Order Number 9130321

A systematic evaluation of the components of frame-of-reference training and their effects on rating error, accuracy, and individual cognitive processes

Hartog, Sandra B., Ph.D.

City University of New York, 1991

U·M·I
300 N. Zeeb Rd.
Ann Arbor, MI 48106

A SYSTEMATIC EVALUATION OF THE COMPONENTS OF FRAME-OF-REFERENCE TRAINING AND THEIR EFFECTS ON RATING ERROR, ACCURACY, AND INDIVIDUAL COGNITIVE PROCESSES

by

SANDRA B. HARTOG

A dissertation submitted to the Graduate Faculty in Psychology in partial fulfillment of the requirements for the degree of Doctor of Philosophy, The City University of New York.

1991

This manuscript has been read and accepted by the Graduate Faculty in Psychology in satisfaction of the dissertation requirement for the degree of Doctor of Philosophy.

2/28/91
Date

Roger E. Millsap
Chair of Examining Committee

3/1/91
Date

Herbert D. Saltzstein
Executive Officer

Walter Reichman, Ed.D.

Donna E. Thompson, Ph.D.

Richard Reilly, Ph.D.

Gerard Kehoe, Ph.D.

Supervisory Committee

The City University of New York

Abstract

A SYSTEMATIC EVALUATION OF THE COMPONENTS OF
FRAMEOFREFERENCE TRAINING
AND THEIR EFFECTS ON
RATING ERROR, ACCURACY,
AND INDIVIDUAL COGNITIVE PROCESSES

by

Sandra B. Hartog

Adviser: Roger E. Millsap, Ph.D.

Research in the performance appraisal field has been moving away from issues that are most salient for organizational use and toward a more theoretical investigation of underlying processes in performance ratings. In this study, Frame-of-Reference training (FOR) was investigated as a technique that can be directly applied to increasing the effectiveness of organizations' performance appraisal systems. FOR training was chosen because its methodology exemplifies many of the separate elements employed in performance appraisal training programs, and thus the results may be generalizable to other research in this area. This study was also undertaken to clarify many inconsistencies in this research paradigm, including the generalizability of effectiveness indices, operational and

conceptual ambiguities across training programs, and elements in the research design itself (e.g., topic and order effects of videotaped performance vignettes).

The components of FOR training, as originally conceptualized by Bernardin and Buckley (1981), were systematically investigated to determine the differential influence of each on error and accuracy effectiveness indices. Trainees were 250 college students across 12 intact classes. Two classes were randomly selected for each of six training conditions. The difference between conditions was the omission of one training component each. The effectiveness indices were measures of leniency, interrater reliability, two indices of halo error, elevation, differential elevation, differential, stereotype, distance and leniency accuracies, and a measure of absolute value halo.

The results indicate that the relationship between the effectiveness of the various components involves an interplay between the types of rating judgements required, the integration of similar or different types of information, and the underlying cognitive strategies promoted by each component. When FOR training fares best it appears due to the feedback of expert true scores and behavioral rationales, along with participant discussion. When FOR training leads to inaccuracy, different combinations of components differentially influence the accuracy

indices. The results also suggest that the choice and presentation of stimulus materials are very important design considerations. Additional results demonstrate that: rater attitudes are related to accuracy and dimension ratings; there are differential relationships between identical indices of observation and evaluation accuracies; and, the different evaluation accuracy indices do not generalize across measures.

Contributions of this project are discussed, as are limitations. Numerous suggestions for future research are offered.

ACKNOWLEDGEMENTS

It seems that the majority of my graduate training has been spent fighting deadlines and working late into the night. There is poetic justice that it is now the eleventh hour, and I am in the hot pursuit of a 24 hour copy shop. However, this project cannot be completed without the acknowledgement of some very special individuals.

I, unlike many students, had the good fortune to emerge from my graduate training with only pleasant memories and a feeling of competence. I credit this positive experience to the faculty members of the CUNY Graduate Center subprogram in Industrial/ Organizational Psychology. As a whole, they are amongst the most supportive, nurturing, and educative group I have ever had the pleasure of knowing. Their doors were always open. Specifically, I would like to express my gratitude to Joel Lefkowitz who saw clear to inviting me in and making sure I had a good seat.

Walter Reichman was instrumental in my training from early on. He was always willing to share his vast store of knowledge that went far beyond classroom education. He offered encouragement when I was down and advice when I needed it. By providing a view of his life I knew what I wanted. I look forward to continuing to work with him.

Donna Thompson was a godsend. Her insight into all aspects of the research process provided much needed guidance. Her feedback was always accurate and delivered without any punches. Her encouragement and emotional support always made me believe this would someday end and she would be there cheering. Her voice will always echo in whatever work I produce.

The hardest person to thank is Roger Millsap. He was, and always will be, my mentor and my friend. From him I learned to love the research process and view it as an intellectual challenge. He taught me to listen to what the data was saying and how to make it tell more. He suffered through my moods uncomplaining and never made me feel that I was asking too much (even when I am sure I was). I owe him more than I can ever express.

If it were not for my friends and colleagues, Mark Kornspan, Margaret Geoghan, Lynn Gracin, Lori Poveromo and many others, the trip would not have been as fun. Their support was immeasurable; they were always there whether I wanted them to be or not.

To my family I can finally say "I'M DONE!" While you may have never understood why I did what I did, or for that matter what I did, you were behind me 100%. Since I was a child you encouraged me Thank you.

Howard Epstein, my husband, best friend and colleague is last in this list because for the past year he was willing to

take this position. He allowed me the time and means to pursue this goal and only rarely complained. He guided me through the rocky spots and glowed with my achievements. He provides the foundation of my life and without him all accomplishments would be hollow.

TABLE OF CONTENTS

TABLE OF CONTENTS
(continued)

TABLE OF CONTENTS
(continued)

TABLE OF CONTENTS
(continued)

TABLE OF CONTENTS
(continued)

TABLE OF CONTENTS
(continued)

TABLE OF CONTENTS
(continued)

TABLE OF TABLES

TABLE OF TABLES
(continued)

CHAPTER I

INTRODUCTION

Background

The origin of performance appraisals can be traced back to World War I. It was then that Walter Dill Scott convinced the U.S. Army of the benefits of a performance evaluation method he had developed for use with military officers. After the war, performance appraisals came into vogue as a means for evaluating hourly employees. In this type of system a rational wage structure was instituted wherein in-grade wage increases could be based upon merit. "Merit" was evaluated vis-a-vis a performance appraisal system, then termed a "merit-rating system."

With the growing sophistication of corporate America in the 1950's, industry placed great interest upon developing a systematic mechanism for evaluating the performance of its technical, professional, and managerial staff. It was felt that a rigorous evaluation system would provide insight into the strengths and weaknesses of staff. This insight could then be used for training, development, and advancement. The systematic evaluation of an employee's job relevant behavior used to determine effectiveness at work, and future potential, came to be known as a performance appraisal. The emphasis was now directed toward employee development and away from merit rating.

Throughout the decades since World War I, research on performance appraisals has occupied one of the largest areas of interest for Industrial/Organizational psychologists as evidenced by the 100's of published articles in professional journals. Scientists very often use performance appraisal information as a criterion measure in test validation (Landy & Trumbo, 1980) and other laboratory research. Organizations use performance appraisals for salary administration, performance feedback, and employee development (Cleveland, Murphy, & Williams, 1989), to determine employee training needs (Levine, 1986), and for promotion and transfer decisions (Campbell, Dunnette, Lawler, & Weick, 1970). According to Bernardin and Villanova (1986) 90% of American organizations have some type of performance appraisal in their overall human resources effort. In a recent survey of 106 members of the Society for Industrial and Organizational Psychology who were employed in private industry, all but four reported that their organization had a formal performance appraisal system (Cleveland, Murphy, & Williams, 1989).

Given that performance appraisal systems hold a unique position between applied and pure research they are open to inquiry and evaluation from many different directions. Through this widespread investigation they have continued to be fined-tuned and modified. Many extensive literature reviews exist (e.g., Landy & Farr, 1980; Saal, Downey, & Lahey, 1980; Smith, 1986; Spool, 1978) documenting their progression. The literature

reviews serve to consolidate large bodies of research, discriminate trends, and reveal areas needing further investigation. However, in an effort to consolidate and generalize information these reviews may unintentionally lead to a false sense of confidence in what initially appears to be consistent or interlocking patterns of results. The lack of generalizability across research has been noted by other reviewers (i.e., Smith, 1986) as has the lack of external validity important for organizational use (i.e., Banks & Murphy, 1985; Bernardin & Villanova, 1986). One goal of the present review is to identify inconsistencies in this area of research and to suggest how they may be contributing to ambiguity in how to apply these results and ambiguity in identifying elements attributing to results.

There has been a call in the literature to narrow the gap between researchers and practitioners in the performance appraisal field (Banks & Murphy, 1985). While the field has experienced many advances, research is moving away from issues that are most salient for industrial use (e.g., determining the best methods for selecting valid and appropriate data for different appraisal purposes) and toward a more theoretical or abstract investigation of the underlying psychological and cognitive processes in performance ratings. What is missing is the investigation of performance appraisal techniques and

procedures that can be directly applied to increasing the effectiveness of organizations' performance appraisal systems.

For this research to be considered helpful and to be integrated by an organization, we need to concentrate on the bottom line. That is, we need to identify specific elements of the performance appraisal system that can be improved through research and thus, in this manner, have a direct instrumental utility on applied uses. The area of rater training in increasing the accuracy of performance evaluation judgements provides an ideal opportunity to meld research aims of both the practitoner and the scientist. Unfortunately, practitioners have not accepted as useful many of the conclusions drawn from training research and therefore have not integrated the results into their systems.

It is suggested that this lack of integration may have failed to occur for a variety of reasons. There are many training programs that appear in the performance appraisal literature. Four general orientations exist. There are those that focus on rating error, those that teach observational strategies, those that differentiate dimensions and standards of behavior, and those that involve intensive orientation to performance appraisal ratings through group discussion of rating errors and practice and feedback. However, the process and content of these general types of training programs are often blurred so that the conclusions do not differentiate which

elements contributed to the results (Smith, 1986). Often it is unclear whether it was the training method (i.e., feedback) or the orientation and intent of the training (i.e., differentiating standards of behavior) that resulted in correct performance ratings.

No one best program has yet to emerge. While there has been a steady progression of developments, consistent training paradigms or an agreed upon set of variable operationalizations have been lacking. Many of the research designs lack adequate control or comparison groups. The instruments used to measure the dependent variables are applied inconsistently. Elements of the measurement situation itself that may be influencing outcomes, are not investigated. Methodologies and operationalizations often lack generalizability across studies.

Thus, we are not offering practitioners a clear path to the application of this body of research. It seems appropriate that we take the time now to consolidate this area of research and differentiate and assimilate effective and ineffective methods and processes (content) of performance appraisal training. The basic orientation of this project is to present a systemmatic approach to assessing the effectiveness of performance appraisal training which examines many of the confounds of previous research. The goal is to identify specific components that effect different types of rating judgements and to suggest how the linkage between the two might differentially influence the effectiveness of various applied purposes (i.e., salary

increases, promotion, training and development, group ratings) of performance appraisals. Thus, another purpose of this research is to determine which components contribute or detract from different rating abilities and to explore the best combination of training techniques for different rating purposes.

Frame-of-Reference (FOR) training (Bernardin & Buckley, 1981), which attempts to identify dimensions and standards of behavior, was chosen as the focus of study in this research since in its methodology it exemplifies many of the separate elements of the general performance appraisal training programs. Furthermore, in a recent review of the performance appraisal literature FOR training appears to enjoy positive results in relation to increasing rater accuracy (Murphy & Cleveland, 1990). The FOR training model, as proposed by Bernardin and Buckley (1981), is developed in such a manner that each of the six elements of the program can be separated from the larger whole, and therefore, the contributions of each component in terms of method and content can be individually evaluated. Another purpose of this study is to provide results which will generalize to past and future research by identifying the individual training components and processes that have a causal effect on performance appraisal ratings.

A fourth purpose of this project is to investigate elements of the laboratory research design itself to determine whether there are aspects of the design, apart from the training, that

are influencing results. The design in this study contains many aspects that are common across the performance appraisal training research, for example the use of videotaped performance vignettes, the rating scales, and different topics presented in the vignettes. These elements have not been investigated specifically in the performance appraisal training research in order to determine their effect on training outcomes. It is the belief here that these variables may be confounding the demonstration of clear trends, or at the very least contributing to the outcomes. The identification of these variables may lead to greater clarity regarding the effect of the training methodology.

A final focus of this investigation is to compare results across the various operationalizations of error and accuracy that exist. It has been demonstrated in the literature that the operationalizations of error and accuracy indices are not always comparable across measures (Fisicario, 1987; Becker & Cardy, 1987; Murphy & Balzer, 1986). The suggestion has been made that the various operationalizations may actually be measuring different rater abilities. This seems to have direct implications for the different processes and content of FOR training components. The investigation of the pattern of results for the different effectiveness indices may yield some meaningful information related to the contribution of each training component to increases in different types of rating

judgements. The pattern of results may also aid the understanding of how increasing one type of rating judgement affects another.

The utility of the different aspects of this research is to gain some clarity in the performance appraisal training literature. By doing such it is hoped that the results will offer practitioners a clear path to selecting different methods of training dependent upon the purposes for appraisal use. It is also hoped that this project will serve as a guide for researchers in issues to consider when conducting training research.

CHAPTER II

PERFORMANCE APPRAISAL TRAINING

As stated in Chapter I, many comprehensive reviews of the performance appraisal research exist (e.g., Ilgen & Feldman, 1983; Landy & Farr, 1980; Saal, Downey, Lahey, & Farr, 1980; Smith, 1986; Spool, 1978). The purpose of this chapter is not to provide yet another review, but rather to point to areas of the research that are in need of greater clarity and which, through their discrimination, suggest areas of future study. This review is limited primarily to issues involved in the pursuit of developing and evaluating rater training programs. It covers all the published research in this area from 1952 to the present, specifically addressing areas of concern in performance appraisal training for performance evaluation. Other aspects of the performance appraisal process which include a) rater and ratee interactions, b) the rating format, c) the rating context and, d) the results of the ratings (Landy & Farr, 1980), are beyond the scope of this review except for where they intersect with research on training.

The structure of this chapter is to first provide a brief overview of the history of performance appraisal training to demonstrate the direction of developments from the initial idea of rater training to the more recent emphasis on different training techniques and the processes underlying these methods.

Second, a content analysis of the research is presented to document similarities and differences across the research that suggest questions about the generalizabilty of results. Third, psychometric indices of rating effectiveness are presented and discussed. The stages of this review lead to the rationales and hypotheses which are the basis of the current study.

History

As indicated in Chapter I performance appraisal research spans decades. Yet, the majority of research that concerns training raters in conducting performance appraisals is less than fifteen years old. Prior to 1975, and the now landmark study conducted by Latham, Wexley & Pursell (1975), training raters in the use of performance appraisals was an oddity. In fact, the first research conducted on performance appraisal training (i.e., Levine & Butler, 1952) while successful, stood apart from other avenues of interest for over 20 years. During this time the performance appraisal literature was filled with acknowledgment and concern for different psychometric errors committed in the pursuit of judgmental measures in general (Spool, 1978), and performance appraisal in particular (Landy & Farr, 1980). In general, these errors were pursued through the investigation of individual difference variables, situational effects, and the manipulation of rating formats. The idea of training raters to alleviate errors in judgment was an idea yet to come. Even today much of

the research seems to consider training more of an experimental manipulation than a necessity. In spite of the lack of attention there has been a steady, albeit slow, progression of research on performance appraisal training.

Length of training. Initially, rater training was pursued in an attempt to eliminate rating errors such as halo, leniency, contrast effects, unreliability, and invalidity (Borman, 1975; Borman & Dunnette, 1975; Brown, 1968; Latham et al., 1975). Early efforts in this regard resulted in the development of training programs generally of short duration, about 5 minutes to 15 minutes (Borman, 1975; Brown, 1968; Thorton & Zorich, 1980; Vance, Kuhnert, & Farr, 1978). With these brief programs Borman (1975) and Brown (1968) demonstrated a reduction in halo, yet Borman failed to find an increase in the validity or reliability of the performance measures.

It was proposed that more extensive and intensive training was necessary to induce change in these psychometric errors (Ivancevich, 1979; Latham et al., 1975). Researchers began investigating the success of longer training sessions. Latham et al. (1975), Bernardin (1978), Borman (1979) and Ivancevich (1979) were among the first to extend training to between 1 to 14 hours. In general, they found that longer training programs were successful in reducing halo and leniency. It appeared that longer training sessions were better able to bring about a reduction in judgment errors, however, the incremental value of

the additional training time was not investigated. Thus, there was no knowledge of whether the reduction in error had utility when contrasted with the time, cost, and effort required for these expanded training programs.

Duration of training. While training seemed effective in decreasing error immediately following the training sessions, the duration of the effect was still unknown (Latham et al., 1975). Four studies investigated this issue with conflicting results. In one study, Latham et al. (1975) found that training effects remained after a six month period, yet this was only in the most intense of three training groups. Conversely, three other studies failed to support these results (i.e., Bernardin, 1978; Ivancevich, 1979; Warmke & Billings, 1979), and found that differences in the degree of error among training conditions dissipated over time. The suggestion was made that refresher courses would be helpful to reacquaint trainees with potential errors in performance appraisal ratings (Ivancevich, 1979).

Training techniques. With the demonstration of training as an effective way to reduce rater errors, training techniques were emphasized. Researchers began comparing different techniques (i.e., Bernardin & Walter, 1977; Ivancevich, 1979; Latham et al., 1975; Warmke & Billings, 1979). Research spanned many different ideas including: investigating the differences between the intensity of training programs - for example lecture, discussion, or workshop methods; methods designed for more active trainee

involvement such as participation in scale construction or diary keeping; increasing raters familiarity with the scales before rating behavior; or, simply encouraging careful observation and note taking. The goal was to reduce error. However, the implicit rationale was that error reduction could be accomplished through facilitating a deeper level of understanding or cognitive processing of the relevant information. In general, results demonstrated that the more intense or involved the training, the greater the rating error reduction.

Rater accuracy. A focus on rating accuracy arose midway through the investigation of training techniques as a method to reduce error. Borman (1975) made an important point regarding the potential "side effects" of reducing psychometric error, suggesting that this reduction could lead to a parallel reduction in the reliability and/or validity (accuracy) of the performance appraisal ratings. Previously error was considered the main element contributing to inaccuracy in performance ratings therefore the emphasis in training was to reduce error by reducing the amount of halo, leniency, or range restriction in ratings. The assumption was that eliminating these from the ratings would raise the accuracy of the ratings, since the ratings would then be closer to the true performance levels exhibited by ratees. Borman suggested that this assumption may have been incorrect. He reasoned that eliminating psychometric error may actually decrease the precision of ratings. For

example, cautioning against similar ratings across dimensions may decrease halo, yet multiple areas of factual strengths or weaknesses would be eliminated. In actuality the consistency in ratings across dimensions may be accurate representations of the true performance levels. Thus, these psychometric indices were not truly "errors" in the sense of being wrong or incorrect ratings.

Borman proposed that what was needed was to compare trained raters' scores against some known criterion value or a "true score". Towards this end he introduced the notion of an "accuracy" score or "true" score which could serve as a predetermined or manipulated measure of a ratee's behavior. The rationale here is that if there is a baseline of accuracy or a known "true" score, there is then a known standard with which to measure training success.

This methodology was used in two studies (i.e., Borman, 1975; 1979). Six vignettes were constructed which manipulated ratee behaviors. "Experts" in the field of performance appraisal rated the vignettes and an average of these ratings became the "true" performance scores with which to judge trainees against. Borman's implementation of true scores was seized upon by the majority of researchers in the area and, until late, has been considered the accuracy evaluation methodology of choice.

The incorporation of an accuracy measure of effectiveness has brought interesting conclusions. It appears that when assessing accuracy, as opposed to assessing a reduction of

psychometric error, accuracy of observations or judgments is not always increased through training. A decrease in error may not lead to a concomitant increase in accuracy. In an effort to explain the conflicting results between accuracy and error, Bernardin and Pence (1980) suggest that training programs may be doing nothing more than teaching raters how to replace one erroneous response set, or response distribution, with another. Bernardin and Buckley (1981) suggest that "... what raters are learning during training on psychometric error is a response set of low mean ratings with low intercorrelations across dimensions (p. 206)." This does not necessarily correspond to more accurate ratings, especially when one considers that most jobs are comprised of interrelated tasks, knowledge, skills and abilities. It is more likely the case that performance ratings would be intercorrelated. For example, the more seasoned employee should receive ratings at the high end of the scale across dimensions of behavior.

Rater accuracy training. Based on this line of reasoning, Bernardin and his colleagues suggested a training program that led raters to be better observers. Initially calling this technique Rater Accuracy Training (RAT) they proposed that training should focus on the dimensionality of jobs, the importance of accurate evaluations, and the development of stereotypes of effective and ineffective behavior (Bernardin & Pence, 1980). In this manner, typical rating errors would be discussed through the use of behavioral example and attention

would not be focused on the psychometric representation of errors. The rater would be given an overview of the criticality of the rating process and encouraged to be a more observant and accurate rater. The underlying goal here is to avoid merely the replacement of one incorrect response set with another. Unfortunately, initial research on this program yielded ambiguous results (Bernardin & Pence, 1980). While RAT resulted in greater accuracy than traditional error training, using Borman's (1975) expert true score methodology, it was no more accurate than the omission of any training. Additionally, traditional error training reduced psychometric error yet had lower accuracy scores than a RAT and control group.

In a later article, Bernardin and Buckley (1981) formally proposed what they now called Frame-of-Reference training (FOR). This training consisted of 6 steps and involved the active involvement of the trainees in developing a common frame of reference of appropriate and inappropriate behavior for a particular job, and the identification of different levels of behavior through practice, feedback and discussion. Investigations of this training method demonstrated increases in accuracy indices and decreases in error indices (Athey & McIntyre, 1987; Hedge & Kavanagh, 1988; McIntyre, Smith, & Hassett, 1984).

More recent training research has concentrated on comparisons of different training techniques designed to enhance observation or decision making skills and to investigate

differences in these approaches with respect to accuracy and error (Hedge & Kavanagh, 1988). In addition, the effect of motivation on performance judgments (e.g., McIntyre et al., 1984) has also received attention. Whereas Zedeck and Cascio (1982) demonstrated results indicating differential rating effects when perceived purpose of performance ratings were manipulated, McIntyre et al. (1984) found only a slight effect (less than 5% of the variance) on error and accuracy measures.

