<td>1:00</td>
<td>So, our objective ... (points to chalkboard) ... find out ... change in temperature of solvent ... changes the rate of solution ... construct a bar graph to show our data.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Now ... think back ... what we had as an objective ... go through steps of doing a science fair project.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Why did we do that ... anybody remember?</td>
<td>Some students raise hands</td>
</tr>
<tr>
<td>1:30</td>
<td>Because there's going to be a science fair! ... and you know ... want you to do a science fair project We'll be doing this ... between now and ... science fair so that ... second nature to you.</td>
<td></td>
</tr>
</tbody>
</table>
Now, yesterday we wrote... lab report... project report... get those out, please.

Everybody have your paper?

Now, you remember, we said writing up science fair project... like writing lab report, except... more detail... more steps...

What the first step we did on this write-up? Joe?

Okay, the question... and what was that?

Okay. How does the change in temperature change... solution.

We went on... to the next step in our project.

The hypothesis... can you read it for me?

Okay... so our hypothesis states... increasing the temperature... cause rate of solution to speed up.

Okay, so that was our question and hypothesis.

The next thing... Ouji?

Okay... list our materials.

And all of you had that list.

Then we needed to design our experiment, and how did we go about doing that?

First of all, what do we call the next step?... David?

The procedure.

Now, whenever we work out that... be careful... keep as many things... constant.

Now, in our... experiment... one thing to change.

What one thing... allowing to change?

Shane?

We're allowing the temperature to change.

What... call thing that changes? Joe?

A variable. Okay.

And... in our experiment... only one variable.

We don't want... other things changing.
4:15 What other things that might change other than temperature? Laura?

**Laura responds**

Color could change. Would that affect our experiment if it did?

**Class responds.**

No, it might even change in the process of solution. Suzy?

**Suzy responds.**

Okay the amount of water.

4:30 We decided on a set amount of water when we set up our experiment. How much?

**Class responds.**

Three hundred ml of water. and I have that amount.

What's something else that would be a variable other things? David?

**David responds.**

Okay the size. We're assuming the tablets same size.

5:00 If you were using something else as your solute sugar or salt have to measure that wouldn't you? We're letting the company keep amount constant, and we hope they did.

5:20 The first thing before we collect our data somebody to record data for us.

Okay, Shane do that?

**Some hands raised.**

5:30 Just stay up there okay? Now, we want to do first part of our experiment, which was which part?

Okay and we were going to start at room temperature for us.

**Student responds.**

6:00 Okay, Rachael.

(Gives thermometer to Rachael.)

**Rachael goes to lab table at front.**

(Instructs Rachael how to hold thermometer.) Can you see it okay?

**Rachael begins to measure temperature of water in beaker.**

**Rachael responds.**
Okay, 21° C.

6:30  Okay! Thank you.

6:40  Now we need somebody . . . to do the timing for us.
Okay, Mitzi . . . you want to do that?
Okay, I'm going to have Mitzi time . . . with a stopwatch, and . . . use something like that . . . watch with second hand. (Hands stopwatch to Mitzi.)

7:00  If you need accurate timing . . . use some way . . . really accurate. (Gives stopwatch instructions.)
Now, I'm going to count 3 . . . 2 . . . 1 . . . go.
(Drop the tablet into the water.)

7:30  Is it all through? What was our time?
Okay, 31 seconds.
Write that in your data table.
Thank you, Mitzi.

8:00  By-the-way, let me mention . . . hotplate is hot . . . could burn . . . be real careful.

8:15  Okay, now second temperature . . . cold temperature . . .

8:30  (Adds ice to water in beaker and pours off excess.)
All right . . . need volunteer . . . to read temperature again.

9:00  Okay, Melissa.
(Repeats previous procedure in experiment.)

9:30  10° Celsius is our next temperature. Everybody got that?

10:00  (Experiment continues as tablet dissolves.)

10:30  (Experiment continues.)

10:30  Is it done?
Okay, 47 seconds
11:00 Now, for our third one... hope water is still warm... and I need... yes?

I don't know. Who can say... what makes it jump around? Wade?

That's probably exactly right.

11:30 Suzy?

Well, why would you think? Remember when we talked about... there's more than just solution going one... also chemical reaction.

11:45 Now, what happens to the rate of chemical reaction... whenever you add heat to it?

It speeds up. So probably... bubbles are being released faster... makes it bounce around more. Okay.

12:00 All right. Now I need someone... read temperature for me. Okay, Amy.

Now this is the warm water. (Experiment repeated.)

What's the temperature of our warm water?

63° Celsius. Everybody hear that?