Training and scale format. The 1970's produced a number of studies, separate from training issues, comparing the superiority of one rating scale over another. Unfortunately, this line of inquiry failed to demonstrate firm conclusions (Kavanagh, 1982; Landy & Farr, 1980). In the mid 1980's this issue was reapproached within the training literature. Lee (1985) and Pulakos (1986) suggested that there was an interaction between performance appraisal scale format, and thus the rating task, and training. Pulakos (1986) found that accuracy was greater for congruent rater training and performance appraisal scale format. For example, she demonstrated that observational training was more effective with an observational rating format and evaluative training was better with an evaluative rating format. This line of research, which also includes Athey and McIntyre (1987), can be viewed within the increasing spectrum of research with a cognitive orientation. The implicit or even explicit assumption driving this work is that rater training and appraisal format

can be tied to different levels of cognitive processing facilitating the evaluations.

Summary and conclusions. From this historical overview of the performance appraisal training research some general conclusions can be drawn. First and foremost it appears that training individuals in performance appraisal judgments of any type is advantageous in both laboratory and organizational settings. The only study not to find a positive effect of training when operationalized as a reduction of error or an increase in accuracy, was Vance, Kuhnert, and Farr (1978). However, this study differed most from others in that the rating stimulus was comprised of tape recorded interviews as opposed to the more common stimuli of videotapes or live individuals. Additionally, the training was a brief written form of error training. A second finding is that error training appears effective in reducing psychometric error. This training has the most consistent effect upon reducing halo. Third, training has differential effects depending upon the intended purpose of training or the dependent variables measured. It has been demonstrated that a reduction in psychometric error need not accompany an increase in accuracy. Often the same types of training that reduce psychometric error decrease accuracy. There is also the suggestion of an interaction between training and rating scale format, such that the goals of training (i.e., greater observation, awareness of the multidimensionality of

behavior) are best assessed by a scale that is oriented towards those goals (i.e., behavior frequency, behavioral dimensions). Furthermore, the longer the duration of the training program, and therefore the more involved the program, the greater effect it has upon psychometric error. Unfortunately, while results are mixed, training seems unable to sustain effects over time.

Content Analysis of Studies

The generalized results presented above need to be viewed within a clear description of the entire body of performance appraisal training research, since it is the belief herein that inconsistencies across research studies confound what could be more clearly established results in the literature. In this section a content analysis of the training studies is presented in order to demonstrate the similarities and differences in this body of research. Table 1 provides an overview of the 19 studies in the published research that specifically pertain to performance appraisal training. This analysis is not presented in order to refute accepted conclusions, but rather to point to the need for greater clarity of elements of the research paradigms and to direct attention to specific issues that need to be assessed in relation to their influence on training results.

Table 1

Content Analysis of Training Studies

Researcher	Brown (1968)	Borman (1975)	Latham, Wexley, & Pursell (1975)
Subjects	120 Nurses	90 Managers	60 Managers
Stimulus	Own subordinate nurses	Written vignettes of first line supervisors	Videos of job candidates
Scale	Trait	BARS	Overall ratings
IV	Training, trait ratings, relationship between rater & ratee	RET	Workshop, group discussion, & control group training
DV	Halo	Halo, interrater reliability, validity	Contrast, similar-to-me, 1st impression
IV	2x2 post-test only	Counterbalanced one group pre/post	Post-test only control group
Results	Trained raters less halo, no effect for intensity of relationship	Decreased halo across ratees, lower reliability, validty unchanged	Workshop group had no errors, control group had all errors but 1st impression, discussion group had 1st impression error

Table 1 continued

Researcher	Bernardin & Walter (1977)	Bernardin (1978)	Vance, Kuhnert, & Farr (1978)
Subjects	156 Students	80 Students	125 Students
Stimulus	Actual Instructors	Actual Instructors	Taped interview of job candidates with differing performance levels
Scale	BES	BES or Summated	Graphic & behavioral scales
IV	Amount of training, scale exposure, error knowledge	Longitudinal rating, comprehensive or abbreviated training, scale type	Training, scale type
DV	Leniency, halo, interrater reliability, discrimination across ratees	Error knowledge, halo, leniency	Accuracy, halo, leniency, intraclass correlation, confidence score
Design	Posttest only control group, some measures 10 weeks delayed	Pre/Post control group	Post-test only
Results	Group with most training and scale exposure had least halo & leniency, & most interrater reliability	Comprehensive group best at Time 1 on errors, abbreviated group better than control at Time 1, no difference over time	No effect for training, behavioral scale was better psychometrically and had better accuracy

Table 1 continued

Researcher	Borman (1979)	Warmke & Billings (1979)	Ivancevich (1979)
Subjects	123 Student	52 Nurse supervisors	66 Engineering supervisors
Stimulus	Video of manager counseling employee	Own staff nurses	Videos of above & below average performers
Scale	BARS, behavioral summary, trait scale, numerical rating, & personal characteristics	Graphic ratings and present organizational method	BES
IV	RET vs no training, scale format	Lecture, discussion & control group; experimental vs administrative ratings	Intense, discussion & control group training
DV	Halo, validity & accuracy	Halo, central tendency, leniency & interrater reliability	Halo, leniency
Design	Posttest only	Posttest only control group	Pretest with multiple posttests, & control group
Results	Training reduced halo, no effect on accuracy, inconsistent scale effect	Experimental rating group with scale construction training reduced halo & central tendency, lecture group reduced central tendency, discussion group increased central tendency; administrative ratings had more halo	Intense group reduced error, discussion group had less halo than control group, no difference on leniency; effects dissipated over time

Table 1 continued

Researcher	Thorton & Zorich (1980)	Pursell, Dossett, & Latham (1980)	Bernardin & Pence (1980)
Subjects	170 Students	6 Supervisors	72 Students
Stimulus	Video of leaderless group discussion	Own subordinates	Written vignettes
Scale	BOS	BOS	BES
IV	Behavioral observation, behavioral observation & error training	Workshop training	RET, RAT & control group training
DV	Knowledge test	Validity coefficient	Leniency, halo, & accuracy
Design	Posttest only control group	One group pre-post-test	Post-test only control group
Results	Highest accuracy obtained through behavior observation & error training	Increase in validity coefficient	RET group had lowest leniency & halo but less accuracy; no significant difference between RAT & control

Table 1 continued

Researcher	Zedeck & Cascio (1982)	Pulakos (1984)	Davis & Mount (1984)
Subjects	130 Students	108 Students	402 Managers
Stimulus	Written vignettes	Videos of manager counseling employees	Own subordinates
Scale	BARS	BARS	None
IV	Training; purpose (merit raise, development & retention)	RET, RAT, RET/ RAT, & control	Computer Assisted Instruction (CAI), CAI with behavior modeling, & control
DV	Accuracy, discrimination between raters, policy rating	Halo, leniency, accuracy	Managerial learning, managerial job performance, appraisal discussion effectiveness
Design	Posttest only control group	Posttest only control group	Posttest only control group
Results	Purpose had a significant effect, no difference in discrimination between raters	No relationship between error & accuracy; RET & RET/RAT reduced halo; RAT & RET/ RAT reduced leniency; RET & control group equal on leniency; RAT most accurate	Trained managers more knowledgeable; employees were more satisfied; no significant differerence between trained & untrained for quality of documentation

Table 1 continued

Researcher	McIntyre, Smith & Hassett (1984)	Pulakos (1986)	Athey & McIntyre (1987)
Subjects	164 Students	144 Students	108 Students
Stimulus	Videos of instructors	Video of manager talking with subordinate	Videos of instructors
Scale	BOS	BOS, BARS	BOS
IV	RET, FOR, FOR/RET, & control; rating purpose	Evaluative, observational & control group training; rating task	FOR, INFO only & control group; group size
DV	Accuracy, leniency, halo	Accuracy	Retention of training, reten- of pretraining knowledge, accuracy, halo, leniency, & arousal
Design	Posttest only control group	Posttest only control goup	Posttest only control goup
Results	FOR most accurate & reduced most 'true halo; RET reduced most halo; no strong effect for purpose	Accuracy was greatest for convergent training & rating task	FOR improved reten- tion of training & had greater distance accuracy and less halo; group size effected pre- training knowledge only; INFO group retained more infor- mation than control group but was no less accurate

Table 1 continued

Researcher	Hedge & Kavanagh (1988)
Subjects	52 Supervisors
Stimulus	Video of manager with problem subordinate
Scale	None given
IV	RET, observational, decision-making, & control group training
DV	Leniency, halo, range restriction, accuracy, & attitude measures
Design	Pre/Post control group
Results	Only RET reduced leniency & halo, observational group had increased halo & more range restriction than other groups; DM group had no effect on error; RET had less accuracy & DM group had most; no difference between observational or control groups

Rating scale format. As mentioned, the performance appraisal literature is rich with investigations of the psychometrically superior rating scale (Landy & Farr, 1980), but none have yet been discovered. Rating scales employed in the training research vary widely. Four studies used a graphic rating scale, five used behaviorally anchored rating scales, four used behavioral observation scales, five used behavioral expectation scales, two used summated ratings, five studies manipulated scale type as one independent variable, and two studies either did not employ a rating form or did not indicate the type of rating scale utilized. Herein lies a problem. Due to the great diversity in rating scales it is difficult to generalize across training results. When the rating formats differ it becomes unclear whether the same results would have been found with a different scale, or conversely, that the results are scale dependent. The different formats require different types of inferences or judgements from the raters. Additionally, the training programs teach different behaviors. These behaviors may or may not be related to the subsequent rating judgements trainees are asked to provide.

Often researchers have, unintentionally, compared training programs (e.g., RET and RAT, or RET and decision-making) which require different cognitive tasks against a single rating scale. The rationale behind many of the training programs is implicitly or explicitly in support of various different aspects of

cognitive processes. Yet, the vehicle that is used to gather the dependent measures, the rating scales, may not be consistent with the rationale behind the training or the cognitive processes investigated. This methodological confound is particularly salient given Pulakos' (1986) research which supports the idea that training congruent with the rating task is more effective in increasing accuracy. Other evidence supporting the identification of this inconsistency as a problem can be drawn from those studies that found divergent results across scale types (i.e., Vance et al., 1978; Pulakos, 1986), and from comparison between Bernardin and Pence (1980) and Pulakos (1984) where divergent results between errors and rater accuracy training were found. Pulakos (1984) using a BARS format found that rater accuracy training reduced leniency whereas Bernardin and Pence (1980), using a BES format, found no difference for leniency between rater accuracy training and a control group.

This lack of format consistency may result from several factors. First, stimulus materials are borrowed and loaned frequently in this area of research. As such, if one researcher relies on the stimulus materials of another then the scale format may only be a consideration of convenience and not necessarily linked with the goals of the training. Therefore, while there may be consistency in scale use across studies, they are likely being used inconsistently with the training content. Secondly, given the inconclusiveness of research seeking a superior format,

the choice of rating scale may not appear important unless there are explicit hypotheses concerning this variable such as in Pulakos' (1986) research. Third, the development of rating forms is extremely burdensome in both time and money. Therefore, in field research the use of a pre-existing scale specific to the host organization may seem more expedient but again, may not be linked to the training or may be idiosyncratic to the organization. Or, it may not be feasible to introduce a new performance appraisal form in an organizational setting.

While these real opportunities to use existing scales remain and constraints over developing new scales exist, this leads to potential confounds. It is not possible to compare training outcomes across studies when the rating tasks or judgmental processes are inconsistent with one another. In synthesizing this literature it becomes apparent that many studies incorporate or model the training after Latham et al.'s (1975) workshop group, or Borman's (1975) error training program. Whereas the original studies used one type of rating format, other studies either do not indicate the scale type used (i.e., Hedge & Kavanagh, 1988; Warmke & Billings, 1979), or if different from the original they do not necessarily provide a rationale in support of the difference, which makes one wonder if they ever gave the variable careful consideration. Additionally, many studies employ RAT or FOR training, and use behavioral observation or behavioral expectation scales, and/or even a

knowledge test as vehicles for gathering the data. An attempt is then made to generalize across these disparate operationalizations of the stimulus materials. There can be no comparison between such diverse studies unless this diversity is dealt with in some systematic and exploratory manner.

Design. The majority of the studies (14) collected only posttest measures and four studies were comprised of a single group without any control or comparison group. Twelve studies compared differences between types of training, whereas in five studies training was compared against a no training condition. Thus, the problems here are the omission of a control group in some of the designs, appropriate comparison groups, or when neither groups are present the collection of posttest measures without any known true performance score. Therefore, it is difficult to rule out threats to internal validity. With the exclusion of a control group it is difficult to know whether the results occurred due to multiple testing when a pretest-posttest design was used (e.g., Borman, 1975; Pursell et al., 1980), or that practicing with the scales resulted in change, or whether the results occurred explicitly due to the training. In fact, results in this area have demonstrated situations where the control and experimental groups were no different on measures of quality of documentation (Davis & Mount, 1984), accuracy (Athey & McIntyre, 1987; Bernardin & Pence, 1980), leniency (Ivancevich, 1979; Pulakos, 1984), or range restriction (Hedge & Kavanagh, 1988).

Another inconsistency is the timing of the posttests. In some designs the longitudinal effects of training are specifically explored through either a delay in ratings or multiple posttests (i.e., Bernardin, 1978; Bernardin & Walter, 1977; Ivancevich, 1979; Latham et al., 1975). However, in other studies an unintended consequence of their design is a delay in ratings (i.e., Borman, 1979; Davis & Mount, 1984; Pursell et al., 1980; Warmke & Billings, 1979). The length of time between training and subsequent ratings have been anywhere from one week to six months. While not acknowledged by the investigators, performance ratings results gathered immediately after or during training may not be comparable to ratings gathered after a separation of time from training. For example, in the former case results may be capitalizing on recall, and in the latter results may be suffering from decay. In fact, longitudinal studies, with the exception of Latham et al. (1975), have found that training effects do not persist over time.

Materials. The rating stimulus, the performance examples to be appraised, are also varied throughout the literature. Across studies it can be seen that six studies were conducted with real ratees and the remaining 13 studies used either video, audio, or written ratee vignettes. Most often, behavioral vignettes are used. These vignettes are either written (i.e., Bernardin & Pence, 1980; Vance et al., 1978; Zedeck & Cascio, 1982) or videotaped with actors playing the different roles (i.e., Athey &

McIntyre, 1987; Borman, 1975; Borman, 1979; 1975; Hedge & Kavanagh, 1988; Ivancevich, 1979; Latham et al., 1975; Pulakos, 1984; Pulakos, 1986; McIntyre et al., 1984; Thorton & Zorich, 1980). The behaviors are reflective of instructors presenting a lecture, supervisors talking with subordinates, or individuals involved in a leaderless group discussion. Furthermore, many studies that use videotaped vignettes share these videos across studies.

The rationale for using videos is to be able to control and manipulate performance presented to trainees. Apart from the increased rigor it provides and the control of possible confounds, videotaped performances also allow for the generation of "true" scores to be utilized as accuracy measures. However, within those studies that use vignettes most require trainees to apply their training directly to vignette ratings. Whereas some studies that use videotaped supervisor/subordinate interactions distance the rating process and task by asking trainees to predict what the actor/supervisor's ratings of the actor/subordinate would be (i.e., Borman, 1979; Latham et al., 1975). Additionally, through the use of either type of vignettes behavior is only viewed at one time and only behavior relevant to the rating situation is presented.

Field studies use the subject's own employees as stimuli to judge the effectiveness of the training programs (Bernardin, 1978; Bernardin & Walter, 1977; Brown, 1968; Davis & Mount, 1984; Ivancevich, 1979; Pursell et al., 1980; Pursell et al., 1980).

Unfortunately, across the literature field studies provide criteria of success (e.g., accuracy and error) that are operationally different from lab studies but are often conceptually similarly. In the case of field studies accuracy cannot be assessed unless it is operationalized as a cognitive measure (i.e., a knowledge test). Halo and leniency can only be discerned in a limited sense since it actually may be capturing true variance as opposed to error variance. Thus, the dependent measures arrived at through the use of these divergent materials are not necessarily comparable.

Setting. Comparison of results between contrived manipulations in a laboratory setting and actual ratings in an applied setting are difficult. Four studies that used actual ratees were conducted in the field, whereas two field studies were conducted under laboratory-like conductions. The remaining 13 studies were conducted in the laboratory. In the latter situation subjects are given a very narrow environment, presented with a scenario specifically designed for a certain effect, and their time is spent only on the task at hand (Banks & Murphy, 1985). In an applied setting the appraisal is not the only task at hand and the very nature of the appraisal situation, and behavior being rated, is taking place in a very "noisy" and diverse environment. Additionally, motivational factors differ between the laboratory and the field, as do the process issues. In a field situation there is an ongoing relationship between the

rater and ratee that will continue after the appraisal as well as generally require feedback of the appraisal information. This aspect of the process is missing in laboratory research. Unfortunately, the majority of the performance appraisal research takes place within a contrived laboratory setting. Even at a very basic level it is difficult to generalize across these different settings.

Purpose. Performance appraisals by their very design are intended for organizational or administrative use. It can be assumed that the intended and actual purpose of a performance appraisal in an applied setting is different than that used in a research setting. Hand in hand with purpose it would seem necessary to consider the rater's motivation for completing the ratings. Fromkin and Streufert (1976) point to the importance of identifying boundary variables that can impede the external validity or generalizability of laboratory research. Ironically, in this area of research this concern is often overlooked or given perfunctory acknowledgement in discussing the implications of the study. Only two studies conducted in a laboratory setting specifically investigated or manipulate the motivation of the raters (i.e., McIntyre et al., 1984; Zedeck & Cascio, 1982). As noted above their results conflict with one another, whereas Zedeck & Cascio (1982) found that the perceived purpose of the ratings affected accuracy and rater discrimination, the other research did not establish a very strong effect. Additionally,

Warmke and Billings (1979) found that differences in training effected experimental ratings but these differences disappeared on subsequent administrative ratings. They suggested that training effects may not generalize to adminstrative ratings unless important contextual variables (e.g., performance pay contingencies, political pressures) are taken into consideration. Bernardin and Villanova (1986) stongly urge the continued investigation of motivation and purpose as potentially significant individual level variables that affect the generalizability of this body of research.

Training Paradigms

As seen through Table 1, the training paradigms were varied throughout the research. Nine employed some type of rater error training, four employed some form of rater accuracy or frame-of-reference training, two employed an intensive training program, three used an observational training strategy, one study developed computer-assisted training, one employed behavior modeling, two used training as it related to scale construction, and one used training in decision-making. Table 2 presents descriptions of the training programs in each of the 19 studies.

Table 2

Description of Performance Appraisal Training

Author	Condition	Description
Brown (1968)	Trained and Untrained	Trained group told about different types of scales, procedures, and problems in obtaining sound ratings and constant errors; given practice on rating scales; discussed own ratings. Duration was 1 hour.
Borman (1975)	Halo Training and No Training	Read description of halo; demonstrated what halo looks like; told not to give overall ratings but should pick out strengths and weaknesses. Duration was 5-6 minutes.
Latham, Wexley, & Pursell (1975)	Workshop, Discussion, and Control	Workshop group observed video of manager making observational errors; discovered own errors; received feedback; practiced. Duration was 14 hours. Discussion group trainer defined and gave examples of errors; trainees generated personal examples. Duration was 9 hours. All groups given detailed job descriptions and job requirements.
Bernardin & Walter (1977)	Four training groups: RET with Diary Keeping and Scale Orientation; RET with Diary Keeping, and Brief Written Instruction	Group 1 received RET training; distributed BES scale; kept behavioral diary for 10 weeks prior to actual ratings. Group 2 received same training as Group 1 but no scale orientation; given rating scale dimensions at training. Group 3 received same training as Groups 1 and 2 immediately prior to actual ratings.

Table 2 continued

Author	Condition	Description
		Group 4 received brief written instructions immediately prior to actual ratings.
Bernardin (1978)	RET with Diary Keeping and Halo Training	Group 1 received same training as Group 2 in Bernardin & Walter (1977); also given practice data and discussed evaluations.
		Group 2 received same training as Borman (1975); also reference made to the scale dimension ratings.
Vance, Kuhnert, & Farr (1978)	Written RET and No Training	Both groups given job description. Trained group received instructions on errors and numerical descriptions.
Borman (1979)	Workshop Training with Expert Feedback and No Training	Training same as in Latham et al. (1975); also had feedback of true scores and behavioral rationales; ratings collected 1-2 weeks after training. Duration was 3 hours.
Warmke & Billings (1979)	Lecture, Discussion, Scale Construction, and Control	Lecture group received lecture on ratings errors; included definitions, graphic examples, and examples on how to avoid errors. Duration was 2 hours.
		Discussion group received modified version of Latham et. al.'s (1975) discussion session. Duration was 3 hours.
		Scale construction group developed new rating scale; no discussion of error; most time spent on behaviorally defining dimensions. Duration was 4 hours.

Table 2 continued

Author	Condition	Description
Ivancevich (1979)	Intense, Discussion, and Control	Intense group viewed 30 min. video of employee showing behaviors on scale; rated video; discussion on ratings and halo and leniency; discussed BES at length. Duration was 14 hours. Discussion group received same training; did not include video participation. Duration was 14 hours. All groups received the BES and a users manual prior to training.
Thorton & Zorich (1980)	Behavioral Instruction, Error Instruction, and Control	Control group told to take notes and informed they would be questioned after viewing 45 min. video. Duration was 3 minutes. Behavioral instruction group had same instructions as above; also told to observe behavioral details, note specific verbal and nonverbal behavior. Duration was 5 minutes. Error instruction group had same instructions as both above; also lecture, description and avoidance tips on eight errors. Duration was 15 minutes.
Pursell, Dossett, & Latham		Training session similar to Latham et al. (1975); also given additional errors; rated behaviors prior to training.
Bernardin & Pence (1980)	RET, RAT, and Control	RET group followed Bernardin (1978) and Borman (1975). RAT group discussed multidimensionality of jobs, and need to distinguish performance; stressed fair and accurate ratings; generated and defined dimensions; discussed examples of high, medium, and low behavior. No scale orientation.

Table 2 continued

Author	Condition	Description
Zedeck & Cascio (1982)	RET with Workshop Training	Trained group received explanation and example of rating errors; also 2-3 hours of outside reading; practiced appraisal through role playing, discussion, and feedback. Duration was 5 hours.
Pulakos (1984)	RET, RAT, RET/RAT, and Control	RET group followed Latham et al.'s (1975) workshop; no discussion on dimensions; participated with scales and videos; critiqued own ratings. Duration was 1.5 hours. RAT group was lectured on multi-dimensionality of jobs, given rating scale, discussed behaviors at different performance levels; participated with scales and videos; behavioral rationales discussed; received feedback on accuracy. Duration was 1.5 hours. RET/RAT group received combined and shortened version of each. Duration was 1.5 hours
Davis & Mount (1984)	Computer Assisted Training (CAI), CAI and Behavior Modeling, and Control	CAI consisted of a 6 chapter text and four 30 min. learining activities. Duration was 6 hours. Behavioral modeling training was concerned with conducting the performance appraisal discussion and receiving feedback about strengths and weakness. Included lecture, video's, role playing, feedback and discussion. Duration was 1.5 days.

Table 2 continued

Author	Condition	Description
McIntyre, Smith, & Hassett (1984)	RET, FOR, RET/FOR, and Control	RET group received rating scale; dimensions read aloud and discussed; given typical errors and graphic examples; group discussion on how to avoid these errors. Duration was 15 minutes. FOR group had same beginning as above; viewed videos; rated behaviors; given true score feedback and behavioral rationales. Duration was 30 minutes. COMB group had RET and then FOR. Duration was 45 minutes. All groups received rating scale and learned dimensions.
Pulakos (1986)	Evaluative, Observational, and Control	Evaluative group same as Pulakos (1984). Duration was 1.5 hours. The observational group was lectured on attending to behaviors; focus on counting relevant behaviors; given list of behaviors corresponding to each scale dimension; memorized, rehearsed and wrote behaviors for each dimension; given list of questions to review when viewing tapes; received true score feedback and discussed. Duration was 1.5 hours.
Athey & McIntyre (1987)	FOR, Information Only (INFO), and Control	FOR group was same as McIntyre et al. (1984); also same information as INFO group. Duration was 30 minutes. INFO group received visual and oral presentation of performance items and behavioral components of scale dimensions; given scale and behaviors for each performance level. Duration was 20 minutes.

Table 2 continued

Author	Condition	Descriptions
		Control group received brief explanation of scale.
Hedge & Kavanagh (1988)	RET, Observation, and Decision-Making	RET group similar to Pulakos (1984); but no scale orientation or practice. Duration was 3.5 hours.
		Observation group was instructed on importance of being good observer; told to take notes; presented current performance appraisal and defined dimensions using observational keys; discussed systematic observational errors and how to avoid them; twice viewed videos, rated behavior, received true scores; taught strategies of observation; through case study taught correct and incorrect observational strategies. Duration was 3.5 hours.
		Decision-making group received lecture on intuitive and formal DM strategies and costs and benefits of both; lectured on common DM errors and demonstrated inappropriate judgements; had discussion session with video exercise and still life scenes from work; developed list of observations and inferences; discussed differences between two. Duration was 3.5 hours.

While at the initial reading it may appear as though there is a large degree of overlap among the training paradigms, closer inspection reveals divergent approaches or elements within each study. There are many inconsistencies, or weak translations and operationalizations of training components across programs. These inconsistencies may account for the lack of strong conclusions about the effects of performance appraisal training. While investigators often claim to be following the approach of other researchers there are design changes that alter either the change process (the intent of the training) or how it is done (the method of training). As such, the training components (i.e., feedback, practice) are not being systematically manipulated and the process effects of the training are not being systematically explored. It is often difficult to distinguish the unique contributions of either method or content of the training and thus the results do not provide as much information as they might had these distinctions been built into the investigation.

Throughout the performance appraisal literature the most often discussed and investigated training paradigms are of four types. The purpose of this section is to compare and contrast the similarities and distinctions within these training paradigms.

Rater error training. Rater Error Training (RET) is usually attributed to Borman (1975). It is the most common and basic of the training paradigms. Original conceptualizations (i.e., Bernardin & Walter, 1977; Borman, 1975; Warmke & Billings, 1979)

involve lectures on psychometric errors including some combination of halo, leniency, central tendency, contrast effects, or similar-to-me error. A graphic and numerical demonstration of these errors is presented and trainees are admonished to avoid these errors. At the conclusion, trainees conduct performance appraisals on the stimulus material, most often videotaped behavioral vignettes.

Some reformulations of this program incorporate training on scale dimensions (i.e., Bernardin, 1978; Bernardin and Walter, 1977; Ivancevich, 1979; McIntyre et al., 1984), and practice and feedback (i.e., Bernardin, 1978; Bernardin & Pence, 1980; Hedge & Kavanagh, 1988; Pulakos, 1984; Zedeck & Cascio, 1982). Additionally, some RET programs include pre-training ratings (i.e., Bernardin, 1978; Borman, 1975; Hedge & Kavanagh, 1988; Ivancevich, 1979) and others do not (i.e., Bernardin & Walter, 1977; McIntyre et al., 1984; Pulakos, 1984; Vance et al., 1978; Zedeck & Cascio, 1982). Given these important methodological differences comparisons across trainees cannot be made. When attempting to generalize results it becomes unclear whether effects were due to the intent of RET or rather they were due to familiarizing trainees with the scales, employing practice and feedback, sensitization to scales while gathering pretraining ratings, and so forth. The trainees' experiences are no longer similar since they are now in some manner systematically

different from one another across studies. There is now the potential of criterion contamination.

Workshop training. Another popular training strategy is Latham et al.'s (1975) workshop. As part of the training exercise trainees are given detailed job descriptions and job requirements of managers/actors they observe on videotape during an interaction with a job candidate/actor. The vignettes depict managers in the process of making observational errors. For example, the trainees witness a manager interview a candidate in which the manager is so obviously impressed with the candidate's non-job related background that it completely colors his view of the candidates job related credentials (halo). Trainees then develop their own ratings and suggest ratings they think the managers/actors would give the job candidate/actor. Within the context of the observational errors made, trainees receive feedback on their own observations, practice rating behaviors, and discuss the errors.

Many studies claim to be using Latham et al.'s (1975) workshop program. Yet again, methodological differences occur. Their program lasted 14 hours, whereas others ranged from one and a half hours (Pulakos, 1984) to eight hours (Pursell et al., 1980) with various times in between. Additionally, the errors discussed in the various programs are not consistent across studies or reflective of the original performance appraisal errors. Hedge and Kavanagh (1988) specifically discuss one of

their training programs as paralleling Latham et al.'s (1975) with the addition of some training components. Careful observation reveals few similarities including lectures on observation, note-taking, scale orientation, feedback of true scores, and training in observational strategies. However, they point to the shorter duration of training as the major difference between the two programs. It can be seen that across research studies claiming to use the workshop technique the content and methodologies are very divergent.

Performance dimension training. Among the more recent and popular programs is what, for our purposes, will be considered performance dimension training. This incorporates many recent training programs and is best typified by a training design where trainees discuss the multidimensionality of performance and the need to distinguish performance dimensions, as well as levels of performance (Athey & McIntyre, 1987; Bernardin & Pence, 1980; McIntyre et al., 1984; Pulakos, 1984, 1986). Trainees then practice with the scales by rating performance vignettes, and receive feedback on their rating accuracy through presentation and discussion of expert ratings of the same behaviors.

Performance dimension training combines both RAT and FOR. The latter was first proposed, but never researched, by Bernardin and Buckley (1981) and the former derives from both Bernardin and Pence (1980) and Pulakos' (1984) work. Specific to RAT, in Bernardin and Pence's (1980) program, a group of trainees

generate and define performance dimensions, but in Pulakos' (1984) program the dimensions are defined for the trainees. In Bernardin and Pence's (1980) research trainees never see the rating scale, but in Pulakos' (1984) training program the scales are discussed and used for practice. Additionally, in her program trainees practice rating behavioral vignettes and receive feedback and discussion on the accuracy of their ratings, along with behavioral rationales of the true performance scores. So while RAT derives from two sources there are important processual and methodological differences between each.

FOR training was made operational through the work of McIntyre et al. (1984) and Athey and McIntyre (1987). However, even in this case the operationalization differs from the original conceptualization. The original training program proposed by Bernardin & Buckley (1981) involves: 1) giving trainees job descriptions of the job to be rated and discussing as a group the duties and qualifications necessary to perform the job; 2) performing three trial performance appraisals and writing out the justifications for their ratings; 3) having trainers feedback the correct rating based on expert scores and relaying the expert's rationales for each performance score; and, 4) generating a discussion that focuses on the discrepancy between the true ratings and the trainees ratings. McIntyre and his colleagues synthesized these components into a shortened version by only having one practice session, not requiring any trainee

justifications, and not employing any discussion of discrepancies between the expert ratings and trainee ratings. Additionally, in another study Athey and McIntyre (1987) also distribute scale items and behaviors for each level of performance. Therefore, large differences exist between the actual training and the proposed FOR program. Based on these large methodological and process discrepancies across performance dimension training programs it is difficult to determine whether it was the dimension training or the methodology that influenced the trainees.

Observational training. Finally, another current program is observational training. Whereas dimension training focuses on understanding performance dimensions, this method stresses careful observation (Bernardin & Walter, 1977; Hedge & Kavanagh, 1988; Pulakos, 1986; Thorton & Zorich, 1980). Generally, trainees are given the rating scales, and performance dimensions are discussed in terms of key observational points and relevant behaviors. Trainees may or may not be instructed to take notes when viewing performance. Scale practice is employed by rating videotaped vignettes and receiving feedback on the accuracy of their observations. Programs differ in whether or not trainees discuss this feedback as well as are instructed on explicit strategies of careful observation.

Many different training formats are utilized in these programs. Thorton and Zorich (1980) employ the simplest methodology involving an instructional set regarding

observational strategy and note-taking behavior. This program lasted five minutes. Bernardin and Walter (1977) incorporate an observational strategy in one training program by requiring participants to maintain observational diaries, and use this strategy along with practice and feedback in another study (Bernardin, 1978). The two studies most representative of the observational approach include Pulakos (1986) and Hedge and Kavanagh (1988). However, differences between the two are great. Pulakos (1986) uses dimension definitions and corresponding behaviors to train participants. Raters are required to memorize the dimensions and specific behaviors, are tested on their recall, and are given an observational strategy to follow. Subjects then practice with this method and receive feedback on the results as well as participate in a discussion. Conversely, Hedge and Kavanagh (1988) combined training in observational strategy with RET, practice, and feedback with true scores. As with the three paradigms presented above the attribution of effects are difficult to discern.

To summarize, within training paradigms there are widely divergent operationalizations that purport, both implicitly and explicitly, to be facilitating the same training approach. But as is evident throughout this discussion this is clearly not the case. The differences between the programs cover ideology and methodology and therefore should prevent comparisons across studies. The studies differ in relation to: 1) the amount of

information they provide about the rating task; 2) the amount of information trainees are asked to integrate and organize; 3) the types of cognitive activities that are called upon during training and different levels of cognitive processing required in the rating task; 4) the degree of familiarity and practice participants have with the stimulus materials; 5) and the method of presentation.

Training Results

Smith (1986) presents a comprehensive review of the performance appraisal training outcomes where he dichotomizes studies by the method and content of training presentation and discusses trends in the results. Methods include lecture presentation, group discussion, and practice and feedback. Training content is grouped as RET and includes the typical discussion of common rating errors; Performance Dimension Training where the purpose is to familiarize raters with the rating dimensions by reviewing the scale, having raters participate in scale construction, or providing descriptions of job requirements; and Performance Standards Training where the purpose is to instruct raters in a common frame of reference for judging performance by demonstrating job behaviors and feeding back the true performance scores. Differences exist between his review and the one herein in terms of which elements of training are being identified as method issues and which are being identified as content issues. It is the belief here that his

organization of training content is confounded with method. This becomes evident through his difficulties in discriminating conclusions regarding content results. However, even with these discrepancies, the review of his conclusions adds understanding to this literature.

Using his categorization schema, Smith (1986) reaches the following conclusions. In general, the lecture method was unsuccessful at increasing accuracy unless it was combined with one, or both, of the other two methods. It was successful at reducing halo, but no more so than the practice and feedback method, and least successful at reducing leniency. The discussion method was the most successful at reducing leniency and halo. The discussion method also increased accuracy, but in the three studies that measured this variable, discussion was employed along with practice and feedback. Practice and feedback was the most successful method for increasing accuracy and fairly successful at reducing leniency.

The conclusions for the content of training are more complicated. RET training seems most effective at reducing halo; the majority of studies that employed this technique demonstrated a reduction in halo. The results for leniency and accuracy are more confused. While one RET study (Bernardin, 1978) reduced leniency, only two studies replicated this result (Pursell et al., 1980; and Ivancevich, 1979). Additionally, those two studies did not use RET in isolation but appear to have included

significant segments of training that incorporated practice and feedback. The same contrast is found for accuracy measures. Two RET studies demonstrated a negative effect on accuracy while two studies resulted in an increase in accuracy indices. However, the studies that generated an increase in accuracy with RET also included practice and feedback (Pulakos, 1984), or performance dimension training (Fay & Latham, 1982).

Only one study (Bitner, 1948) used a pure form of Performance Dimension Training and this demonstrated a reduction in leniency. All the remaining studies combined the content of this training program with either Performance Standards Training or RET training. When combined with RET it resulted in a reduction in leniency. When combined with Performance Standard Training the two resulted in an increase in accuracy, and a decrease in leniency. Additionally, all three methods reduced leniency. Performance Standards Training, when used alone, reduced leniency and halo. Again, in combination with Performance Dimensions Training it increased accuracy.

Smith's (1986) article is meritorious in its attempt to synthesize a diverse body of literature with very divergent operationalizations of both the independent and dependent variables. Unfortunately, his efforts may have the wrong effect. What he is presenting is an acceptance of the differences in this research in an attempt to draw together some trends and general conclusions. However, by being less accepting and more critical of the research we can perhaps move forward and begin to address the

discrepancies and confounds. By doing such the training research can be further improved by clearing away confounds that may be impeding the identification of salient, and consistent, results.

Rating Effectiveness

Performance appraisals, as have been discussed here, are prone to errors of judgement inherent in any subjective rating system. The purpose of this section is to briefly review the history and controversies in this area and again point to inconsistencies in this area of the literature. The different psychometric operationalizations of accuracy and error indices used in the performance appraisal training literature are listed in Table 3.

Errors

Errors as discussed in the performance appraisal literature differ from common usage of the term and refer here to systematic bias, not random measurement error. The issue of errors in subjective rating data was investigated by Thorndike in 1920. He developed the error concept of "halo," or the high intercorrelations between behavioral dimensions due to a ratee evaluation as generally good or generally bad which permeates all performance dimensions. During the next 10 years various researchers brought to attention other judgmental rating errors such as a lack of interrater agreement, leniency/severity, central tendency, range restriction, and several other criteria (Saal, Downey, & Lahey, 1980) as measures of performance appraisal quality.

Table 3

Accuracy and Error Operationalizations

Index	Operationalization
Halo	
	Intercorrelations among dimension ratings over ratees for an individual rater.
	Results of a principal component factor analysis of the dimension intercorrelation matrix.
	Variance or standard deviation of an individual rater's ratings of an individual ratee across all dimensions.
	Rater x Ratee interaction in a Rater x Ratee x Dimension ANOVA.
Leniency	
	Comparison of mean dimension ratings with scale midpoints.
	Rater x Ratee x Dimension ANOVA with evidence of a rater main effect.
	The degree of skewness of the frequency distribution. A negative skew equals leniency and a positive skew equals severity.
Central Tendency/Restriction of Range	
	The standard deviation of the ratings assigned to all ratees within a particular performance dimension.
	The proximity of the mean dimension ratings to the scale midpoint.
	The degree of kurtosis of the frequency distribution of the dimension ratings for multiple dimensions.
	Rater x Ratee x Dimension ANOVA with evidence of a ratee main effect.

Table 3 continued

Index	Operationalization
Interrater Reliability	
	The standard deviation of the ratings assigned to a particular ratee by several raters for a single dimension.
	An intraclass correlation coefficient.
	Rater x Ratee x Dimension ANOVA without a Rater x Ratee interaction.
Elevation Accuracy	
	Accuracy of the average rating, over all ratees and dimension. Refers to the accuracy of the rater.
Differential Elevation	
	Accuracy of the mean rating for each ratee across all dimensions. Refers to distinctions among ratees in overall performance.
Stereotype Accuracy	
	Accuracy of the average ratings given to each job dimension across all ratees. Refers to the accuracy of rater's rating for a particular dimension.
Differential Accuracy	
	Accuracy with which ratees are rank ordered on a given dimension. Refers to differences among ratees in patterns of performance.
Distance Accuracy	
	Absolute average deviation of rater's ratings from true scores. Reflects the level difference between trainees and experts.

Table 3 continued

Index	Operationalization
Differential (Correlational) Accuracy	The accuracy of rater's discrimination among ratees on a number of performance dimensions.
Correlational Halo Accuracy	The observed dimension intercorrelations for each rater's rating of each ratee minus the true dimension intercorrelations.
Absolute Value Halo	The mean difference between the variance of obtained ratings per ratee and the variance of expert ratings for that ratee computed across ratees. A positive value reflects more halo than should exist.
Leniency Accuracy	The true score for each dimension subtracted from the mean rating for each dimension across ratees for each rater. A negative value reflects leniency and a positive value reflects severity.

The literature is rich with investigations of these criteria. Therein is where problems lie. Once again, there is a lack of agreement over conceptual and operational definitions. For example, Saal et al. (1980) provide an excellent review of the conceptual and operational definitions of common rating errors in the literature up until 1977. They point to four different operationalizations of halo, three conceptual and three operational definitions for leniency/severity, three different operational definitions for restriction of range which they distinguish as separate from central tendency, and five different operational defintions of interrater reliabilty. Yet in the literature these terms are often used and operationalized interchangeably (Saal et al., 1980).

This lack of agreement then leads to false conclusions when trying to identify trends in the literature or generalize across studies. Recently, research (e.g., Becker & Cardy, 1986; Murphy & Balzer, 1981; and Sulsky & Balzer, 1988) has demonstrated that these different operationalizations lead to different conclusions about the existence of rating errors in particular data sets. Murphy and Balzer (1981) found that a variance measure of halo, reflecting intraratee rating differences across dimensions, and a correlational measure of halo, looking at the intercorrelations of each rater's appraisal across dimensions and ratees, were independent of one another. Becker and Cardy (1986) found differential relationships between two type of halo and accuracy

measures. Central tendency and restriction of range are similarily confounded. Whereas central tendency refers to ratings that cluster at the mean of the rating scale, restriction of range refers to ratings that are not spread out along the rating scale and which could cluster around any point on the scale. Central tendency can encompass restriction of range, but restriction of range does not necessarily conote central tendency. Additionally, leniency/severity is used to refer to a rater's tendency to assign ratings above or below some true ability level across all ratees, as well as a shift in mean ratings consistently biased in the positive or negative direction. Thus when trying to interpret effects and develop a critical understanding of the pattern of training results, the divergent conceptual and operational indices may lead to misinterpreted or overlooked patterns.

Accuracy

Measures of halo, leniency/severity, central tendency, and interrater reliability provide only indirect evidence regarding rating quality. The underlying goal in this line of research is to determine the quality or "goodness" of subjective performance ratings. "Goodness," or accuracy as it has come to be called is the degree of "closeness" between the rating and some "true score," however operationalized, and is the actual focus of attention or study. Cronbach (1955) proposed that subjective accuracy, as generally discussed, consists of the sum of four

different components assessed across ratings on multiple dimensions. Each component has a different conceptual meaning and operational definition. Elevation refers to raters overall level of rating or the distance between their average ratings across all ratees and dimensions and the average true score rating across all ratees and dimensions. Differential elevation refers to the accuracy with which raters identify overall differences among ratees resulting in a correct rank ordering of ratees. Stereotype accuracy involves differences between performance dimensions across individual ratees, or the accuracy with which raters evaluate performance within a dimension across ratees. The last component, differential accuracy indicates discrimination between ratees for each performance dimension, or raters accuracy in identifying ratee performance profiles. These indices can only be assessed against some known true measure of a ratee's performance. This assessment has now become possible through a methodology originally developed by Borman (1977).

Another accuracy index is Borman's (1977) measure of differential accuracy. This is based on a correlation of rater scores with expert's scores for each dimension. Some researchers (e.g., Fisicaro, 1988; and McIntyre et al., 1984) prefer to term this "correlational accuracy" to clearly identify it as different from Cronbach's (1955) measure of differential accuracy. They felt that Cronbach's (1955) measure, in ANOVA terms, is based on a ratee by dimensions interaction and thus involves mean

differences. Borman's measure is based on correlational information and in this measure the mean difference between rater and expert is lost (Becker & Cardy, 1986). A sixth accuracy measure is termed distance accuracy (Fisicaro, 1988; McIntyre et al., 1984; Sulsky & Balzer, 1988; Vance et al., 1978) and refers to the average absolute deviation of subject ratings from true scores (Sulsky & Balzer, 1988).

Recently, in a review of the research on accuracy measures Sulsky & Balzer (1988) develop the point that these measures are not based on a common conceptual accuracy definition. They examine, empirically, the different operationalizations using two data sets. Consistent with other research (e.g., Becker & Cardy, 1986; Fisicaro, 1988; and Murphy & Balzer, 1981) they found weak relationships between different accuracy measures and suggest that different accuracy measures may yield different results such that each index is measuring a different type of rating ability. Thus, it is important to understand the potential differential pattern of results between accuracy indices. Different training programs may increase some types of accuracy but fail to increase or negatively affect other types.

True Scores

Each accuracy component involves comparison of a rater's score with some known true score. The true score can be an objective criterion measure if one is available, but as is the case most often, the true score consists of corresponding

ratings generated by a group of "experts." As is typical throughout this literature, there are multiple methods for developing true scores. Borman's (1977) method was the first developed and is the most common approach. It involves a group of experts who are given enhanced, or repeated opportunities, to view a ratee's behavior and are throughly familiarized with the rating scale and potential rating errors. These experts then rate the ratee's performance and the average of these scores, across raters, is taken to represent a "true" performance score. Experts are generally upper level graduate students in industrial/ organizational psychology who have taken courses in performance measurement and are sensitive to potential rating errors and the issues involved in generating accurate ratings.