12:30 Okay, Now I need somebody... to do timing at 63°... Kent?

(Experiment continues)

13:00 Okay, what did you get?

Okay... 23 seconds.

All right. Thank you.

13:15 Now... if we can stop... getting these temperatures... I want to be sure you understand ranges... in Celsius scale.

13:30 What is boiling temperature, Celsius? Mitzi?

100° Celsius... and we found room temperature to be... what? Amy?
Amy responds.

21° Celsius

So that ... gives you a range ... Since we use Fahrenheit scale ... wanted to be sure you know that.

Okay. Now ... very hot ... water.  
Do I have a volunteer ...?

14:00  
Okay, Rachael.

Now, this time ... I'll hold the thermometer and you read.

14:30  
Okay, 90° Celsius.

(Experiment repeated.)

15:00  
Okay ... 16 seconds this time.

15:15  
See what happened? ... didn't get much of a change this time.

15:30  
Okay, you want to ... take your seat Shane?

We have ... accomplished ... first objective ...

15:47  
Now, our second objective ... bar graph of our data. 
I have graph paper for you.

16:00  
(Distributes graph paper.)

16:15  

16:30  
And then ... straight edges ... protractors work well. 
(Distributes protractors.)

16:42  
Now we want to ... graph of our data. 
Remember ... doing a science fair project here. Okay? 
Now ... we want to try to make data as showy as we can.

17:00  
We're going to ... make a graph that will really show up. 
For now ... just do a form of our bar graph.

(Moves closer to students.)
Okay. Now ... two things we need to put on our graph ... horizontal line ... vertical line.

17:30  
What ... two kinds of numbers ... Amy?

The temperature of the water and ... the speed ... and we measured that in ... seconds.
Since this is a bar graph... and we only used four temperatures... each bar represents one of our temperatures.

(Gives options for layout of graph.)

(Models layout on chalkboard.)

(Continues to model layout.)

Now we need to fit numbers on the graph... What have to do...? Suzy?

Okay... have to fit... on it.

(Checks students progress.)

Okay... want you to see... thing Paul has done... Brian doing the same thing. (Shows paper to class.)

See how he did that?

(Continues modeling of graph construction.)

Do I need to give you... minutes to get your numbers on?

Anyone not finished putting numbers...?

Now, the only other thing we need to decide... how much space to use...?

Do we want to put temperature in order or start with lowest... go from there? Suzy?

Okay... think that's a good point. Start with lowest... show increase.

(Continues to model process of making bar graph.)

I think I can help you best if... work with you individually on that part. (Monitors and helps if necessary.)
26:00 (moves back to front of class.) Let's pause for a few minutes... go on to last part of science fair write-up. What's the very last part we have to do? Shane?

The conclusion.

Remember the rule... for writing the conclusion?

Shane responds.

26:30 What do we have to do in the conclusion? Joe?

Joe responds.

You have to answer the question. Okay. Who has the question? Mitzi?

Mitzi responds.

Okay. How does change... affect rate of solution?

26:45 We went on to the hypothesis, and what was that? Okay, Matt?

Matt responds.

27:00 Okay. We said we think that.... Now we need to write a conclusion. Did increasing... speed up... rate of solution? Melissa?

Melissa responds.

Not every time did it?

Now... was our hypothesis entirely wrong?

Class responds.

27:30 No... wasn't entirely wrong was it... but not completely right. So, we need to write a conclusion to agree with the hypothesis... what ways correct... what ways it was not.

Last two things I want you to do today... write a conclusion for lab report... finish your graph.

28:00 If... not finished... to your satisfaction, keep it... finish it at home or in study hall... turn it in tomorrow at beginning of class.

But if you have it... ready to turn in... turn it in today.

28:20 - 45:00 Students work independently. (Teacher moves around room monitoring student progress.)
Teacher Performance Evaluation (ISURF)

Teacher Performance Rating Scale

1 - Level 1  Performance is highly unacceptable. The teacher does virtually nothing of educational value related to this criterion.

2 - Level 2  Performance is not at an acceptable level. The teacher is deficient enough to be ineffective for this criterion.

3 - Level 3  Performance is acceptable. The teacher demonstrates adequate skill for this criterion.

4 - Level 4  Performance is of high quality. The teacher is above average for this criterion, but not good enough to serve as a model for others.

5 - Level 5  Performance could serve as a model for other teachers. The teacher demonstrates a high proficiency for this criterion.