A second method is to generate a mean score of all raters' scores (Bernardin & Pence, 1980) as a measure of the true performance rating. A third method uses scores produced through generating a ranked set of behavioral critical incidents by one set of experts and then having a second set of experts rate the ratees performance against these critical incidents (Becker & Cardy, 1986; Borman, 1975; Zedeck & Cascio, 1982). This procedure is very similar to the development of behaviorally anchored rating scales (Smith & Kendall, 1963). Finally, McIntyre et al. (1984) and Athey & McIntyre (1987) have used Borman's (1977) method but adapted it by generating dimension

consensus scores among small groups of experts as true scores, rather than taking the average of the independent expert's scores.

The precision of the true score methodology has been questioned by Sulsky and Balzer (1988). In the consensus method there is a question regarding the potential effect of initial agreements or disagreements and the magnitude and direction of these effects on the resulting true scores. In the Borman (1977) method, true scores are derived from an average of expert scores. As noted by Borman (1977) and Sulsky and Balzer (1988), experts do not necessarily agree on their ratings, however, this information is lost in the final true score, as it also is with the consensus method. Furthermore, even if raters do agree this could be the result of some systematic bias, as opposed to a true performance rating (Sulsky & Balzer, 1988). When an average over all raters is used it is often derived from a sample of unskilled raters (undergraduates) and therefore, it is difficult to consider this an "expert" generated or "true" score.

In the only study to date which explicitly examines the current use of expert generated true scores Smithers, Barry, and Reilly (1989) provide evidence supporting this methodology. They compared the accuracy of expert (graduate industrial/ organizational students) true scores with nonexpert (undergraduate students) scores and found that the experts were consistently more accurate when evaluated against an objective criterion measure. They also found evidence which demononstrated

a high correlation between objective true scores and true scores produced through mean expert ratings. However, it should be noted that the rating task for determining true scores and protocol was somewhat different than the usual task and protocol. True score ratings were developed based on objective performance levels, however the performance ratings were developed such that ratees were compared against one another.

Less Error, Greater Accuracy?

A recurring question in the literature involves the degree of association between indices of error and accuracy (e.g., Cooper, 1981; Fisicaro, 1988; Murphy & Balzer, 1986; Pulakos, 1984). Recall that indices of error were used to determine the quality or "accuracy" of subjective performance ratings. The assumption was that the less error that existed in ratings the greater the accuracy of the ratings. Therefore, for example when conducting research on performance appraisal formats or training programs if it could be demonstrated that errors were reduced the logic was that the ratings were consequently more accurate.

However, while investigating the relationship between halo and accuracy Cooper (1981) found results that were completely contrary to the currently held assumptions. In a reanalysis of data from five studies he demonstrated a weak but positive relationship between halo and accuracy. Given low correlation coefficients (median value of .15) between these measures he suggested that the relationship between halo and accuracy was

weak enough to be considered somewhat inconsequential. This study, along with results from studies by Murphy and Balzer (reported in Becker & Cardy, 1986) and Pulakos (1984) led to a position in the literature which tended to support a clear division between measures of accuracy and error and the belief that the absence, or presence, or rating errors had no direct consequences for rating accuracy.

Fisicaro (1988), however, points out that the analyses conducted by Cooper (1981) were confounded. Some of the correlations generated by Cooper (1981) were between halo and accuracy and others were between halo and inaccuracy. Fisicaro (1988) when correcting for this inconsistency demonstrates support for the original assumption of a negative relationship. He also generates correlation coefficients of a greater magnitude than Cooper's (1981), which suggests a stronger relationship between the two measures than originally believed. These results were supported for all accuracy measures discussed above, with the exception of differential elevation and two different absolute halo measures (ratees standard deviations and dimension intercorrelations).

Similarly, Becker and Cardy (1986) also found that the relationship between halo and accuracy depends on the operationalizations used to generate the indices. Sulsky and Balzer (1988) suggest that the differences found between various

experimental manipulations may be dependent upon not only the manipulation but the error and accuracy operationalizations as well.

It has been shown in several studies mentioned herein that the measures of rating quality may have no relationship to one another. This includes comparision between error indices, accuracy indices, or across error and accuracy indices. Given that each index may be measuring different abilities, or generating different information in terms of the effectiveness of the manipulation it may seem fruitful at this stage to assess as many different operationalizations as possible within one data set. The differential results may reveal more meaningful information than just one measure alone. Additionally, the training may actually be promoting one type of accuracy versus another since each accuracy component is potentially measuring a different ability (Murphy, et al., 1982).

The Present Study

Based on the above review it seems appropriate that what is necessary in this area of research is the consolidation and systematic evaluation of training paradigms already in existence. Performance appraisal training research is in desperate need of synthesis to more directly explore the effects of individual components of the training (methods and intentions), in an attempt to isolate these effects. We can then begin to understand the differential contribution of the component parts

of the process underlying training. This will contribute to a greater understanding of the effective elements of performance appraisal training and the determination of components to include in future training programs. Additionally, from an applied perspective this can greatly help in streamlining a time consuming and expensive area of training, as well as contributing to more accurate and error free performance ratings.

To meet these research aims, FOR training has been chosen as the vehicle with which to systematically investigate common methods of training presentations and the content of performance appraisal training. Given the possibility of employing any of the many training paradigms in this research domain FOR training was chosen for several reasons. First, based on an extensive literature review FOR training appears to demonstrate the most promise of leading trainees to more accurate ratings (Athey & McIntyre, 1987; McIntyre et al., 1984; Murphy, 1990; Smith, 1986). Second, as originally conceptualized it has clearly defined component parts (Bernardin & Buckley, 1981), which have never been investigated individually. Third, each component facilitates different cognitive tasks and can be linked to cognitive changes in relation to levels of processing (Athey & McIntyre, 1987). Fourth, as a training platform it incorporates many of the principles and methodologies of other programs and therefore may provide information beyond the limits of one training method or content area. These component parts can be

linked to the different methods discussed herein. Therefore, by manipulating each component evidence is also gathered that may generalize to other training programs. Finally, given the instrumental value of each component of the training program, identification of the strength or weakness of a component of FOR training would have a direct value from an applied perspective.

Using the original FOR training program proposed by Bernardin and Buckley (1981) there are five steps (described above) representing the separate training components. In this research one additional component has been added which has also been used previously by McIntyre et al. (1984) and others. The added component involves distributing the rating scale and discussing the dimensions before completing the practice exercises. This component seems to be included in studies that demonstrate greater accuracy in ratings and is suggested by Smith (1986). Also, it is a signficant feature in many of the other training programs and therefore its inclusion may be illuminating. This study also provides for generalizability by utilizing the identical FOR training materials employed in the work of McIntyre and his colleagues including the videotaped behavioral vignettes and rating scale.

This project capitalizes on studies demonstrating the necessity of considering trainee motivations (e.g. Zedeck & Casio, 1982). In order to increase student trainees involvement the trainees will be told that the individuals they observe will

actually receive their feedback and evaluations. Also the student trainees will be told that the individuals are being considered for employment and that their review will be considered in the decision.

The measurement situation itself will also be investigated to determine if there are factors in the design that differentially affect the performance ratings. In an effort to control for elements of the design the training stimuli will be counterbalanced. In this manner the training stimuli, individually, can be studied to determine its role in training effectiveness. This is important at two levels. First, as mentioned previously the stimuli (the videotaped behavioral vignettes) are often used across studies but apart from determining their validity in demonstrating differences in behavioral dimensions, their impact on raters has not been established. Secondly, if they are having an effect on results, collapsing across vignettes may be mitigating the subsequent effects of the training components. The measurement situation will be investigated in relation to vignette lecture topic effects and effects of the order of topic presentation. The knowledge of these design elements' contributions to the research results may lead to a clearer understanding of the elements of the training itself.

Another aspect of confusion in this literature concerns the accuracy and error measures employed. It was pointed out above that these measures differ in conceptualization and operationalization. Additionally, it was noted that different operationalizations can result in different conclusions about the results of studies, and that similar measures of the same concept may be completely independent of one another. Therefore, it seems important to employ multiple operationalizations of the error and accuracy measures to this research, as opposed to relying on one particular operationalization which may or may not evidence any results. This provides a unique opportunity to compare the different error and accuracy measures against one another within the same data set and across separate conditions. It will be illuminating to compare different measures of the same concepts against the various component parts of FOR training.

Based on the preceding synthesis of the performance appraisal literature the following hypotheses are offered for investigation in the present study.

HYPOTHESIS I

FOR training that incorporates all of the components of training will be more effective at increasing accuracy and decreasing error than training that incorporates only some of the components of training.

HYPOTHESIS II

The training conditions that require trainees to be more active will be more effective in increasing accuracy and decreasing error than those conditions that do not require as much active participation.

HYPOTHESIS III

Of the conditions that require less active participation, those that include feedback of true scores and behavioral rationales, and discussion will be more effective in increasing accuracy and decreasing error than those that do not include this component.

HYPOTHESIS IV

Trainees with the greatest accuracy scores on the memory observation scale for what was observed will also demonstrate greater accuracy on the behaviorally based performance appraisal scale than those trainees with poorer accuracy on the memory observation scale.

No hypotheses will be offered for the investigation of the differences in error or accuracy indices, or for the investigation of the factors in the measurement situation. Both these aspects are exploratory in nature and given the uniqueness of their investigation in the literature it seems more appropriate to avoid constraining their exploration to explicit hypotheses.

CHAPTER III

METHODOLOGY

Design Overview

A 2 x 6 x 2 incomplete factorial design with fractional replication was employed in this study. The factors and levels were: Lecture topic (crowding and stress (CS) vs. the self fulfilling prophecy (SFP) x Conditions of FOR training (see below) x Pairing of levels of performance viewed during training (high vs. low) and topic. Intact classes of college students were randomly assigned to one of six training conditions. Two classes were randomly assigned to each condition. The study took place during normal class time. Regardless of the training condition, all trainees practiced rating two lecture vignettes (CS and SFP) with a behavioral rating scale. At posttest, trainees viewed two more lectures on the same topics. However the lecturers were different than those seen during training, yet consistent across all conditions. Trainees completed the same performance ratings scale as during training as well as an additional Behavioral Frequency Scale for each vignette. After all stimulus materials were completed they then responded to a demographic questionnaire. The rating true scores and corresponding behavioral rationales were gathered from expert raters prior to running subjects. The true scores and behavioral rationales were used as feedback during training and employed in the data analyses.

Rating Scale

The Instructor Evaluation Scale is a behaviorally based measure developed by Costin (1974) to specifically assess the performance of college lecturers. Critical incidents were derived from a review of research on lecturer behavior and through essays written by graduate and undergraduate students. This procedure resulted in 40 items. A second sample of 300 student rated these items against the "best lecturer they had in a college course." Factor analysis of the data resulted in four factors (i.e., organization, clarity of communication, elocutionary skill, and intellectual stimulation) with a total of 23 items that had a .400 or better factor loading.

This scale has been modified by McIntyre et al. (1984) and Athey & McIntyre (1987) for use in performance appraisal training research. They modified the response format from a frequency rating to an evaluative one, where raters respond to positively worded statements on a seven-point agree-disagree scale. Additionally, 11 of the items were dropped and three items were modified for use with videotaped performances resulting in a 12 item scale. An item measuring overall performance was also added. In a pilot of the modified scale, McIntyre et al. (1984) demonstrated an internal consistency reliabilty coefficient of .87 and an average interitem correlation of .37.

A combination of the modified and original scales were used in this project (See Appendix A). The evaluative response format supplied by McIntyre and his colleagues was adopted here, as were the 13 items. Two additional items from Costin's (1974) original scale were also included that seemed appropriate for evaluation in the present context. These were: 1) He encouraged questions during the lecture; and, 2) He made you interested in the material. The resulting 15 item scale was used to gather true score ratings and was considered the performance appraisal instrument. It was used during training and to collect posttest data.

Lecture Vignettes

The videotape performance vignettes were those developed by Murphy, Garcia, Kerkar, Martin, and Balzer (1982). Each vignette involved a person (male) delivering a lecture on either stress and the effects of crowding (CS) or on the self-fufilling prophecy (SFP). The vignettes were approximately 8-10 minutes long. The vignettes had scripted lectures for each content area. However, the organization and throughness of each lecture, the question and answer period, and the dynamics of the presentation had been systematically varied across topic areas. Drama students performed the role of lecturer and had been given explicit scripts and stage directions for the delivery of either an effective or ineffective presentation. Each of four drama students delivered two lectures on the same topic (either CS or the SFP) however their performances were varied in each vignette.

The effect of this manipulation resulted in three different levels of overall performance (low, medium, and high) within each topic, and two different levels of ratings for each actor within a topic. Therefore, as originally designed by Murphy et al. (1982), and as occured in this study each actor received one group of high ratings and one group of lower ratings. This manipulation allowed for the investigation of the effect of pairing lecture topic and level of performance while keeping the lecturer constant.

The expert ratings conducted on the eight lecturer vignettes prior to training resulted in a rank order of levels of performance of lecture vignettes across topics. Table 4 shows the experts' true score ratings broken down by actor, topic and rated level of performance. Only six lecture vignettes were needed here (two high, two medium, and two low levels of performance). However, within the eight performances it was necessary to match similar levels of high and low performance across actors and topics at pretest, and to match medium levels of performance across the remaining two actors at posttest in such a manner that trainees did not view the same actor twice and that performance levels were similar across conditions. Therefore, all trainees viewed similar levels of high and low performance across topics, the lecturer remained constant within topics at training (e.g., the same actor delivered a high stress lecture performance and a low stress lecture performance), and

Table 4

Expert True Score Ratings by Performance Level, Actor, and Topic

Topic	Performance Ranking		
	High	Medium	Low
CS			
Actor 1	39	61[a]	*
Actor 2	52[a]	*	88[a]
SFP			
Actor 3	*	69[a]	78
Actor 4	43[a]	*	89

Note. Ratings are the sum across all dimensions.

* No tape for this combination.

a Tape selected for FOR training.

the trainees viewed four different lecturers across the four rating occassions (two at training and two at posttest). Table 4 notes the final selection of the six lecture vignettes that best serve the goals of this study.

Expert True Scores

Lecturer performance true scores were gathered from expert raters who were given an enhanced opportunity to view the behavior on the videotape vignettes. Experts are generally considered those individuals familiar with the job (Athey & McIntyre, 1987; Borman, 1977; McIntyre et al., 1984), in this case lecturer, and with the performance appraisal literature (Borman, 1977; Murphy, Garcia, Kerkar, Martin, & Balzer, 1982; Pulakos, 1984). The experts in this study were 9 upper level industrial/ organizational psychology graduate students. They had completed all classwork leading to their degrees, including at least one class in psychometrics and two classes in personnel psychology. In order to ease scheduling constraints the experts were divided into two groups of five and four.

The procedure used to generate the expert scores was a combination of the techniques of Borman (1977), Murphy et al. (1982) and McIntyre et al. (1984). Thus, experts viewed the videotapes in a group, individually rated the performances and then discussed their ratings, and in doing so arrived at a consensus rating for each behavioral dimension. Each group met

for two sessions and rated four vignettes at each session. The order of vignette presentation was randomized for both groups.

Specifically, the expert sessions conformed to the following sequence. Before the experts rated the videotapes, the rating scales and behavioral dimensions were introduced and discussed. Experts were introduced to the content, context, and background of the videos and given time to review the rating scales. The videotapes were shown once and the experts were told to take notes on the lecturer's behavior as it related to the scale dimensions. When they felt comfortable with the material and behaviors presented they individually rated the vignette performance using the rating scales provided. They rated each vignette with both Costin's (1974) Instructor Evaluation Scale (discussed above) and with Murphy et al.'s (1982) Behavior Frequency Scale (discussed below). The Instructor Evaluation Scale also required a written behavioral rationale justifying each behavioral dimension rating.

Next, the experts discussed their ratings on the Instructor Evaluation Scale (Costin, 1974) within the group, and each expert was asked to give a specific behavioral rationale for why they chose a particular rating point. For each rating and rationale the experts' answers were elicited randomly within the group to avoid any type of group conformity behavior. The ratings and rationales for each dimension were discussed until consensus was reached. In this study consensus was defined as the point in the discussion when no more than a one point difference in ratings

across experts had been reached. For example, if some experts chose a performance rating of 4 and others chose a performance rating of 5 this would be considered consensus agreement. If, however, some experts chose a performance rating of 4, others chose a performance rating of 5, and still others chose a rating of 3, this was not considered consensus agreement. In cases such as these, the experts had to convince one another of the appropriateness of their rating until consensus was obtained. In most cases the experts would elaborate on their behavioral rationales and return to the videotape performances to review particular points of discrepancy. This procedure was carried out eight times, once for each vignette.

Each group of experts received the same instructions and materials and care was taken to facilitate all four group sessions in exactly the same manner. However, since ratings were gathered from two different groups of experts it was necessary to assess the degree of intergroup agreement. Following the procedure used by McIntyre et al. (1984) the consensual scores for each group were correlated across the 15 dimensions for each behavioral vignette. This resulted in eight Pearson Product-Moment correlations (one for each vignette) with a range of .44 to .94. The median correlation of .54 is taken as an index of interrater agreement. This correlation compares with McIntyre et al.'s (1984) results of .64.

There are several reasons that may contribute to this somewhat low median correlation. The range of ratings tended to be restricted with the majority of ratings ranging from 3 (slightly agree) to 7 (strongly agree), which may have attenuated the correlation. The number of items in the correlation was small given that the ratings were comprised of 15 items. Also, the process of arriving at the consensus ratings, within two groups, may have unknowingly generated some group conformity. However, the directon of this conformity may have been different in the two groups. For example, one group tended toward the more favorable end of the continuum, agreeing to a score of 3 (slightly agree), whereas the other group tended towards the more negative end, agreeing to a score of 5 (slightly disagree). Both scores appears to have resulted from the same behavioral rationales but became a choice of deciding between a glass half full or half empty.

While the experts rated eight performances only six of these were used in the remainder of this study (See Table 4). By dropping the two lecturer vignettes that were not used the intergroup agreement index was increased to .58. This increase comes closer to McIntyre et al.'s (1984) index of .64. However, given that this index is lower than what was expected it was important to assess other aspects of the validity of their ratings.

Another method of assessing the degree of intergroup agreement was to correlate each dimension rating across the eight behavioral vignettes for the two expert groups. As such, 15 Pearson Product-Moment correlations were developed that investigated the degree of relationship between the expert groups for the rank-ordering of ratees within a single dimension. This index was stronger than the previous index. It demonstrated a range of .61 to .95, with a median correlation of .83.

As an indication of the "expertise" of the experts the true scores were validated using the method proposed by Kavanagh, MacKinney, & Wolins (1971) and as modified by Borman (1978). This is the method most often used in the literature (Borman, 1978; Murphy et al., 1984, Murphy & Balzer, 1986; Murphy, Martin, & Garcia, 1982; Pulakos, 1984; 1986). Specifically, Borman's (1978) method differs from Kavanagh et al.'s (1971) by assuming ratees and dimensions to be fixed effects, and raters to be a random effect. These assumptions seemed appropriate for this study since the ratees and dimensions were prearranged and the raters were randomly selected. Also, accepting these assumptions allowed for the direct comparison of results with Borman (1978) and Murphy et al. (1982). Analysis of variance (ANOVA) was used in a Rater x Ratee x Dimension design. The analysis was conducted by entering each rater's dimension ratings for each of the eight lecture vignettes as an individual data point. In this manner the analysis was performed on what can be called the

experts' observed scores as opposed to the consensual ratings. This procedure is intended to measure the degree of convergent and discriminant validity to assess whether raters agree in their ratings and do, in fact, discriminate between ratees across dimensions. The intraclass index for the ratee main effects provides a measure of convergent validity and the intraclass index for the Ratee x Dimension interaction is interpreted as a measure of discriminant validity.

The intraclass index for the ratee main effect was .58. As an index of convergent validity it compares favorably with indices from other research. The intraclass index for the Ratee x Dimension interaction was .40. This index of discriminant validity also compares well with indices from other research. Table 5 compares the intraclass indices herein with results from Borman (1978) and Murphy et al. (1984).

Further evidence supporting the use of these scores was gained by examining the variance components of the Rater x Ratee x Dimension ANOVA. Kavanagh et al. (1971) explain that these measures are useful for making inferences of meaningful effects. Table 6 presents F values, the variance components, and variance indices for each source. The variance component of 1.57 for the ratee effect shows considerable convergent validity, or agreement in ratings between the experts. The variance component of .74 for the Ratee x Dimension interaction, while half of the convergent validity component, is still a strong indication of a good deal of discriminant validity among the experts' ratings.

Table 5

Comparison of Intraclass Indices Across Studies

Effect	Present	Borman (1978)	Murphy et al. (1984)
Ratee (Convergent Validity)	.58	.69	.57; .70
Ratee x Dimension (Discriminant Validity)	.40	.58	.21; .47

Note. In the Murphy et al. (1984) study, two performance appraisal ratings scales were employed.

Table 6

F Values, Variance Components, and Intraclass Indices of Experts' Ratings

Source	df	F	Variance Component	Index
Ratee	2,56	27.23**	1.57	.58
Ratee x Dimension	56,98	6.87**	.74	.40
Ratee x Rater	56,78	7.07**	.46	.29
Error			1.13	

** $p<.01$.

This measure is larger that the Ratee x Rater interaction and indicates that there is more rater discrimination between ratees than there is halo within ratings of ratees.

Given the accumulation of the different pieces of information described above it was concluded that true scores could be derived from the consensus ratings of the experts. True scores were operationalized as the mean consensus ratings of the two groups for each ratee on each dimension.

Independent Variables

The independent variables were the different components of FOR training, lecture topic, and the pairing of lecture topic and level of vignette performance observed during training.

Frame-of-Reference Training

The stages of the training program were broken down into separate components to explore the unique contribution of each stage. The six components were operationalized and organized as follows:

1) a job description was distributed and trainees participated in a discussion of the duties and qualifications necessary to perform the job of lecturer;

2) the Instructor Evaluation Scale was distributed and the definitions of the dimensions and the scale itself were discussed;

3) participants viewed two videotaped lectures on either CS and the SFP and individually practiced rating the behaviors of the lecturers on the Costin (1974) and Murphy et al. (1982) rating scales. Each videotaped lecturer presented either an outstanding or poor performance.

4) after rating behaviors, trainees individually wrote out justifications, or behavior rationales, for each rating.

5) based on expert scores, trainees received feedback on the correct ratings for each vignette.

6) The experts' behavioral rationales for the true scores were also relayed. This was accomplished by actually playing back the video when necessary to demonstrate certain behaviors and verbally describing behaviors. The discussion then focused on any discrepancies between the correct ratings and the participants' ratings. Particular attention was paid to types of behaviors.

Overall, the procedure for FOR training followed as closely as possible to the original proposal of Bernardin and Buckley (1981) and to that which has been conducted by McIntyre et al. (1984) and Athey and McIntyre (1987). However, some differences should be noted to clearly differentiate the training here from that conducted elsewhere. These distinctions are important since they can affect differences in results and were analyzed separately, as were all the components. In McIntyre's research the last component of FOR training, which involved discussion of the behavioral rationales and the discrepancies

between the experts' and the trainees' ratings, was omitted. Here, as proposed by Bernardin and Buckley (1981) in one condition trainees received all the training components, whereas in the other five conditions the components were systematically manipulated. Second, Bernardin and Buckley (1981) suggest the presentation of three levels of performance (high, average, and low) during the practice sessions. In McIntyre's research trainees viewed only one practice vignette of an undetermined level of performance. Trainees in this study practiced with two levels of performance (high and low) in order to demonstrate two different levels of behavior or frames-of-reference. Additionally, trainees viewed both lecture topics during the practice sessions and both lecture topics during posttest performance evaluations. Both these aspects allowed for the evaluation of the effects of topic, performance level, and the interaction of topic and performance level on posttest ratings. Unfortunately, taking into consideration the time limits of the average class, more than two practice sessions were not feasible. Third, an additional component was added to the training which provides for the presentation and discussion of the rating scale before viewing the behavior vignettes. While this component differs from the original FOR proposal, it is consistent with the work of McIntyre and others, and appears to be beneficial to effective training by increasing trainees familarity with the subsequent behavioral dimensions requiring judgement.

Additionally, one of the goals here was to maintain accordance with McIntyre's research whenever possible.

Based on these components the training conditions were comprised of the following.

Condition 1. Trainees received all levels of training

Condition 2. Trainees received all levels but 6 (discussion and behavioral rationales - they did receive true scores)

Condition 3. Trainees received all levels but 5 and 6 (no feedback of true scores or discussion)

Condition 4. Trainees received all levels but 4 (did not write out behavioral rationales)

Condition 5. Trainees received all levels but 2 (discussion of the rating scale)

Condition 6. Trainees received all levels but 1 (job description and discussion of duties and qualifications)

All trainees received the same introduction and instructions to the study (See Appendix B). Briefly, they were told that they were being asked to participate in a study on performance appraisal training. Performance appraisals were defined and their attention was directed towards situations where they might have had experience with a performance appraisal. Discussion then focused on teacher evaluations and four purposes were detailed. These were: 1) to provide a medium for the student's point of view; 2) to help identify a teacher's strengths and weaknesses; 3) for tenure decisions, or in the case of adjunct

faculty (whom they would be dealing with) to help determine whether or not they would be rehired.

Trainees were told they were taking part in developing a training program in an attempt to increase their accuracy in conducting student evaluations, and the training program would eventually be used by the College. As part of this study, four candidates for part-time faculty positions had agreed to let the College videotape them while presenting a lecture on one of two topics. The candidates had a minimum of two weeks to prepare their lecture, and had been told that students would be evaluating their performance. The evaluations would be fed back to themselves and to the Chairman of the Psychology Department. The trainees were told that their ratings would, in effect, have an influence on the potential hiring of the candidates they were evaluating. This point was reinforced during the course of the study. Finally, they were briefed as to the overall procedure for the study.

Anecdotal evidence occured that can be interpreted as a manipulation check of the trainees' belief in the elements of the training program and their belief that the lecturers were real job candidates. Often during the discussion session trainees would speak about the lecturers with true conviction that they existed, saying that they seemed nervous or expressing concern over whether they would get the job, or asking further questions on how long they had had to prepare, whether or not they had

taught before, and so on. On a number of occassions trainees approached the researcher after training and ask if they could leave a personal note with some feedback for the lecturer. Additionally, while trainees were debriefed they were not initially told that the lecturers were fictitious job candidates. Weeks after training, many trainees approached the researcher and inquired as to which candidate had been hired by the College or whether the following semester more students would be trained in conducting performance appraisals. Only three trainees ever asked if the lecturers were real job candidates and this occured after the collection of posttest information. Based on this information it appears that in most cases trainees were properly motivated and invested in the process of performance appraisal training, and completed the ratings to the best of their abilities.

Level of Performance

The videotaped lectures used here have been used extensively in previous performance appraisal research and, as discussed above, have been shown to have clearly identifiable performance differences across and within vignettes (Athey & McIntyre, 1984; McIntyre et al., 1984; Murphy et al., 1982; Murphy, Balzer, Kellam, & Armstrong, 1984). In a pilot study 14 graduate students, blind to the manipulation of performance level, correctly identified the high and low performance dimensions across and between vignettes (Murphy et al., 1984). Murphy et al. (1982) reports a high median correlation across tapes between

mean expert ratings and intended true scores of $\underline{r}$ = .84 for a performance evaluation scale and $\underline{r}$= .89 for the Behavioral Frequency Scale discussed above. Using these tapes allowed greater generalizability of the results of this study. Specific to this researcher's purpose, the tapes have also been used by McIntyre and his associates, therefore, the results of this project can be used in direct comparison to that previous research.

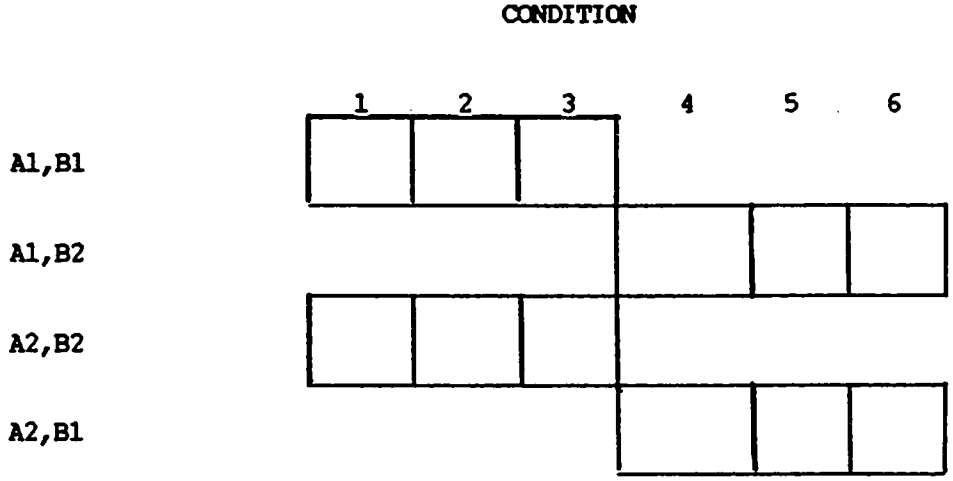
As discussed above (See Lecturer Vignettes) the videotapes can be grouped into outstanding, average, and low performance levels based on the "true scores" generated by the expert raters. During training, trainees viewed one lecturer of low performance and a different lecturer who demonstrated high performance. The order of presentation for level of performance remained constant across conditions such that trainees always viewed the high performer first and the low performer second, regardless of topic.

Lecture Topic

The lectures also differed by topic which included both CS and the SFP. Topics were counterbalanced in the training program such that one group in each condition viewed the CS lecture first and the SFP lecture second, while the second group received the opposite order of presentation. In the collection of posttest ratings, vignette topics were also counterbalanced such that one class received the SFP lecture first and the CS lecture second. The second class received the opposite order. All classes viewed medium levels of performance during the posttest.

Design Summary

To summarize the design, within each of six training conditions the presentation was as follows.



Where:

Pre-test pairing order (A):

A1= (high CS, low SFP)

A2= (high SFP, low CS)

Post-test pairing order (B):

B1= (medium SFP, medium CS)

B2= (medium CS, medium SFP):

and, conditions are as described above.

This design allows for the examination of main effects and two-way interactions, but it precludes analyses of three-way interactions.

Demographics

A demographic questionnaire (See Appendix C) was administered in order to investigate the possible effects of individual level variables. The questionnaire consisted of 15 items which elicit information typical to these surveys (e.g., age, academic major, etc.), as well as information on areas that might in some way affect the training experience. These items included questions related to their perceptions of stress and crowding, prior knowledge of the topic areas, and prior experience with performance appraisal ratings. This questionnaire was administered at the end of the study.

Trainee Attitude

A six item attitude questionnaire was administered after every rating occassion. This questionnaire was included to solicit information about the lecturer and lecture topic separate from their performance (See Appendix D). This was included to investigate attitudes that could influence the ratings and pertained mostly to the lecture topic itself. Two questions specifically addressed the lecturer.

Dependent Variables

Accuracy

Seven different accuracy operationalizations were used in this study in order to compare the different types and their various operationalizations. Two of these indices, absolute value halo and leniency accuracy, investigate aspects of error

but do so with the knowledge of true scores and thus can be interpreted as accuracy/error indices. The formulae are in Appendix F.

Elevation (EL). Accuracy of the average rating, over all ratees and dimensions (Cronbach, 1955). This measure investigates the accuracy with which the rater approximates true score estimates of the overall performance of ratees. (See formula 1).

Differential Elevation (DE). Accuracy of the mean rating for each ratee across all job dimensions (Cronbach, 1955). This investigates the distinctions among ratees in overall performance. (See formula 2).

Stereotype Accuracy (SA). Accuracy of the average rating given to each job dimension across all ratees (Cronbach, 1955). This investigates the accuracy of a rater's ratings for that dimension. (See formula 4).

Differential Accuracy (DA). Accuracy with which ratees are rank ordered on a given dimension (Cronbach, 1955). This investigates differences among ratees in patterns of performance. (See formula 6).

Distance Accuracy (DISTA). Absolute average deviation of trainees ratings from true scores (McIntyre et al., 1984). It is sometimes refered to as the square block metric and investigates how close or far away the trainees are from the experts. (See formula 8).

Absolute Value Halo (ABVH). The absolute value of the variance between observed and true halo for each ratee prior to averaging across ratees (McIntyre et al., 1984; Vance et al., 1978). This measure investigates the amount of halo trainees' evidence in relation to the amount of halo experts' determine in the ratees performance (See formula 11).

Leniency Accuracy (LENA). The mean true score for each dimension subtracted from the mean rating for each dimension across ratees for each rater. A high positive difference score would indicate greater leniency (McIntyre et al., 1984). (See formula 12).

Error

Five different measures of error were used in this study in order to investigate aspects of rating effectiveness that might be differentially identifiable from accuracy. Various error indices were selected based on their frequency in the research and the uniqueness of the information they contributed.

Halo. Two different operationalizations of halo were used in this study in order to compare the different types.

Variance Halo. The variance of a particular rater's ratings of a particular ratee across all dimensions (Saal et al., 1980).

ANOVA Halo. A Rater x Ratee x Dimension ANOVA, which would indicate halo through the absence of a Ratee x

Dimension interaction. This investigates the extent to which a rater considers each dimension separately in rating ratees (Borman, 1979).

Leniency. One operationalization of leniency that leniency that is common in the literature was investigated.

Midpoint. Comparison of a rater's mean dimension ratings with scale midpoints. This investigates whether raters' ratings center on average ratings across ratees (Saal et al., 1980).

Interrater Reliability. Comparison of the variance of ratings assigned to a particular ratee by raters within a condition for each dimension. This investigates convergent validity across raters and how well they discriminate across ratees (Borman, 1979).

Memory Test

This was a test of the behavior observed on the performance vignettes. This was operationalized as a measure of behavioral frequency developed by Murphy et al. (1982a) specifically for use with the videotapes employed in this research. Eight graduate students content analysed the videotapes "listing all teacher behaviors that contributed, negatively or positively, to their overall evaluation of each ratee." After streamlining the list of behaviors to reduce redundancy and ambiguity, the remaining 30 behaviors were rated for clarity and relevance. Then, using these items 45 undergraduates rated the videotapes. The data were then factor analyzed. After dropping low items two

dimensions remained which were labeled speaking style and clarity. Two additional items were dropped which left 10 items remaining in the Behavior Frequency Scale. Cronbach's alpha are .89 and .86, respectively for each dimension. This rating scale was administered during the posttest only. It was used as a measure of knowledge and/or memory for what was seen during the vignettes (See Appendix E).

Subjects

In all there were 250 trainees across the six conditions. Each condition was comprised of two randomly selected classes. The smallest condition contained 38 trainees and the largest had 49. There were 96 males and 154 females with a modal age of 20-22 years. The majority of the sample were night students accounting for 62.2% (n=156) of the total and 36.4% (n=91) were day students. Thirty-four percent (n=85) of the day students were enrolled full-time and 27.2% (n=68) of the night students also had full-time status. The class breakdown was fairly evenly divided with seniors accounting for the smallest percentage (20.8%, n=52) and juniors achieving the largest (26.8%, n=67). The largest percentage of the sample was employed full-time (56.4%, n=141), and only 8% (n=20) were not employed at all. In terms of familiarity with performance appraisals 67.2% (n=168) had participated in their own performance appraisal and 39.2% (n=98) had formally reviewed another's performance. The largest

majority of the sample had no prior formal knowledge of either stress and crowding or the self fulfilling prophecy (60%, n=150, and 63.2%, n=158, respectively).

Data Analyses

Data analyses were conducted in several stages after the development of true scores. As a first step, the measurement situation was analyzed to determine the presence of topic and order effects. Second, the error and accuracy indices were calculated and used to investigate the specific hypotheses of this study as they relate to differences among the training conditions. Third, memory accuracy indices were developed and used to determine differences in accuracy as affected by differences in memory, or knowledge, for what was observed on the lecture vignettes. Fourth, the relationships among the various error and accuracy indices themselves were investigated to determine whether the different operationalizations of the same constructs evidenced the same results. Fifth, individual differences among trainees were studied to determine any potential influence these may have had on the results.

Topic and Order Effects

The influence of aspects of the measurement situation were assessed separately for topic and pretest and posttest order. Posttest topic effects were first analyzed as a 2 x 6 x 2 (Pretest x Condition x Topic) factorial design using a multivariate analysis of variance (MANOVA) with posttest topic as

a repeated measure and condition as a between groups factor. Topic effects were also analyzed as 2 x 6 (Pretest x Condition) MANOVA with topic and dimension ratings as within subject factors and pretest topic/performance level and training condition as between group factors. This allowed for the analysis of differences in dimension ratings between topics. Differences in posttest topic ratings were investigated separately for main effects and interactions of topic, pretest topic/performance level, posttest order, conditions, and behavioral dimension ratings. When significant main effects and interactions were demonstrated they were explored further through post hoc analyses.

Order effects were analyzed separately by lecture topic and for the two lecture topics combined, to examine pretest and posttest order main effects, and interactions between pretest order, posttest order and conditions. Three MANOVA between-subjects designs were employed for each topic. These included a 2 x 6 (Posttest order x Training condition), a 2 x 2 (Pretest topic order/level of performance x Posttest order), and a 2 x 6 (Pretest order x Training condition) MANOVA with dimension ratings as the dependent measure. When significant main effects and interactions were demonstrated they were explored further through post hoc analyses.

Error and Accuracy

Error and accuracy indices were calculated as defined above. For the training condition comparisons, planned contrasts were

performed to investigate the specific hypotheses above. The error indices were investigated using analysis of variance and Student t-tests. When significant differences between conditions occured these were explored further through post hoc analyses.

The relationships between the different accuracy measures were investigated through indice intercorrelations. Additionally, where differences between conditions and indices existed these were explored further.

Individual Differences

When significant differences on individual level variables occured across or between groups these were used as a grouping factor on subsequent analyses. In this manner subject differences were explored to determine their influence on the results.

Memory Effects on Accuracy

Accuracy measures on the Behavior Frequency Scale were trichotomized into high, medium, and low scores based on the cummulative frequency of accuracy values for each category and in this manner operationalized as memory accuracy indices. For the analysis of Hypothesis 4, a MANOVA was conducted for each memory accuracy index. The dependent measures in the MANOVA's were the seven accuracy indices developed from the Instructor Evaluation Scale and the independent factor was each memory index which then consisted of three levels. Therefore, a separate MANOVA was conducted for each memory index.

CHAPTER IV

RESULTS

The Measurement Situation

Overview

The analyses pertaining to the investigation of elements of the measurement situation were exploratory in nature. Since little is known about the effects of topic and pretest order/level of topic presentation and posttest topic order it was important to investigate whether these elements of the measurement situation affected trainees posttest performance appraisal ratings. Specific questions were investigated for each of these two design elements (e.g., topic and topic order/level of presentation). Topic effects, and order effects for pretest order/level of topic presentation and posttest order effects will be presented separately.

Topic Effects

Topic effects were first analyzed through a 2 x 6 x 2 (Pretest x Condition x Topic) factorial design using a doubly repeated measures MANOVA design with posttest topic as a repeated measure factor and condition and pretest topic/level as between group factors. The dependent measures for this analysis were the 15 dimension ratings of a given posttest lecture topic. Table 7 contains the results of this analysis. Table 8 provides the means for the posttest ratings broken down by behavioral dimension, topic, and condition.

Table 7

Multivariate Analysis of Variance Results of Topic Effects

Effect	Wilk's Lambda	F
Topic	.50519	14.627***
Training Condition	.64739	1.365*
Training Condition x Topic	.62982	1.455**
Pretest Order x Topic	.85771	2.477**

* p<.05. ** p<.01. *** p<.001.

Table 8

Mean Dimension Ratings by Posttest Topic and Condition

	CS					
	Condition					
Dimension	1	2	3	4	5	6
Seemed interested in topic	2.29	2.86	2.42	2.42	2.55	2.69
Used clear examples	2.26	2.86	2.67	2.42	2.90	2.57
Presented lecture smoothly	2.26	3.40	3.23	3.02	3.10	3.08
Integrated material effectively	2.39	3.14	3.06	2.65	2.97	2.90
Followed an outline	1.97	2.40	2.25	1.97	1.85	2.39
Followed a logical sequence of thought	2.05	2.86	2.58	2.32	2.42	2.58
Encouraged questions	3.03	4.17	3.94	3.75	3.65	3.96
Provided relevant answers to questions	2.42	3.09	2.69	2.70	2.85	2.65
Well prepared	2.00	2.94	2.56	2.22	2.63	3.04
Acted relaxed	2.18	3.29	2.98	2.90	3.47	3.10
Spoke clearly and distinctly	1.82	2.91	2.83	2.27	2.92	2.76
Spoke with rigor	2.76	3.37	3.21	3.10	3.20	3.18
Emphasized important points	2.47	3.31	3.13	3.07	2.97	2.84
Made you interested	2.26	3.66	2.85	3.02	3.13	3.06
Overall	2.76	3.83	3.54	3.38	3.38	3.37

Table 8 continued

	SFP					
	Condition					
Dimension	1	2	3	4	5	6
Seemed interested in topic	2.71	2.43	2.33	2.30	2.40	2.49
Used clear examples	2.39	2.57	2.46	2.17	2.65	2.53
Presented lecture smoothly	2.45	2.60	2.40	2.25	2.55	2.37
Integrated material effectively	2.45	2.77	2.50	2.42	2.22	2.49
Followed an outline	2.42	2.23	2.19	1.90	1.82	2.20
Followed a logical sequence of thought	2.58	2.43	2.09	2.10	2.32	2.47
Encouraged questions	3.47	4.14	3.75	3.75	3.70	3.78
Provided relevant answers to questions	4.17	3.77	3.73	3.77	4.47	3.61
Well prepared	3.16	2.66	2.70	2.35	2.85	2.47
Acted relaxed	2.84	2.43	2.35	2.07	2.65	2.33
Spoke clearly and distinctly	2.26	2.20	2.08	2.05	2.35	2.33
Spoke with rigor	2.74	2.40	2.35	2.27	2.45	2.73
Emphasized important points	2.57	2.23	2.46	2.13	2.25	2.47
Made you interested	2.92	2.74	2.54	2.85	2.65	2.92
Overall	3.58	3.09	3.15	2.80	2.97	3.06

The overarching question regarding the effect of lecture topic (CS or SFP) was whether there were differences in ratings attributable to topic. The MANOVA yielded significant Wilk's lambdas for the main effects of topic (.50519, F=14.627, $p < .001$) and training condition (.64739, F=1.365, $p < .024$) and for the interaction effect of training condition and topic (.62982, F=1.45530, $p < .008$). The topic main effect reveals that trainees were consistently rating the two topics differently even though the topics represent similar levels of performance. Overall, the CS lecture received higher posttest ratings than the SFP lecture. However, this difference can be somewhat explained through differences between the experts' true score ratings. While the experts rated these vignettes similarly the CS lecture was slightly higher than the SFP lecture. By looking at Table 8 the interaction between training condition and topic is clear. In Condition 1, which contained all components of FOR training, the CS lecture received lower mean ratings than the SFP lecture, in all other conditions the CS lecture received higher mean ratings than the SFP lecture. Additionally, the order of the condition differences were not consistent across the two topics. A rank ordering of the overall mean ratings demonstrates that for the CS lecture training Condition 2 resulted in the highest ratings ($\bar{x}$ = 3.21) followed by Condition 6 ($\bar{x}$ = 2.94), Conditions 3 and 5 demonstrated identical overall mean ratings ($\bar{x}$ = 2.93), Condition 4 had the second lowest overall mean ratings ($\bar{x}$ = 2.75) and

Condition 1 had the lowest ($\bar{x}$ = 2.33). Conversely, for the SFP lecture Condition 1 resulted in the highest overall mean ratings ($\bar{x}$ = 2.79) followed by Conditions 2 ($\bar{x}$ = 2.71), 5 ($\bar{x}$ = 2.69), 6 ($\bar{x}$ = 2.68), 3 ($\bar{x}$ = 2.60), and 4 ($\bar{x}$ = 2.60). Thus, it appears that while there was a significant main effect for training condition the ordering of the condition differences were not consistent across the two topics.