<table>
<thead>
<tr>
<th>Teacher Behaviors</th>
<th>Levels of Performance</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Organizes students for effective instruction. (The teacher ... )</td>
<td></td>
</tr>
<tr>
<td>1. There was no evidence of organizing students for effective instruction.</td>
<td>Organized students for effective instruction.</td>
</tr>
<tr>
<td>2. There was little evidence of organizing students for effective instruction.</td>
<td>A variety of techniques to organize students for effective instruction.</td>
</tr>
<tr>
<td>3. There was evidence of effective use of a variety of techniques to organize students for effective instruction.</td>
<td></td>
</tr>
<tr>
<td>4. Organized students for effective instruction</td>
<td></td>
</tr>
<tr>
<td>5. A variety of techniques to organize students for effective instruction were used and could be modeled</td>
<td></td>
</tr>
<tr>
<td>N/0 This behavior was not observed.</td>
<td></td>
</tr>
</tbody>
</table>

2. States instructional objective(s). (The teacher clearly tells students what they are going to learn and why it is important.)

| 1. Did not tell students what they were going to learn and why it was important. | 2. Told students what they were going to learn, but did not explain why it was important. | 3. Told students what they were going to learn and why it was important. | 4. Used multimodal techniques to tell students what they would learn and why it was important. | 5. Used multimodal techniques to tell students what they would learn and why it was important. | N/0 This behavior was not observed. |
3. Provides opportunities for student participation - All students have the opportunity for information exchange during the lesson (i.e., show of hands, group and/or individual oral responses, hands-on experiments, etc.)

<table>
<thead>
<tr>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>N/0</th>
</tr>
</thead>
<tbody>
<tr>
<td>Students were not given opportunities to participate.</td>
<td>Specific students were given opportunities to participate.</td>
<td>All students were provided with opportunities to participate.</td>
<td>All students were encouraged to participate orally and nonverbally.</td>
<td>A variety of techniques were used to involve all students on an equitable basis.</td>
<td>This behavior was not observed.</td>
</tr>
</tbody>
</table>

4. Incorporates effective questioning techniques - The teacher effectively probes students' knowledge and/or ideas by using appropriate questioning techniques.

<table>
<thead>
<tr>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>N/0</th>
</tr>
</thead>
<tbody>
<tr>
<td>No effective questioning techniques were evident during this lesson.</td>
<td>Questioning techniques were generally based on the knowledge and comprehension levels of Bloom's taxonomy.</td>
<td>Effective questioning techniques were evident throughout this lesson.</td>
<td>A variety of effective questioning techniques were evident and questions posed were based on all levels of Bloom's taxonomy.</td>
<td>A variety of effective questioning techniques were evident, questions posed were based on all levels of Bloom's taxonomy, and could be modeled for others.</td>
<td>This behavior was not observed.</td>
</tr>
</tbody>
</table>

5. Checks for student understanding - The teacher incorporates strategies to determine whether students understand the concepts taught, to determine length of instruction, who is ready for guided practice and who needs reteaching, etc.

<table>
<thead>
<tr>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>N/0</th>
</tr>
</thead>
<tbody>
<tr>
<td>No strategies to check for student understanding were included in the lesson.</td>
<td>Few strategies to check for student understanding were included in the lesson.</td>
<td>A variety of strategies to check for student understanding were included for some students in the lesson.</td>
<td>A variety of checks for student understanding involved all members of the class.</td>
<td>Several effective strategies to check for student understanding were included in the lesson and could be modeled for others.</td>
<td>This behavior was not observed.</td>
</tr>
</tbody>
</table>
6. Gives supportive and immediate feedback to students -- The teacher frequently provides knowledge of results to the students relative to their level of success.

<table>
<thead>
<tr>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>N/O</th>
</tr>
</thead>
<tbody>
<tr>
<td>Student attempts to respond were not immediately recognized or acknowledged and teacher comments were not supportive.</td>
<td>Some student attempts to respond were immediately recognized or acknowledged and teacher comments were generally supportive.</td>
<td>Student attempts to respond were immediately recognized or acknowledged and teacher comments were supportive.</td>
<td>A variety of feedback strategies were used to encourage all students to respond and could be modeled for others.</td>
<td>This behavior was not observed.</td>
<td></td>
</tr>
</tbody>
</table>

7. Models effective communications skills -- The teacher speaks and writes appropriately for student ability level and for the learning process.