Another question of interest here was whether posttest topic ratings varied with the pairing of lecture topic/level of performance observed at pretest. It will be recalled that pretest training topics were counterbalanced by level of lecturer performance (high and low) in such a manner that in each condition one group of trainees viewed a high CS lecture performance and a low SFP lecture performance while the second group in the same condition viewed a high SFP lecture performance and a low CS lecture performance. In either case the speaker was the same. Collapsing over posttest order, the MANOVA results in Table 7 indicate that there was a significant Pretest order x Topic interaction (Wilk's lambda= .85771, F=2.447, $p < .002$). Posttest topic mean ratings differed with the topic/level of performance observed during training. Table 9 lists the dimension means for each topic broken down by pretest order. It can be seen that in pretest order A1 the CS posttest ratings were highly elevated compared to the SFP posttest ratings. However, in pretest order A2 the SFP posttest ratings achieved parity with

Table 9

Mean Dimension Ratings by Posttest Topic and Pretest Order

	Pretest Order			
	A1	A2	A1	A2
Dimension	CS		SFP	
Seemed interested in topic	2.93	2.12	2.54	2.34
Used clear examples	2.93	2.28	2.66	2.26
Presented lecture smoothly	3.46	2.57	2.65	2.20
Integrated material effectively	3.23	2.48	2.73	2.20
Followed an outline	2.45	1.83	2.37	1.88
Followed a logical sequence of thought	2.79	2.14	2.61	2.02
Encouraged questions	4.21	3.28	4.22	3.27
Provided relevant answers to questions	3.24	2.18	4.26	3.52
Well prepared	3.16	1.97	3.04	2.31
Acted relaxed	3.47	2.49	2.63	2.23
Spoke clearly and distinctly	3.06	2.11	2.39	2.02
Spoke with rigor	3.48	2.79	2.53	2.46
Emphasized important points	3.21	2.70	2.41	2.30
Made you interested	3.45	2.51	2.85	2.68
Overall	3.70	3.04	3.17	3.02

Note. Trainees in Pretest Order A1 viewed a high CS lecture and then a low SFP lecture. Trainees in Pretest Order A2 viewed a high SFP lecture and then a low CS lecture.

the CS posttest ratings. Therefore, it appears that there was a context effect. Whichever topic was placed in the high performance category at pretest was being rated highly favorably at posttest. Additionally, given the topic main effect which demonstrated higher overall mean performance ratings for the CS lecture topic, by comparing SFP ratings across pretest orders it can be seen that the high condition at pretest seemed to elevate the posttest ratings of whichever topic was put in this condition.

To further explore trends in the differences found between topics a 2 x 6 (Pretest x Condition) repeated measures MANOVA was computed with topic and dimension ratings as within subjects factors, and pretest topic/performance level and training conditions as between group factors. The question of interest here was whether there was a significant difference in dimension ratings between the two topics. The results demonstrated a significant dimension main effect (Wilk's lambda =.27316, $F=42.76$, $p<.001$) and a Topic x Dimension interaction (Wilk's lambda =.5167, $F=15.03$, $p<.001$). Table 10 lists the mean ratings for the dimension ratings by topic. By reviewing Table 10 it can be seen that of the 15 dimensions, 12 resulted in higher mean ratings for the CS lecture, whereas two dimensions demonstrated higher mean ratings for the SFP lecture, and one dimension had equal ratings across the two topics. The dimensions that received the higher mean ratings for the SFP lecture were "he provided relevant answers to questions" and "he was well

Table 10

Mean Dimension Ratings by Topic

	Topic	
Dimension	CS	SFP
Seemed interested in topic	2.54	2.44
Used clear examples	2.61	2.46
Presented lecture smoothly	3.02	2.43
Integrated material effectively	2.86	2.47
Followed an outline	2.15	2.13
Followed a logical sequence of thought	2.47	2.33
Encouraged questions	3.76	3.76
Provided relevant answers to questions	2.72	3.90
Well prepared	2.58	2.69
Acted relaxed	2.99	2.44
Spoke clearly and distinctly	2.60	2.21
Spoke with rigor	3.14	2.50
Emphasized important points	2.96	2.36
Made you interested	2.99	2.77
Overall	3.38	3.10

prepared." This result is consistent with the design of the vignettes, while the tapes are matched on overall level of performance, their profiles across dimensions are not matched.

Order Effects

The influence of pretest topic order/level of performance and posttest topic order on posttest mean dimension ratings were analyzed separately for each of the posttest lecture topics. It should be recalled that in investigating pretest and posttest order effects in this study pretest topic order itself could not be separated from the level of performance viewed for each pretest topic. In all pretest conditions the first lecture viewed always depicted a high performance and the second lecture always depicted a low performance. It is only at posttest that the performance levels were equal and independent of topic. In this manner it was possible to assess condition and pretest order/level of performance effects, as well as posttest topic order effects.

For each topic, order effects were analyzed in a 2 x 6 (Posttest order x Training condition), a 2 x 2 (Pretest topic order/level of performance x Posttest order), and a 2 x 6 (Pretest order x Training condition) MANOVA with dimension ratings as the dependent measures. Table 11 reports the results of these analyses.

Table 11

Multivariate Analysis of Variance Results of Order Effects for CS and SFP

Effect	Wilk's Lambda	F
CS		
Pretest Order	.74418	5.13**
Posttest Order	.92156	1.32
Training Condition	.64433	1.38*
Pretest Order x Training Condition	.68603	1.18
Posttest Order x Training Condition	.55282	1.89**
Pretest Order x Posttest Order	.93380	1.10
SFP		
Pretest Order	.76328	4.79**
Posttest Order	.92018	1.34
Training Condition	.62157	1.50*
Pretest Order x Training Condition	.65091	1.35*
Posttest Order x Training Condition	.53826	1.98**
Pretest Order x Posttest Order	.89075	1.90*

*p<.05. **p<.01.

The first question of interest here was whether the order of viewing tapes at pretest affected posttest ratings. The results demonstrated a significant pretest main effect for both the CS lecture (Wilk's lambda =.7442, F =5.13, $p<.01$) and the SFP lecture (Wilk's lambda =.7633, F =4.79, $p<.01$). It appears that the pretest order did influence posttest ratings. It can be seen in Table 9 that those in pretest condition A1 (high CS, low SFP), in general, demonstrated higher posttest ratings regardless of posttest topic. It may be the case that the slightly elevated performance of the high CS lecture noted by the experts is permeating all other ratings such that those trainees that viewed the CS lecture first are holding on to a context effect and thus raising all subsequent ratings.

The results of the MANOVA did not indicate a significant posttest order main effect for either posttest topics. It does not seem to matter whether trainees rated the CS lecture or the SFP lecture first at posttest. Therefore it is reasonable to conclude that there were no carryover or practice effects. There was a significant Pretest x Posttest interaction for the SFP lecture, yet the interpretation of this interaction is unclear. Additionally, none of the univariate F's were significant for any of the dimension ratings. So while the interaction appeared significant it seems to have been an artifact of the analysis and is difficult to explain. This pattern of results was true for the interaction of pretest and conditions as well and is equally uninterpretable.

As indicated in Table 11, there were significant Condition x Posttest interactions for both lectures, indicating that the differences in ratings across conditions varied with posttest order. The nature of this interaction can be seen through Table 12. Trainees who received posttest order B1 consistently rated both lectures more severely in the fourth through sixth conditions whereas those who received the opposite order rated both lectures more leniently in these conditions. However, while the pattern of results is clear, interpretation is difficult.

Hypothesis Testing

Condition Effects

The three hypotheses investigated in this stage of the analyses were that: 1) training that included all the components of FOR training would lead to more accurate results and less error than those conditions that did not include all FOR training components; 2) those training conditions that involved more active participation by the trainees would be more accurate and have less error than those that did not; and, 3) those conditions that involved less active trainee participation but included feedback of expert true scores and behavioral rationales would be more accurate and have less error than those that did not include these components.

Table 12

Mean of Topic Dimension Ratings by Condition and Posttest Order

	CS					
	Condition					
Dimension	1	2	3	4	5	6
Posttest Order B1						
Seemed interested in topic	2.61	3.27	2.76	2.29	2.26	2.17
Used clear examples	2.61	3.04	2.64	2.36	2.26	2.04
Presented lecture smoothly	2.83	3.69	3.44	2.82	2.74	2.39
Integrated material effectively	2.89	3.46	3.08	2.57	2.47	2.35
Followed an outline	2.28	2.69	2.32	1.79	1.47	2.04
Followed a logical sequence of thought	2.39	3.08	2.60	2.29	1.79	2.14
Encouraged questions	3.72	4.15	4.20	3.54	3.00	3.23
Provided relevant answers to questions	2.89	3.35	3.16	2.50	1.74	2.26
Well prepared	2.50	3.23	2.84	1.89	1.68	2.30
Acted relaxed	2.67	3.46	3.04	2.50	2.84	2.30
Spoke clearly and distinctly	2.39	3.15	2.92	2.04	2.21	2.17
Spoke with rigor	3.06	3.73	3.36	3.04	2.63	2.78
Emphasized important points	2.72	3.65	3.04	3.11	2.16	2.70
Made you interested	2.61	4.12	3.08	2.93	2.26	2.65
Overall	2.89	4.08	3.76	3.29	2.79	3.00

Table 12 continued

Dimension	CS Condition 1	2	3	4	5	6
Posttest Order B2						
Seemed interested in topic	2.00	1.67	2.04	2.75	2.81	3.15
Used clear examples	1.95	2.33	2.70	2.58	3.48	3.04
Presented lecture smoothly	1.75	2.56	3.00	3.50	3.43	3.69
Integrated material effectively	1.95	2.22	3.04	2.83	3.43	3.38
Followed an outline	1.70	1.56	2.17	2.42	2.19	2.69
Followed a logical sequence of thought	1.75	2.22	2.57	2.42	3.00	2.96
Encouraged questions	2.40	4.22	3.65	4.25	4.24	4.58
Provided relevant answers to questions	2.00	2.33	2.17	3.17	3.86	3.00
Well prepared	1.55	2.11	2.26	3.00	3.48	3.69
Acted relaxed	1.75	2.78	2.91	3.83	4.05	3.81
Spoke clearly and distinctly	1.30	2.22	2.74	2.83	3.57	3.27
Spoke with rigor	2.50	2.33	3.04	3.25	3.71	3.54
Emphasized important points	2.25	2.33	3.22	3.00	3.71	2.96
Made you interested	1.95	2.33	2.61	3.25	3.90	3.42
Overall	2.65	3.11	3.30	3.58	3.90	3.69

Table 12 continued

Dimension	SFP Condition 1	2	3	4	5	6
Posttest Order B1						
Seemed interested in topic	2.89	2.54	2.28	2.29	2.37	2.22
Used clear examples	2.82	2.73	2.32	2.11	2.26	2.39
Presented lecture smoothly	2.83	2.65	2.40	2.04	2.37	2.04
Integrated material effectively	2.89	2.85	2.64	2.18	1.95	2.26
Followed an outline	2.61	2.46	2.20	1.68	1.58	1.87
Followed a logical sequence of thought	2.94	2.62	2.16	2.07	1.89	1.96
Encouraged questions	4.28	4.15	4.24	3.64	3.16	3.13
Provided relevant answers to questions	5.06	3.88	4.12	3.44	3.95	3.70
Well prepared	3.72	2.92	2.92	2.25	2.11	2.30
Acted relaxed	3.00	2.50	2.28	1.86	2.37	1.96
Spoke clearly and distinctly	2.56	2.35	2.08	1.86	2.26	2.09
Spoke with rigor	3.06	2.35	2.12	2.21	2.47	2.57
Emphasized important points	2.72	2.31	1.96	1.96	2.00	2.26
Made you interested	3.28	2.88	2.28	2.71	2.63	2.74
Overall	3.67	3.12	3.04	2.79	2.63	3.04

Table 12 continued

Dimension	SFP Condition 1	2	3	4	5	6
Posttest Order B2						
Seemed interested in topic	2.55	2.11	2.39	2.33	2.43	2.73
Used clear examples	2.00	2.11	2.61	2.33	3.00	2.65
Presented lecture smoothly	2.10	2.44	2.39	2.75	2.71	2.65
Integrated material effectively	2.05	2.56	2.35	3.00	2.48	2.69
Followed an outline	2.25	1.56	2.17	2.42	2.05	2.50
Followed a logical sequence of thought	2.22	1.89	2.00	2.17	2.71	2.92
Encouraged questions	2.67	4.11	3.22	4.00	4.19	4.35
Provided relevant answers to questions	3.28	3.44	3.30	4.50	4.95	3.54
Well prepared	2.65	1.89	2.45	2.58	3.52	2.62
Acted relaxed	2.70	2.22	2.43	2.58	2.90	2.65
Spoke clearly and distinctly	2.00	1.78	2.09	2.50	2.43	2.54
Spoke with rigor	2.45	2.56	2.61	2.42	2.43	2.88
Emphasized important points	2.42	2.00	3.00	2.50	2.48	2.65
Made you interested	2.58	2.33	2.83	3.17	2.67	3.08
Overall	3.50	3.00	3.26	2.83	3.29	3.08

Note. Trainees in Posttest Order B1 received the CS lecture first and the SFP lecture second. Trainees in Posttest Order B2 received the SFP lecture first and the CS lecture second.

Accuracy

Seven different operationalizations of accuracy were investigated. It is commonly agreed in the literature that there are many different components to accuracy, and it is suggested that each component captures different information about the "correctness" of an individual's or groups' performance ratings. Or, that in comparison to some known "true score" performance ratings may be correct for one operationalization of accuracy and incorrect for another. For the purposes of this research seven of the most common conceptual and operational definitions of accuracy were used to test hypotheses 1 through 3. The accuracy indices assessed were the four Cronbach accuracy measures, namely DA, SA, DE, and EL, and DISTA, LENA, and ABVH (see pp. 91-92). Accuracy indices were calculated over ratees. That is, the indices were collapsed over topic since accuracy is typically assessed via multiple ratees.

In relation to accuracy, a post hoc decision was made to investigate these hypotheses in two different ways. For the purposes of locating differences in accuracy results among the training conditions accuracy indices were first analyzed through the aggregate scale measures. That is, an accuracy index was calculated for each of the seven measures using all 15 performance appraisal behavioral dimensions on the Instructor Evaluation Scale. However, it was also believed that each of the 15 behavioral dimensions were looking at unique aspects of

performance. To collapse over these dimensions might therefore obscure the detection of raters' abilities to accurately rate behavior if certain dimensions or combinations of dimensions were more easily detected or rated than others. Based on this assumption a second method of analyzing the condition effects on accuracy was to break the 15 behavioral dimensions into three different factor. Thus the ratings for each factor were also calculated for each of the seven accuracy indices. The factorial structure was developed a priori based on the experts' behavioral rationales for each scale dimension. The behavioral rationales were content analyzed and grouped according to the behavioral aspects attended to in arriving at their ratings. Three factors emerged that shared common behavioral rationales. The first factor contained items where the behavioral rationale used to score them referred to personal aspects of the lecturer, for example, facial expression, tone and body language. The second factor contained items where the behavioral rationales used for rating judgements referred to more cognitive aspects of the lecturer's performance, for example, number of examples presented, transitions between topics, and clarity of presentation. The third factor contained items that shared both personal and cognitive behavioral rationales. The number of items in each factor of the Instructor Evaluation Scale were 6, 5, and 3 respectively. The listing of items for each factor can be found in Table 13.

Table 13

A Priori Factor Structure of the Instructor Evaluation Scale

Dimensions
Factor 1 - Personal Aspects
1. He seemed interested in the topic
7. He encouraged questions
10. He acted relaxed
12. He spoke with vigor and enthusiasm
13. He emphasized important points by raising his voice
14. He made you interested in the material
Factor 2 - Cognitive Aspects
2. He used clear examples to explain abstract ideas
4. He integrated the material effectively
6. He followed a logical sequence of thought in his lecture
8. He provided relevant answers to questions
11. He spoke clearly and distinctly
Factor 3 - Personal and Cognitive Aspects
3. He presented the lecture smoothly
5. He followed an outline
9. He was well prepared

To analyze the effects of the six training conditions on accuracy, planned contrasts were performed. In this analysis the sum of squares associated with the different conditions were partitioned into separate contrasts, thus resulting in six different sum of squares, each with 1 degree of freedom. Simple contrasts were then performed comparing the different conditions against each other with the accuracy measures entered into the equation as dependent measures as specified in hypotheses 1 through 3. For example, in Hypothesis 1 the accuracy indices for Condition 1 were contrasted with each of the accuracy indices for Conditions 2 through 6. Specifically, for each hypothesis two different series of MANOVA contrasts were performed. The four Cronbach accuracy dependent measures were entered into one MANOVA while the remaining three accuracy measures (e.g., DISTA, LENA, ABVH) were entered into a second MANOVA. The four Cronbach measures are the most common accuracy indices in the literature while the other three indices are more recent additions. In order to maintain the distinction between indices the separation of the accuracy measures into groups of four and three dependent measures was undertaken. For the investigation of each hypothesis the MANOVA contrasts procedure was undertaken four times. One series of planned contrasts were performed for the two sets of aggregate accuracy measures (e.g., those from the 15 dimension scale) and one series was conducted for the two sets of the trichotomized factor accuracy measures (e.g., those from each of the three factor scales).

Error

Oneway analyses of variances (ANOVA's) were employed to test differences in Variance Halo between Condition 1 and each of the other conditions. This variance estimate is the most common halo index in the performance appraisal training literature. Given its popularity it was decided to employ this index for the investigation of both aggregate and trichotomized factor scales. It was believed that the exploration of halo in both manners might demonstrate differences in the pattern of halo errors. The separation of the performance appraisal scale into facets of personal attributes of the lecturer's presentation, cognitive aspects of the lecture, and a third facet that encompasses both can lead to further understanding of where halo occurs in behavioral ratings, or what behavioral aspects of performance contribute to general impression effects. A variance estimate of halo, calculated within raters and ratees across dimensions, was entered as the dependent measure with the different training conditions as the independent factor as specified in the hypotheses. In two of the ANOVA's, halo estimates were considered separate for each topic and in the third ANOVA halo was collapsed over topic. This procedure was performed twice, once for the aggregate ratings and once for the trichotomized factored scale.

Another method of investigating halo was employed through performing separate Rater x Ratee x Dimension repeated measures MANOVA's for each condition. This measure was only investigated for the aggregate performance appraisal scale measures. Dimension ratings for each topic were entered as dependent measures with ratee as the independent factor. Topic and dimension ratings were entered as within subject factors. For each MANOVA, evidence of a lack of halo would be indicated by a significant Ratee x Dimension interaction. While this value is influenced by the total variance in a rater's ratings, it is useful through providing an individual difference measure reflecting the extent to which a rater considered each dimension separately in rating individual ratees (Borman, 1979).

Midpoint leniency was assessed through conducting a Training condition x Dimension x Topic MANOVA with repeated measures on dimensions and topics. A condition main effect would provide evidence of differences in leniency between conditions indicating that there were condition differences in relation to the scale midpoint. This measure was only investigated for the aggregate performance appraisal scale measure.

Interrater reliability was assessed for each topic by comparing the variances across raters for each behavioral dimension for within conditions. Differences between conditions were then assessed through the Hartley (1940, 1950) Homogeneity of Variance test which involved the calculation of an Fmax

statistic. This test has a slight positive bias which results in the rejection of homogeneity more often when condition sizes are unequal, as was the case here. This measure is based on analysis of individual dimensions.

The results for each hypothesis will be discussed separately.

HYPOTHESIS 1

Accuracy

This hypothesis was investigated by comparing Condition 1 which included the entire FOR training program against the other five training conditions where one or more components were absent. Within the MANOVA, five simple planned contrasts were performed for each accuracy index comparing Condition 1 against Conditions 2, 3, 4, 5, and 6. Pretest topic order and training conditions were entered as factors. Of the 10 multivariate tests performed on the condition contrasts (five contrasts per MANOVA, one MANOVA for the Cronbach measures and one for the other three indices) only Condition 3 evidenced a significant contrast with Condition 1 (Wilk's lambda =.9588, F =2.55, $p < .04$). The univariate analysis indicated a significant difference between the two conditions for DE ($F(1,238) = 7.886$, $p < .001$). Thus, the raters in Condition 1 were significantly more accurate than those in Condition 3 in regard to their ability to discriminate between raters in overall behavior ($\tilde{x} = .28$ vs. $\tilde{x} = .43$, respectively). Whereas Condition 1 contained all FOR training components, in Condition 3 the feedback of expert true scores and

behavioral rationales, and discussion was omitted. It appears that these components contributed toward the differentiation of a ratee's overall behavior. These components are what may be contributing to the development of a "frame-of-reference" and used to determine distinctions between performance. The multivariate analysis revealed a significant Wilk's lambda for the interaction between pretest effects and Conditions 1 and 2 (.9458, F= 4.5088, $p < .004$). The univariate analysis demonstrated a significant effect for ABVH ($F(1,238) = 12.75$, $p < .001$). A review of the mean breakdowns indicated that the raters in Condition 1 who received pretest order A1 (high CS, low SFP) evidenced greater correspondence with true scores regarding the amount of true halo (as defined by ABVH) present in the lecturer's performance than those in Condition 2 pretest order A1 ($\bar{x}$= .6511 vs. $\bar{x}$= .9014, respectively). Yet, this finding was reversed when raters received pretest order A2 (high SFP, low CS). In this situation those in Condition 2 had greater accuracy on this index than raters in Condition 1 ($\bar{x}$= .6657 vs. $\bar{x}$= .9791, respectively). It appears that the pretest effect found in the test of the measurement situation permeated the ABVH accuracy index. Trainees in Condition 2 did not receive feedback containing behavioral rationales or discussion. It may be that the distinctions in high and low performance in pretest order A1 were easier to detect. Therefore, when training involved pretest order A2 the information contained in the behavioral feedback and

discussion actually influenced those in Condition 1 to incorrectly judge interrelationships at posttest, while those in Condition 2 who formed their own behavioral rationale based frames-of-reference did not envision the same interrelationship between dimensions and were actually more accurate in regard to the abilities necessary for ABVH accuracy.

Out of five different comparisons on seven different dependent measures (e.g., 35 different possible results) only two reached statistical significance. While the significant results were in the predicted direction it cannot be considered strong evidence in support of Hypothesis 1, since these could have occured as Type I errors if the null hypotheses were true.

Turning to the results of the three factor accuracy indices the evidence is somewhat stronger in support of Hypothesis 1. For these analyses Condition 1 was contrasted with all other conditions and the 21 accuracy indices were entered simultaneously into each of the MANOVA's (e.g., the MANOVA for the Cronbach measure and the MANOVA for the additional accuracy indices) as dependent measures. Thus, the MANOVA's had 12 and nine dependent measures, respectively. Of the five multivariate tests performed on the condition contrasts for the Cronbach measures all were significant. Of the five multivariate tests performed on the condition contrasts for DISTA, LENA, and ABVH two were significant. Table 14 provides the results of these analyses. Investigation of the univariate results presented in Table 15

Table 14

Planned Contrast Results of the Trichotomized Accuracy Indices

Training Condition	Wilk's Lambda	F
	MANOVA 1[a]	
1 x 6	.89909	2.12**
1 x 5	.87428	2.72**
1 x 4	.87978	2.58**
1 x 3	.86797	2.88**
1 x 2	.91090	1.85
	MANOVA 2[b]	
1 x 6	.95113	1.31
1 x 5	.87803	3.55**
1 x 4	.95380	1.24
1 x 3	.87811	3.55**
1 x 2	.93695	1.72

[a] MANOVA 1 contained the trichotomized accuracy indices of DA, SA, DE, and EL resulting in 12 dependent measures.