<table>
<thead>
<tr>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>N/O</th>
</tr>
</thead>
<tbody>
<tr>
<td>Efforts to communicate with students were ineffective, inappropriate, or inaccurate.</td>
<td>Efforts to communicate with students, were inconsistent, with occasional instances of ineffective communications occurring.</td>
<td>Efforts to speak and write appropriately for the learning process were effectively modeled.</td>
<td>A variety of effective speaking and writing techniques were incorporated into the lesson and were modeled for students.</td>
<td>A variety of effective speaking and writing techniques were incorporated into the lesson and were practiced by the teacher and students.</td>
<td>This behavior was not observed.</td>
</tr>
</tbody>
</table>

8. Displays a thorough knowledge of curriculum and subject matter -- The teacher supports instructional objective(s) by using appropriate examples and incorporates accurate and up-to-date information.

<table>
<thead>
<tr>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>N/O</th>
</tr>
</thead>
<tbody>
<tr>
<td>Displayed an insufficient knowledge of curriculum and subject matter.</td>
<td>Displayed a limited or incomplete knowledge of curriculum and subject matter.</td>
<td>Displayed an adequate knowledge of curriculum and subject matter.</td>
<td>Displayed a thorough knowledge of curriculum and subject matter and related concepts being taught with topics and activities introduced previously.</td>
<td>Displayed a thorough knowledge of curriculum and subject matter and could model a variety of techniques to transfer previous learning to the current lesson for others.</td>
<td>This behavior was not observed.</td>
</tr>
</tbody>
</table>
9. Ensures student time on task - - The teacher uses available class time to achieve the instructional objective(s).

<table>
<thead>
<tr>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>N/0</th>
</tr>
</thead>
<tbody>
<tr>
<td>No techniques to ensure student time on task were in evidence.</td>
<td>Techniques to ensure student time on task were generally ineffective or inconsistently applied.</td>
<td>Techniques to ensure student time on task were evident and effective.</td>
<td>Classroom procedures and rules related to student time on task were posted and consistently applied.</td>
<td>A variety of techniques to ensure student time on task were in evidence and students were clearly accountable for their learning.</td>
<td>This behavior was not observed.</td>
</tr>
</tbody>
</table>

10. Demonstrates sensitivity in relating to students - - The teacher contributes to students' self-esteem by the manner in which he/she relates to them.

<table>
<thead>
<tr>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>N/0</th>
</tr>
</thead>
<tbody>
<tr>
<td>Was unresponsive or adversarial in attending to student needs.</td>
<td>Inconsistently demonstrated sensitivity in attending to student needs.</td>
<td>Demonstrated sensitivity in attending to student needs.</td>
<td>Demonstrated sensitivity in attending to student needs and encouraged students to share interests, ideas, and concerns.</td>
<td>Demonstrated sensitivity and could model a variety of techniques to build and enhance student self-esteem for others.</td>
<td>This behavior was not observed.</td>
</tr>
</tbody>
</table>
PRODUCTIVE TEACHING TECHNIQUES

Communicates effectively with students.

For at least three lessons in the next unit the teacher will request that students evaluate the clarity of the lessons.

1. Plan the lesson.
2. Ask students to respond to teacher questions by using a "Clear Teaching Checklist."
3. Review student responses and summarize.
4. Discuss summary with evaluator.
5. Repeat steps 1-3 for subsequent lessons.

Define the time in terms of weeks, months, or other segments for each step of the procedures.

A progress check that could include formal or informal observations, a work sample, etc.

Summary of student responses.

Please check one or more of the following:
- District policy
- Building procedures
- Research-based model
- Other (Please specify)

The evaluator will compare the summary of student responses with the standard.

Fully
Partially
Not accomplished
AREA: PRODUCTIVE TEACHING TECHNIQUES

CRITERION: Communicates effectively with students.

PIC: During the next three months the teacher will evaluate his/her clarity of speech at least four times.

PROCEDURES:
1. Define criteria to use in determining clarity of speech.
2. Plan a 15-minute lesson that requires the teacher to speak frequently.
3. Teach the lesson and audiotape it.

TIMELINE: Define the time in terms of weeks, months, or other segments for each step of the procedures.

MONITORING: A progress check that could include formal or informal observations, a work sample, etc.

EVIDENCE: Audiotape of lesson.

STANDARD: Please check one or more of the following:
   District policy
   Building procedures
   Research-based model
   Other (Please specify)

APPRAISAL METHOD: The evaluator will compare the audiotape of the lesson with the standard.

INDICATORS OF ACCOMPLISHMENT:
Fully
Partially
Not accomplished
AREA: PRODUCTIVE TEACHING TECHNIQUES

CRITERION: Communicates effectively with students.

PIC: During the next semester the teacher will implement a plan for increasing the degree of variety in responses to students.