[b] MANOVA 2 contained the trichotomized accuracy indices of DISTA, LENA, and ABVH resulting in 9 dependent measures.

* $p<.05$. ** $p<.01$.

Table 15

[a] F Values of Univariate Analyses of Condition Contrasts for Trichotomized Accuracy Indices

[b] Accuracy	Condition Contrasts				
	1 x 2	1 x 3	1 x 4	1 x 5	1 x 6
DA1	.70	5.14*	.05	2.08	1.26
DA2	.04	.80	.19	2.81	1.80
DA3	.50	1.68	.38	.35	2.09
SA1	.03	1.90	1.27	1.08	.51
SA2	.30	2.95	.02	.81	2.00
SA3	4.58*	1.19	5.14*	7.82**	2.36
DE1	8.14**	15.75**	11.52**	12.27**	8.54**
DE2	2.16	6.58**	1.34	1.40	1.40
DE3	13.76**	11.73**	17.31**	10.78**	15.37**
EL1	1.42	1.01	1.31	1.59	2.22
EL2	.80	2.68	1.88	1.91	1.11
EL3	.30	1.77	.47	.54	1.39
DISTA1	.91	.22	1.22	.09	1.43
DISTA2	.01	.05	.78	.13	.22
DISTA3	3.57	6.89**	4.80	5.92**	5.23
LENA1	1.47	1.08	1.29	1.56	2.18
LENA2	1.41	1.48	.47	3.90*	1.61
LENA3	.18	.71	.03	.07	1.03
ABVH1	3.99	9.93**	.26	2.31	.23
ABVH2	.01	2.31	.26	1.80	.06
ABVH3	1.02	1.59	.03	.23	.38

[a] F Value df=1,238.

[b] Accuracy Factor 1 includes the six items pertaining to personal aspects of the lecturer. Accuracy Factor 2 includes the 5 items pertaining to cognitive aspects of the lecture. Accuracy Factor 3 includes the 3 items pertaining to both personal and cognitive aspects of the lecture and lecturer.

* $p<.05$. ** $p<.01$.

Table 16

Trichotomized Accuracy Mean Scores[a]

	Condition					
Accuracy	1	2	3	4	5	6
DA1	.48	.52	.58	.48	.55	.53
DA2	.68	.69	.67	.65	.60	.71
DA3	2.06	2.39	2.26	2.05	2.18	2.30
SA1	.82	.80	.89	.75	.88	.79
SA2	.90	.94	1.00	.92	.96	.99
SA3	1.08	.90	.98	.89	.85	.95
DE1	.36	.77	.73	.70	.71	.63
DE2	.38	.53	.62	.52	.50	.49
DE3	.76	.39	.48	.37	.46	.44
EL1	2.26	1.84	2.06	2.11	1.20	1.96
EL2	1.22	.94	1.06	1.15	1.01	1.06
EL3	1.70	1.40	1.48	1.71	1.55	1.49
DISTA1	2.33	2.06	2.25	2.33	2.26	2.14
DISTA2	1.46	1.40	1.47	1.43	1.41	1.48
DISTA3	2.05	1.67	1.73	1.86	1.73	1.76
LENA1	-2.26	-1.83	-2.06	-2.11	-1.99	-1.96
LENA2	-1.16	-.74	-.94	-1.12	-.79	-.92
LENA3	-1.62	-1.30	-1.46	-1.71	-1.53	-1.41
ABVH1	.50	.64	.88	.57	.70	.56
ABVH2	1.10	1.13	1.22	1.08	.97	1.11
ABVH3	1.03	.93	.91	.99	.98	.97

Note. The accuracy scores are averaged across the two posttest tapes.

[a] The lower the value the greater the accuracy.

[b] Accuracy Factor 1 includes the six items pertaining to personal aspects of the lecturer. Accuracy Factor 2 includes the 5 items pertaining to cognitive aspects of the lecture. Accuracy Factor 3 includes the 3 items pertaining to both personal and cognitive aspects of the lecture and lecturer.

demonstrated that there were 19 condition differences across the 21 factorial accuracy indices. The condition means for each accuracy index are presented in Table 16.

There were significant condition differences between Condition 1 and all other conditions for DE1 and DE3. It can be seen that Condition 1 was more accurate than the other conditions for DE1 which refers to raters' ability to accurately discern differences in overall behavior between ratees that relate to personal aspect of the lecturer. Condition 2 was the least accurate. However, when the accuracy index measures raters' ability to discern differences between ratees' overall behavior that relate to a combination of personal aspects of the lecturer and cognitive aspects of the lecture (DE3), Condition 1 was the least accurate. In this case Condition 4, where raters did not write out behavioral rationales, was the most accurate. It may be that the writing of behavioral rationales was the component that led to the most confusion in distinctions between ratees' overall behavior when dealing with two different aspects of performance. However, omitting other aspects of the training also led to greater accuracy than the complete training program. When including all the components and then requiring ratings on two aspects of performance the information presented may have been too overwhelming. The accuracy for discerning differences in overall performance, related only to cognitive aspects of the lecture (DE2), was significantly different between Conditions 1

and 3. Here Condition 1 was again more accurate than Condition 3. In Condition 3 raters did not receive feedback of true scores and behavioral rationales, or participate in any discussion. In this case a possible explanation may be that the cognitive aspects of performance were more difficult to detect without feedback and discussion of the experts' ratings. Therefore, when it came to distinctions in ratees' cognitive performances without these components accuracy was decreased.

There were significant planned contrast effects between Condition 1 and Conditions 2, 4, and 5 for SA3. For this accuracy index Condition 1 was the least accurate and Condition 4 was, again, the most accurate. It appears that accuracy of individual dimension ratings pertaining to both personal and cognitive aspects of the lecturer's behavior and lecture is diminished by including all FOR training components and enhanced most by not writing out behavioral rationales for each dimension. It seems that when rating two different aspects of performance trainees are given too much information in Condition 1 and omitting the behavioral rationales leads to the most clarity.

There were significant planned contrast effects between Conditions 1 and 3 for DA1. By looking at the means in Table 16 it can be seen that trainees in Condition 1 were more accurate in detecting patterns of performance differences between ratees in terms of behavioral distinctions (DA1) than were those in Condition 3 ($\bar{x}$= .48 vs. $\bar{x}$= .58, respectively). There were

significant planned contrast effects between Condition 1 and Conditions 3 and 5 for DISTA3. Trainees in Condition 3 and 5 were closer to the expert true scores on the scale dimensions that were a combination of personal and cognitive aspects of the lecturer and lecture than those in Condition 1. There was a significant planned contrast effect between Conditions 1 and 3 for ABVH1. Again, trainees in Condition 1 were closer to the experts' ratings for personal aspects of the lecturer in terms of absolute halo than those in Condition 3. Finally, there was a significant contrast effect between Conditions 1 and 5 for LENA2. However, in this case trainees in Condition 5 were closer to the experts' ratings of true severity than trainees in Condition 1. In Condition 5 trainees did not receive any discussion of the rating scale dimensions prior to training. This seems to have contributed to less severity in ratings of cognitive aspects of ratees' performances and profile ratings that were closer to the experts' profile ratings.

Error

Error analyses were conducted separately for each topic except where noted. The results of the analyses for Variance Halo are presented in Table 17. The aggregate scale measure of the halo estimates for the CS lecture demonstrated a training condition main effect ($F(5,244) = 2.4348$, $p < .04$). There were no training condition differences for the SFP lecture or for the combination of the two topics on the aggregate scale measures.

Table 17

F Values for Univariate Analyses of Variance Halo Estimates

Topic	F (5,244)
CS	
aggregate	2.44*
Factor 1	2.88*
Factor 2	.90
Factor 3	.88
SFP	
aggregate	1.34
Factor 1	.98
Factor 2	1.93
Factor 3	2.76*
CS/SFP COMBINED	
aggregate	1.79
Factor 1	2.95**
Factor 2	1.36
Factor 3	.94

Note. Factor 1 includes the six items pertaining to personal aspects of the lecturer. Factor 2 includes the 5 items pertaining to cognitive aspects of the lecture. Factor 3 includes the 3 items pertaining to both personal and cognitive aspects of the lecture and lecturer.

* $p<.05$. ** $p<.01$.

The results for the three factor measures demonstrated a different pattern of results. The factor scale items encompassing personal aspects of the lecturer (Factor 1) resulted in significant training condition main effects for the CS lecture ($F(5,244) = 2.88$, $p < .02$) and for the combined topics ($F(5,244) = 2.95$, $p < .01$). The three scale items that contained elements of both personal and cognitive aspects of the lecturer and lecture (Factor 3) demonstrated a significant training condition main effect for the SFP lecture. Tukey post hoc analyses were performed to investigate the direction of the training condition differences. There were significant differences between Conditions 1 and 3 ($p < .05$), for the CS aggregate (σ^2=.754 vs. σ^2=1.19, respectively) and Factor 1 (σ^2=.754 vs. σ^2=1.50, respectively) indices. These condition differences were also evident for Factor 1 within the combined topics (Condition 1 σ^2=.956 vs. Condition 3 σ^2=1.71). However, Tukey post hoc analyses for condition differences for Factor 3 of the SFP lecture failed to demonstrate any significant differences across pairwise comparisons.

The differences between Conditions 1 and 3 were not in the predicted direction. Condition 3 evidenced less Variance Halo than Condition 1. Those trainees who did not receive feedback of expert true scores and behavioral rationales, and discussion committed less halo error, as evidenced by greater variance estimates, than those who did receive these components. This pattern seems even clearer for the personal aspects of the

lecturer's performance where differences between these conditions appeared twice. These results imply that the inclusion of these elements of training lead to more halo, as commonly defined, and ratings evidence less halo when omitted even though the goal of these components is to provide the basis for the frame-of-reference. It may be that these elements lead to a general impression effect, which is even more pronounced for personal dimensions of behavior. Additionally, this is in direct contrast to the accuracy results which indicated that Condition 1 was more accurate than Condition 3 on personal aspects of ABVH. ABVH measures the degree of profile similarity for halo between the experts and trainees. While those in Condition 1 were more accurate that those in Condition 3 in this respect, they had more Variance Halo and thus more error in this operationalization. These results directly support the belief in the literature that less error does not necessarily lead to greater accuracy.

For the second operationalization of halo all six conditions demonstrated a significant Ratee x Dimension interaction indicating a lack of this type of halo for each condition. Given that this interaction can be considered an indication of discriminant validity (Kavanagh et al., 1971) the results were further explored in order to compare differences between conditions. Following the procedure reported in Kavanagh et al. (1971) variance components were first estimated for the Ratee x Dimension effects. These were then translated into variance

indices to control for differences in sample sizes between conditions. These indices were then directly comparable between conditions. The univariate F values, variance components and variance indices are reported in Table 18. These indices refer to raters discrimination among ratees in terms of the ordering of ratees differently on different traits. In this manner a larger variance index would indicate greater discriminant validity since there would be greater discrimination between ratees and between dimensions for each ratee. It can be seen that Condition 5, where trainees did not receive any discussion of the rating scale prior to training, demonstrated the highest variance index and thus the greatest discrimination whereas Condition 1 demonstrated the least discrimination. This is interesting since Condition 1 also demonstrated, in general, the highest degree of accuracy on indice measures of factor 1 (e.g., DE1, DA1, and ABVH1). Again, there is a contradiction in error and accuracy as operationalized.

The results for Interrater Reliability demonstrated only three significant differences, using the Fmax statistic, between conditions across the 30 behavioral dimensions (15 for each topic). These occured on the CS lecture for dimension 5 ($F(6,34) = 2.94$, $p < .05$), dimension 8 ($F(6,39) = 2.72$, $p < .05$), and dimension 10 ($F(6,48) = 2.73$, $p < .05$). These dimensions addressed whether the lecturer followed an outline, provided relevant answers to questions, and whether he was relaxed. Thus,

Table 18

F Values, Variance Components, and Variance Indices for Ratee x Dimension Halo

Condition	n	F (df)	Variance Component	Variance Index
1	38	5.21** (14,518)	.073	.100
2	35	5.92** (14,476)	.086	.123
3	48	8.42** (14,658)	.102	.134
4	40	8.85** (14,546)	.104	.164
5	40	12.12** (14,546)	.176	.218
6	49	6.71** (14,672)	.068	.104
2,3	83	181.00** (14,1148)	.816	.687
2,3,4	123	389.02** (14,1701)	.512	.782
5,6	89	233.01** (14,1228)	.308	.716

** $p<.01$.

in general, it can be concluded that there was very similar interrater reliability across all training conditions. The results of this analysis is presented in Table 19.

The results for Midpoint Leniency failed to support Hypothesis 1. There was no condition main effect and therefore all training conditions demonstrated the same degree of leniency. These results coincide with the results for LENA which failed to detect significant differences between Condition 1 and the other conditions, with the exception of LENA2 between Conditions 1 and 5.

HYPOTHESIS 2

Hypothesis 2 stated that trainees who received more active participation in FOR training would have greater accuracy and less error in rating behavior than trainees whose training was less active. To test this hypothesis conditions were collapsed over high and low activity levels. Thus, Conditions 5 and 6 were considered more active since they included writing of behavioral rationales, feedback of true scores and behavioral rationales, and discussion. Conditions 2, 3, and 4 were considered less active since they were missing one or more of these components. Condition 1 was removed from the analyses since it contained both active and passive components.

Accuracy

Two separate MANOVA's (e.g., one MANOVA for the Cronbach indices and one MANOVA for the remaining three indices) were

Table 19

Fmax Statistic for Interrater Reliability Between Conditions

	Condition Comparisons		
Dimension	All (df)	2,3,4 vs 5,6 (df)	2 & 3 vs 4 (df)
		CS	
1	1.57 (6,34)	1.06 (2,88)	1.15 (2,82)
2	1.74 (6,39)	1.26*(2,88)	1.36*(2,82)
3	1.32 (6,34)	1.09 (2,88)	1.12 (2,82)
4	.83 (6,48)	1.16 (2,88)	1.55*(2,82)
5	2.94*(6,34)	1.22*(2,122)	1.41*(2,82)
6	1.75 (6,34)	1.09 (2,122)	1.16 (2,82)
7	1.73 (6,47)	1.06 (2,122)	1.53*(2,82)
8	2.72*(6,39)	1.34*(2,88)	1.63*(2,82)
9	2.30 (6,39)	1.44*(2,88)	1.29*(2,82)
10	2.73*(6,48)	1.33*(2,88)	1.14 (2,82)
11	1.98 (6,47)	1.01 (2,122)	1.56*(2,82)
12	1.55 (6,39)	1.13 (2,88)	1.23 (2,82)
13	2.54 (6,39)	1.04 (2,88)	1.34*(2,82)
14	2.41 (6,39)	1.20 (2,88)	1.61 (2,82)
15	1.61 (6,34)	1.08 (2,88)	1.59 (2,82)
		SFP	
1	2.37 (6,39)	1.16*(2,122)	1.40 (2,39)
2	1.83 (6,48)	1.20 (2,88)	1.60*(2,82)
3	1.57 (6,39)	1.36*(2,88)	1.14 (2,82)
4	2.27 (6,48)	1.01*(2,122)	1.33 (2,39)
5	1.93 (6,37)	1.01*(2,122)	1.44*(2,82)
6	1.84 (6,37)	1.08 (2,88)	1.36*(2,82)
7	1.78 (6,47)	1.01*(2,122)	1.47*(2,82)
8	2.03 (6,37)	1.03 (2,88)	1.11 (2,39)
9	2.25 (6,37)	1.02 (2,88)	1.42*(2,82)
10	2.36 (6,39)	1.44*(2,88)	1.63*(2,82)
11	1.74 (6,48)	1.44*(2,88)	1.03 (2,39)
12	1.83 (6,37)	1.17 (2,88)	1.40*(2,82)
13	1.41 (6,37)	1.12 (2,88)	1.30*(2,82)
14	1.29 (6,39)	1.11 (2,88)	1.12 (2,39)
15	2.55 (6,37)	1.23 (2,88)	1.01 (2,82)

*$p < .05$.

performed for the aggregate accuracy indices with two condition levels (Conditions 2, 3, 4 = 1 and Conditions 5 and 6 = 2) and two other MANOVA's (separated as above) were performed for the trichotomized indices with the same two condition levels. Therefore, a total of four MANOVA's were performed. However, there were no significant condition effects for any of the analyses. No support was found within accuracy indices for Hypothesis 2.

ERROR

The Halo Variance estimates were combined for Conditions 2, 3, and 4 and for Condition 5 and 6. Separate Student t's were performed on the variance estimates for the CS lecture ratings, the SFP lecture ratings, and both sets of ratings combined to test the difference between the two condition groupings. Student t's were used here as opposed to the ANOVA method employed in Hypothesis 1 since there were now only two conditions. These analyses were performed for the aggregate scale measure and the factor scale measures. There were no significant halo differences between the more or less active training conditions for the aggregate or factor scale measures. Therefore, there was no support for Hypotheses 2 in relation to a decrease in Variance Halo in the more active conditions.

Ratee x Dimension Halo was investigated through performing separate Rater x Ratee x Dimension repeated measure MANOVA's. For this hypothesis Conditions 2, 3, and 4 were combined for one

MANOVA and the results were compared to another MANOVA where Conditions 5 and 6 were combined. Both MANOVA's indicated significant Ratee x Dimension interactions. Thus, there was no difference in this type of halo between the active and passive training conditions. These indices were then explored as indications of discriminant validity. By referring to Table 18 it can be seen that the more passive training conditions had a larger variance index than the more active conditions. The more passive conditions, as a whole, did not receive true score feedback, behavioral rationale feedback, did not write out behavioral rationales, nor participate in discussion. The active conditions included these components. These components are proposed to facilitate greater accuracy and are intended to help in developing a frame-of-reference with which to evaluate behavior. However, as evidenced here they demonstrated a decrease in discriminant validity, whereas removing some or all of these components demonstrated greater discrimination between ratees. This is in direct contrast to Hypothesis 2.

The results for Interrater Reliability demonstrated significant differences between the combined passive conditions versus the combined more active conditions for four dimensions on the CS lecture and seven dimensions on the SFP lecture. Table 19 indicates the specific pattern of results for each dimension within the two topics. These results suggest that there are differences between the two conditions. By reviewing the

variance estimates for the two groups it appears that the more active conditions had greater variability and less interrater reliability than the more passive conditions. This would suggest that feedback of true scores and behavioral rationales, participant discussion, and writing out the behavioral rationales leads to less reliability across ratees than when these components are omitted. These results are inconsistent with those for the other error and accuracy indices. It seems illogical that these components should detract from interrater reliability since they should serve to facilitate a deeper and more consistent level of cognitive processing (Athey & McIntyre, 1987). An alternative and more plausable explanation is that these results are an artifact of the analysis. As discussed above the Fmax statistic is sensitive to differences in sample sizes and there were large differences here (n's = 123 and 89). This sensitivity leads to the rejection of homogeneity of variance more often. It is reasonable to conclude that this false rejection occured here. Yet, these resuls as discussed failed to support Hypothesis 2, and demonstrate the inverse relationship.

Training conditions 2, 3, an 4 were combined as were Conditions 5 and 6 for investigating Midpoint Leniency. Thus, there were now two levels of the condition factor. In this MANOVA there was again no condition main effect and there was a significant ratee main effect ($F(1,210) = 20.94$, $p < .001$). These

results failed to support Hypothesis 2. There were no differences in Midpoint Leniency between the more and less active trainee participation conditions.

HYPOTHESIS 3

Hypothesis 3 stated that in general there was a hierarchical importance to the less active training conditions (2, 3, and 4). It was proposed that there would be greater accuracy and less error in the training condition that included feedback of true scores and behavioral rationales (Condition 4) than in those that did not include these components (Conditions 2 and 3).

Accuracy

The identical MANOVA procedures used to test Hypothesis 2 were employed here, except that scores in Conditions 2 and 3 were collapsed over as one level of the condition factor and Condition 4 was entered as the second level of this factor. Therefore, again, four separate MANOVA's were conducted. There were no significant condition effects for any of the analyses. No support was found for condition differences in terms of the accuracy indices and no support was found for Hypothesis 3.

Error

For the analysis of Variance Halo estimates were combined for Conditions 2 and 3. Separate Student t's were again performed on the variance estimations for the CS lecture ratings, the SFP lecture ratings, and both sets of ratings combined to

test the differences between Conditions 2 and 3 versus Condition 4. These analyses were performed for the aggregate scale measures and the factor scale measures. There were no significant differences between the conditions for any of the tests using the aggregate scale. The results for the factor measures were more illuminating. In this case there were differences in halo estimates between the two groups for the factor comprising the personal aspects of the lecturer for the CS lecture ($t(121) = 2.36$, $p < .02$) and the two lecture topics combined ($t(121) = 1.91$, $p < .05$). In both cases the more active training condition which included feedback of true scores and behavioral rationales, and discussion demonstrated more halo than the less active training conditions that did not include these components. However, trainees in the more active condition (Condition 4) also did not write out behavioral rationales. It is unclear whether it was the inclusion of these components, or the exclusion of writing the behavioral rationales that led to greater halo. This is in direct contrast to Hypothesis 3. It appears that in this case, the less active training led to a decrease in halo whereas the more active training led to greater halo. Therefore, there was no support for Hypothesis 3 in relation to the hierarchical importance of the less active training conditions for Variance Halo.

Ratee x Dimension Halo was also investigated through performing separate Rater x Ratee x Dimension repeated measure MANOVA's. For the analysis of Ratee x Dimension Halo Condition 2

and 3 were combined for one MANOVA and the results were compared to the same MANOVA design employing Condition 4. There was a significant Ratee x Dimension effect for the combined conditions ($F(14,1148)$, = 181.00, $p < .05$). It will be recalled that there was also a significant interaction for Condition 4. Thus there was no difference in this type of halo between the less active training conditions. These interactions were then explored as measures of discriminant validity to ascertain whether there were differences between the two groups. By looking at Table 18 it can be seen that the combined conditions had much more discriminant validity as evidenced by the variance index than did Condition 4. In Condition 4 trainees were not required to write out the behavioral rationales for each rating dimension and this could have contributed to this lower level of discriminant validity. In Conditions 2 and 3 some trainees did not receive the experts' behavioral rationales and others did not receive true scores or behavioral rationales. While these steps are considered integral for leading to greater accuracy their omission also seems to have contributed to greater discriminant validity. These results are in direct contradiction to Hypothesis 3. It appears that of the less active conditions, Conditions 2 and 3 combined led to greater discriminant validity, operationationalized as a ratee by dimension effect, than Condition 4.