PROCEDURES:
1. Videotape a segment of classroom instruction showing interaction with students.
2. View the tape and assess the amount of variety used in responding to students.
3. Design a plan to enhance variety to responses to students.
4. Submit the plan to the valuator.
5. Implement the plan.

TIMELINE: Define the time in terms of weeks, months, or other segments for each step of the procedures.

MONITORING: A progress check that could include formal or informal observations, a work sample, etc.

EVIDENCE: Classroom observation data; plan

STANDARD: Please check one or more of the following:
District policy
Building procedures
Research-based model
Other (Please specify)

APPRaisal METHOD: The evaluator will compare the plan and classroom observation data with the standard.

INDICATORS OF ACCOMPLISHMENT:
Fully
Partially
Not accomplished
AREA: PRODUCTIVE TEACHING TECHNIQUES

CRITERION: Communicates effectively with students.

PIC: Throughout the next grading period the teacher will utilize at least three different verbal and nonverbal techniques in each lesson to provide a smooth transition between segments in the lesson.

PROCEDURES:
1. Review literature concerning verbal and nonverbal techniques.
2. Compile a list of verbal and nonverbal techniques which are appropriate for transition between segments in the lesson.
3. Develop a plan to utilize these techniques.
4. Implement the plan.

TIMELINE: Define the time in terms of weeks, months, or other segments for each step of the procedures.

MONITORING: A progress check that could include formal or informal observations, a work sample, etc.

EVIDENCE: List of techniques; classroom observation data

STANDARD: Please check one or more of the following:
- District policy
- Building procedures
- Research-based model
- Other (Please specify)

APPRAISAL METHOD: The evaluator will compare the list of techniques and the classroom observation data with the standard.

INDICATORS OF ACCOMPLISHMENT:
- Fully
- Partially
- Not accomplished
AREA: PRODUCTIVE TEACHING TECHNIQUES

CRITERION: Communicates effectively with students.

PIC: Throughout the semester the teacher will deliver instructions in a clear and concise manner.

PROCEDURES:
1. Review the literature on effective communication.
2. Outline the essentials of effective communication.
3. Develop a plan which includes these essentials.
4. Implement the plan.

TIMELINE: Define the time in terms of weeks, months, or other segments for each step of the procedures.

MONITORING: A progress check that could include formal or informal observations, a work sample, etc.

EVIDENCE: Outline of the essentials of effective communication; classroom observation data.

STANDARD: Please check one or more of the following:
District policy
Building procedures
Research-based model
Other (Please specify)

APPRAISAL METHOD: The evaluator will compare the outline of the essentials of effective communication and the classroom observation data with the standard.

INDICATORS OF ACCOMPLISHMENT:
Fully
Partially
Not accomplished
APPENDIX F. PROCEDURES AND EXPLANATION
GENERAL PROCEDURES AND EXPLANATION OF SURGENCY BIAS STUDY

1. Read modified informed consent statement to subjects.

2. Subjects complete Personal Data Form through number of days of teacher performance evaluation training.

3. Subjects move to one of two assigned rooms to view videotaped preobservation conference.

4. Subjects view videotaped Judy Oliver preobservation conference.

5. Subjects predict overall teaching performance effectiveness on bottom of Personal Data Form.


7. Review contents of packet:
   a. class description p.1
   b. lesson plan p.2
   c. summary of preobservation conference responses p.3
   d. script of lesson p.4-11
   e. Teacher Performance Rating Scale including PICs p.12
   f. graphic responses for ten teaching behaviors p.13-16
   g. Professional Improvement Commitments (PICs) p.17-21
   h. evaluation summary page p.22

8. Refer to documents and complete T.P.R.S.
   a. ten criteria
   b. overall rating
   c. PICs

9. Summarize findings

10. Subjects make suggestions, comments, corrections on forms

11. Collect materials
MODIFIED INFORMED CONSENT STATEMENT TO BE USED WITH BOB NEWSUM'S VIDEOTAPED PREOBSERVATION CONFERENCE AND SUMMATIVE EVALUATION TRAINING EXPERIMENT.

Ph.D. Dissertation
Richard P. Manatt
Advisor and Co-instructor

The following statement will be read to each participant:

"Today you will participate in an exercise in which you will observe a pre-observation conference and evaluate a lesson. Because the training places you in one of two experimental treatment-groups to test the effects, you have the right to refuse to participate in the exercise and post-exercise analysis. We hope you won't refuse, however, because the investigations are important in improving the process of performance evaluation. If you are willing to participate in this research study, please turn in your evaluation materials at the end of the training session.

Submitting the materials will be construed as a modified consent to participate.

If you do not choose to participate, simply retain our materials at the end of the training session."