The results for Interrater Reliability demonstrated significant differences between the combined conditions 2 and 3 versus Condition 4 for 10 dimensions on each lecture. Table 19 reports the results of this analysis. By reviewing the variance estimates between the two groups it appears that in this situation the more active condition demonstrated greater interrater reliability than the passive conditons. Not writing out behavioral rationales seems to have fared better on this index whereas, not receiving feedback of true scores and behavioral rationales seems to have fared worse. This is in direct contrast to the results of this index in Hypothesis 2.

Due to these inconsistencies, as well as the bias of the Fmax statistic for unequal sample sizes, these results seem untenable. It appears that this index should not be pursued for further exploration, as a false rejection of the homogeneity test seems very likely.

Midpoint Leniency was assessed as described above, however, Conditions 2 and 3 were collapsed over and compared to Condition 4. Thus there were two levels to the condition factor in the MANOVA. As in the previous two hypotheses, there was no condition main effect and there was a significant ratee main effect. Thus, these results failed to support Hypothesis 3.

HYPOTHESIS 4

Hypothesis 4 stated that trainees who had greater memory for what was observed during the lecture vignettes would achieve greater accuracy on the behaviorally based performance appraisal Instructor Evaluation Scale. For this hypothesis, the Behavior Frequency Scale was conceptualized as a measure of memory or knowledge for what was observed during the lecture vignettes. To test this hypothesis the same seven accuracy indices were calculated for scores on the Behavior Frequency Scale (true scores were collected on this scale also) and each index was trichotomized into high, medium, and low memory accuracy scores based on cummulative percentages of the trainees' accuracy values. These seven indices were in this manner operationalized as memory accuracy scores (e.g., MDA, MSA, MDE, MEL, MDISTA, MLENA, MABVH). Therefore, the high memory condition for an accuracy index consisted of approximately 33% of the top scorers, the medium memory condition consisted of approximately 33% of the next highest scorers and so forth. Two separate MANOVA's were conducted (e.g., one MANOVA was for the four Cronbach measures and the other was for the remaining three indices) for each memory accuracy index. Again, the memory indices were developed through ratings on the Behavior Frequency Scale and the evaluation accuracy indices were developed through the Instructor Evaluation Scale. Therefore, 14 separate MANOVA's were employed to test this hypothesis. The dependent measures were the seven

Instructor Evaluation accuracy scores and the independent factors were the seven Behavior Frequency memory accuracy indices. For each memory factor there were three levels - high, medium, and low.

Table 20 provides the results of the MANOVAs. The only accuracy memory score not to demonstrate a significant main effect in either MANOVA was MDE. This is interesting since DE exhibited the most consistent trend for differences in training conditions. All other memory accuracy indices demonstrated significant effects in one or both of the MANOVAs. Table 21 presents the results of the univariate tests. It can be seen that the SA memory score achieved significant results with six of the seven Instructor Evaluation accuracy scores. This is also interesting since SA did not reveal any significant differences across the training conditions. Thus, it appears that a certain type of accuracy, in this case for dimension ratings across all ratees is uneffected by training differences, but as a memory index does influence accuracy ratings, whereas another type of accuracy, related to distinctions in overall performance (DE), is very effected by training differences, but as a memory index does not effect accuracy apart from training.

Where significant multivariate main effects were found Scheffe post hoc analyses were performed to investigate further the pattern of differences among the memory accuracy indices and Instructor Evaluation accuracy. Scheffe tests were performed comparing the means on the evaluation accuracy scores against

Table 20

Multivariate Analysis of Variance Results of Memory on Accuracy

Memory Accuracy Index	Wilk's Lambda	F
	MANOVA 1[a]	
MDA	.85969	4.79**
MSA	.54576	21.57**
MDE	.95687	1.36
MEL	.94975	1.59
MDISTA	.66277	13.93**
MLENA	.95794	1.32
MABVH	.64475	14.97**
	MANOVA 2[b]	
MDA	.96123	1.63
MSA	.54039	29.43**
MDE	.98755	.51
MEL	.94947	2.15**
MDISTA	.64227	20.24**
MLENA	.94752	2.23*
MABVH	.62768	21.41**

[a] MANOVA 1 contained the accuracy indices of DA, SA, DE, and EL.

[b] MANOVA 2 contained the accuracy indices of DISTA, LENA, and ABVH.

* $p<.05$. ** $p<.01$.

Table 21

F Values[a] of Univariate Analyses of Variance Results for Memory Accuracy

	Memory Indices						
Accuracy	MDA	MSA	MDE	MEL	DISTA	LENA	ABVH
DA	14.12**	3.70*	2.78	1.15	5.22**	.60	2.41
SA	.79	.54	1.25	1.03	.75	.07	.24
DE	8.48**	7.37**	1.53	1.08	4.37**	.32	.55
EL	3.37*	99.14**	1.36	3.97*	53.55**	3.49	53.72**
DISTA	.94	95.97**	.49	3.68*	64.45**	5.59**	69.34**
LENA	2.67	85.61**	1.92	4.67**	47.28**	3.20*	48.92**
ABVH	.08	3.42*	.10	.78	2.71	1.50	2.61

[a] F Value df=2,247.

* p<.05. ** p<.01.

each level of the memory accuracy scores. Table 22 presents the mean Instructor Evaluation accuracy indices for each level of the memory accuracy indices. It can be seen that there were significant differences within six of the evaluation accuracy indices for different levels of MSA. The only Instructor Evaluation accuracy index not to demonstrate significant differences between different levels of stereotype memory accuracy was SA. The highest stereotype memory accuracy scores also demonstrated the highest Instructor Evaluation accuracy scores for EL, DISTA, LENA, and ABVH. However, this finding was reversed for DA and DE. In these cases the lowest DA and DE Instructor Evaluation accuracy scores were those that achieved the highest stereotype memory accuracy scores. It appears that heightened memory for behavior at the dimension level across ratees increases accuracy across raters, yet this type of stereotyping decreases accuracy when distinctions between ratees are necessary.

There were also significant differences within three of the Instructor Evaluation accuracy indices for different levels of MDA. These occured for DA, DE, and EL. For these accuracy indices there were significant differences between high and low levels of DA memory accuracy and Instructor Evaluation accuracy indices, however, the high MDA group had the greatest accuracy scores for DE and DA but also had the lowest accuracy scores for EL. In this instance it can be seen that heightened memory for

Table 22

Scheffe Post Hoc Mean Differences in Accuracy for High, Medium, and Low Memory Accuracy

	Memory Accuracy			
Accuracy	High	Medium	Low	Significant Group Differences[a]
			MDA	
	n=85	n=85	n=80	
DA	.65	.71	.77	1 vs 2, 1 vs 3, 2 vs 3
DE	.38	.47	.61	1 vs 3
EL	1.73	1.63	1.44	1 vs 3
			MSA	
	n=84	n=82	n=84	
DA	.74	.71	.68	1 vs 3
DE	.56	.54	.36	1 vs 3, 2 vs 3
EL	1.04	1.55	2.22	1 vs 2, 1 vs 3, 2 vs 3
DISTA	1.52	1.82	2.30	1 vs 2, 1 vs 3, 2 vs 3
LENA	-1.02	-1.51	-2.22	1 vs 2, 1 vs 3, 2 vs 3
ABVH	.82	.86	.94	1 vs 3
			MEL	
	n=82	n=85	n=83	
LENA	-1.64	-1.72	-1.38	2 vs 3

Table 22 continued

Accuracy	Memory Accuracy: High	Medium	Low	Significant Group Differences[a]
		MDISTA		
	n=84	n=81	n=85	
DA	.70	.75	.67	2 vs 3
DE	.51	.55	.39	2 vs 3
EL	1.16	1.52	2.12	1 vs 2, 1 vs 3, 2 vs 3
DISTA	1.55	1.84	2.25	1 vs 2. 1 vs 3, 2 vs 3
LENA	-1.15	-1.47	-2.12	1 vs 2, 1 vs 3, 2 vs 3
		MLENA		
	n=80	n=88	n=82	
DISTA	1.75	2.00	1.89	1 vs 2
LENA	-1.43	-1.73	-1.57	1 vs 2
		MABVH		
	n=83	n=82	n=85	
EL	1.18	1.49	2.13	1 vs 2, 1 vs 3, 3 vs 2
DISTA	1.56	1.81	2.27	1 vs 2, 1 vs 3, 3 vs 2
LENA	-1.16	-1.44	-2.13	1 vs 2, 1 vs 3, 3 vs 2

Note. The lower the accuracy score the higher the degree of accuracy.

[a] Group differences were significant at $p < .05$ or less.

distinctions among ratees in patterns of performance increases accuracy for determining overall performance and differentiation in performance between ratees, yet decreases accuracy of the average rating across ratees.

Memory scores for DISTA affected differences in five Instructor Evaluation accuracy scores. For two of these, DA and DE, the lowest memory accuracy scores demonstrated the highest Instructor Evaluation accuracy scores, whereas for the remaining three (e.g., EL, DISTA, and LENA) increased DISTA memory accuracy increased Instructor Evaluation accuracy. Overall, it appears that DA and DE Instructor Evaluation accuracy scores are increased when particular types of memory, those that are collapsed over ratees, are decreased.

High memory scores for EL demonstrated significant differences with only one evaluation accuracy score. Those with high memory accuracy on EL also demonstrated significantly lower accuracy scores on LENA than those with the lowest memory accuracy scores. In general for the other evaluation accuracy indices those raters who demonstrated a high level of memory accuracy also had the highest Instructor Evaluation accuracy scores. Along with SA only ABVH evaluation accuracy scores were not affected by different levels of memory accuracy on any of the indices. These results represent partial support for Hypothesis 4, yet it is clear that different types of memory affect accuracy in systematic ways.

Intercorrelations Among Accuracy Indices

One of the exploratory elements of this research project was to assess the degree of intercorrelation among the seven accuracy indices, regardless of training condition, to determine whether each measure provides unique information. Table 23 reports the results of the accuracy intercorrelations employing the aggregate accuracy indices (e.g., all 15 behavioral dimensions were included). It shows 16 of a possible 21 (or 76.2%) of the correlations to be statistically significant. SA appears to have provided the most distinct and unique information with only two of a possible six significant correlations. Additionally, there was not a significant correlation between DE and ABVH suggesting that when compared to one another they too provide unique information. In general, these results suggest that there is a great deal of overlap in the information offered by these indices. The correlations between DISTA and DE and LENA were very strong ($\bar{r}$ = .92, $p < .001$) and the correlation between EL and LENA was perfect (r= 1.0, $p < .001$), indicating almost complete, or complete, overlap between these measures as operationalized. Seven of the significant correlations were negative. Of the five correlations with DA (DE, EL, DISTA, LENA, and ABVH) all but DE demonstrated a negative relationship. This suggests that the ability to accurately distinguish among ratees on each performance dimension is not related to the ability to

Table 23

Intercorrelations Among Accuracy Indices

Accuracy	DA	SA	DE	EL	DISTA	LENA[a]	ABVH
DA	1.0	.09	.25**	-.30**	-.16**	-.30**	-.18**
SA	-	1.0	-.05	.06	.22**	.06	.14*
DE	-	-	1.0	-.35**	-.14*	-.35**	.00
EL	-	-	-	1.0	.92**	1.00**	.33**
DISTA	-	-	-	-	1.0	.92**	.37**
LENA	-	-	-	-	-	1.0	.33**

[a] LENA was converted to an absolute value for this analysis.

* p<.05. ** p<.01.

judge other aspects of rating accuracy. Negative correlations between DE and EL, DISTA, and LENA also occured. This demonstrates that the ability to accurately distinguish differences among ratees in overall performance is not related to these other aspects of rating accuracy and the opposite is most likely to be true. Raters who can distinguish between ratees in overall performance are not good judges of the other aspects of accuracy. However, in general, the correlations were of moderate magnitude except where noted.

Individual Differences

There were two different sets of individual differences explored in this research. One type of individual difference included typical demographic variables including sex, age, race, college major, year in school, enrollment status, language used most often, and employment status. Additional demographic variables were included that pertained explicitly to factors that might influence performance appraisal ratings as they related to the lecture topics. These included questions concerning living arrangements that might be particularly salient to subsequent attitudes toward the CS lecture. A list of these variables are included in Appendix C. Another group of variables addressed raters' familiarity and experience with conducting and receiving performance appraisals.

A second set of individual difference variables addressed raters' attitudes toward the lecture and lecturer. There were six attitudinal questions in all. These are presented in Appendix D. It can be seen that four questions pertained explicitly to the lecture and lecture topic while the remaining two addressed questions about the lecturer.

Investigation of the influence of the individual difference variables were explored in several ways. The questions of interest here were: 1) whether the attitudes were related to accuracy; 2) whether the attitudes were related to individual dimension ratings; 3) whether accuracy differed by demographics; 4) whether the individual dimension ratings differed by demographics. Furthermore, in order to further explain influences on condition differences two additional questions were addressed. These were: 5) whether there were differences in attitudes by conditions; and, 6) whether there were differences in demographics by conditions. Each question will be addressed separately.

Demographic Differences among Conditions

To investigate differences in demographics across the six training conditions Chi-Square tests were performed. Of the 16 tests performed (age was excluded) nine demonstrated significant differences between training conditions. Table 24 presents the results of the significant Chi-squares. Though these differences

existed, subsequent analyses (to be reported below) did not demonstrate their effect on accuracy or dimension ratings. Additionally, it can be seen in Table 24 that the strength of the relationship, as indicated by Cramer's V was relatively low, ranging from .190 to .255.

Relationship between Demographic Differences and Accuracy Indices

The demographic differences in accuracy were investigated through two separate MANOVA's for each demographic variable with the DA, SA, DE, and EL accuracy indices entered as dependent measures on one MANOVA and DISTA, LENA, and ABVH entered as dependent measures on the other MANOVA. Across all the demographics the only variable to demonstrate a significant main effect was the rater's most frequently spoken language. For the first MANOVA the Wilks' lambda was .9067, $F = 2.0044$, $p < .022$ and .9108, $F = 1.91864$, $p < .030$ for the second. The univariate analyses revealed significant effects for DA ($F(3,244)= 3.8572$, $p < .01$) and ABVH ($F(3,244)= 3.1176$, $p < .016$). While the effect of language on accuracy is not unusual, what is unusual is that for DA the least accurate group was those who spoke English most often and the most accurate was those who spoke Spanish ($\bar{x}= .7205$ vs $\bar{x}= .5907$, respectively). For ABVH, those who spoke English were most accurate and those who spoke Chinese were the least accurate ($\bar{x}= .85$ vs $\bar{x}= 1.13$, respectively).

Table 24

Chi-Square Results of Significant Demographic Differences Between Training Conditions

Demographic	df	X^2	Cramer's V
Enrollment status	25	50.23**	.201
Year in school	30	81.08**	.255
School major	30	74.94**	.245
Living situation	30	50.24**	.201
Live In	30	61.60**	.222
Employment status	10	24.82**	.223
Had performance reviewed	10	18.06**	.190
Formal CS knowledge	10	18.28*	.191
Formal SFP knowledge	10	29.03**	.238

* $p<.05$. ** $p<.01$.

Since there were no other significant effects it can be concluded that, in general, demographic variables did not influence these results. This is surprising since included in these variables were questions about experience with performance appraisals which indicate that those raters who are more experienced were no more accurate than those with less experience. However, a positive finding was that for those who may have been more sensitive to aspects of the CS lecture, this sensitivity did not affect their accuracy ratings any differently than those who were not sensitive to this topic.

Relationship between Demographics and Performance Dimension Ratings

To investigate differences in demographics on dimension ratings a 2 x 15 (Topic x Dimension) repeated measures MANOVA was performed for each of the 16 demographic variables (age was omitted). The only demographic variable to demonstrate significant multivariate results for a dimension by demographic interaction was sex (.8942, $F= 1.9871$, $p < .02$).

There were no significant main effects for any of the demographic variables. The results were interesting for topic effects. Eleven of the MANOVA's demonstrated significant topic main effects, indicating a difference in topic ratings. This is consistent with the topic effects found for the analyses of the measurement situation. However, for five of the analyses (race, major, language, crowded, living situation) there were no topic

differences. Additionally, there were no significant topic by demographic interactions.

Condition Differences in Rater Attitudes

To investigate training condition differences in rater attitudes a MANOVA was performed with the raters' attitudes for each lecture entered as dependent variables and the six different condition levels as the independent factor. The MANOVA failed to yield a significant condition main effect regarding attitudes. Therefore, there were no condition differences in rater attitudes.

Relationship Between Rater Attitudes and Accuracy

To investigate the relationship between rater attitudes and rating accuracy, Pearson Product-Moment correlations were employed. Table 25 presents the correlations between the seven accuracy scores and the rater attitudes for both the CS and SFP lectures. There was no relationship for either topic between rating accuracy and whether raters would like to learn more about the lecture topic. Across the two lecture topics the relationship between the accuracy indices and the rater attitudes were parallel for SA, EL, DISTA, and LENA. There was no relationship between SA and any of the rater attitudes. For EL, DISTA, and LENA, of the significant correlations four were positive and the fifth was negative. It appears that when attitudes became more positive regarding the lecture, the lecturer accuracy also increased. Conversely, the more

Table 25

Intercorrelations Among Rater Attitudes and Accuracy[a]

	Attitude Scale Items					
Accuracy	1	2	3	4	5	6
			CS			
DA	.25**	-.07	.30**	.30**	-.02	.26**
SA	-.11	-.06	.02	-.04	-.04	-.03
DE	.47**	-.25**	.37**	.50**	.05	.49**
EL	-.56**	.30**	-.49**	-.53**	.02	-.56**
DISTA	-.48**	.25**	-.38**	-.41**	.02	-.44**
LENA	-.56**	.30**	-.49*	-.53**	.02	-.56**
ABVH	-.10	-.07	-.12	-.10	.05	-.09
			SFP			
DA	.02	-.04	.05	.10	-.03	.08
SA	.00	-.09	.04	.09*	-.02	.10*
DE	-.16**	.07**	-.16**	-.13**	.03	-.14**
EL	-.34**	.16**	-.24**	-.35**	.02	-.35**
DISTA	-.36**	.18**	-.27**	-.35**	.02	-.35**
LENA	-.34	.16	-.24	-.35**	.02	-.35**
ABVH	.01	.04	-.12	-.17	.02	-.18

[a] Positive correlations indicate a negative relationship between accuracy and attitudes.

interesting the topic itself was, the less accurate the ratings became. It may be that as raters became more interested in the lecture topic they paid less attention to the lecturer. This negative relationship did not hold for DE. In this case interest in the CS lecture topic was positively related to the ability to make accurate judgements between ratees in overall performance. ABVH was correlated with only one attitude for the CS lecture and two different ones for the SFP lecture. Two variables pertaining to the lecture and two pertaining to the lecturer negatively correlated with DA such that as attitudes became more positive the ability to make distinctions among individuals on each performance dimension decreased. It may be that as attitudes become more positive a type of halo in ratings occurs. However, for the SFP lecture this was not true. There were no significant correlations between DA and attitudes concerning the SFP lecture.

Relationship Between Rater Attitudes and Dimension Ratings

For each lecture topic the relationship between rater attitudes and dimension ratings was investigated through the calculation of Pearson Product-Moment correlations. In general, the pattern of correlations was similar across both topics. For both topics there were nonsignificant correlations between raters' attitudes towards learning more about the topic and dimension ratings. This is considered a positive finding since it indicates

that raters' motivation to learn about either topic is separate from their Instructor Evaluation ratings. The only other nonsignificant correlations occured between interest in the SFP lecture topic and six dimension ratings (i.e., he seemed interested in the topic, made clear objectives, encouraged questions, provided relevant answers to questions, and he spoke with vigor and enthusiasm, he spoke clearly and distinctly). All remaining correlations between rater attitudes and performance appraisal ratings were significant and positive (75 out of 90 or 83.3% for the CS ratings and 70 out of 90 or 77.7% for the SFP ratings). These results indicate that, in general, there is a positive relationship between raters' attitudes and dimension ratings. However, the magnitude of the relationships differ across the two topics. For the CS lecture the significant correlations ranged from $\underline{r}$= .21 ($\underline{p}$ < .001) to $\underline{r}$= .70, ($\underline{p}$ < .001) and for the SFP lecture they ranged from $\underline{r}$= .13, ($\underline{p}$ < .05) to $\underline{r}$= .48, ($\underline{p}$ < .001).

CHAPTER V

DISCUSSION

This study was undertaken in order to continue bridging the gap between research and application of performance appraisal training. There were several goals to this project that meld interests of the researcher and practitioner. These included: 1) investigating methodological issues inherent in this body of research; and, 2) systematically manipulating the components of FOR training in order to distinguish the effects of different methods and processes of performance appraisal training and to identify the most efficient combination of both. The results provide further understanding for both research topics. However, the results also indicate that there are many underlying complexities to the rating process, research methodology, and effectiveness indices that operate to disallow any simple interpretations between these factors. These complexities affect the whole. This research attempts to differentiate these influences and offer interpretations to provide greater understanding of important considerations for both those who conduct research in this area and those who implement performance appraisal training in applied settings.

This chapter provides a thorough discussion and interpretation of the results presented in Chapter IV. Each of the different aspects of the project are discussed separately and

suggestions for future research are offered within areas. Finally, limitations of this study are presented, as are a summary of the conclusions and contributions to research and application in this area.

The Measurement Situation

Many of the laboratory studies in performance appraisal training are conducted within the design framework utilized here (trainees who rate one or more videotaped vignettes of instructors, managers, or workers demonstrating varying levels of behavior). Traditionally, certain elements of the design (e.g., the training program, vignettes observed, rating purpose, rating scale format, etc.) are manipulated in order to determine their influence on error and accuracy indices of rating effectiveness. While this type of manipulation was employed here, investigation of the design elements was undertaken to determine whether the basic training evaluation research framework was inadvertently contributing to, or perhaps confounding, the identification of training effectiveness. Thus, the effect of vignette topics, and order of pretest topic/level of performance, and posttest order were investigated. These included analysis of: 1) whether there were differences in posttest ratings attributable to topics; 2) whether posttest topic ratings varied with the pairing of lecture topic/level of performance viewed at pretest; 3) whether there were differences in dimension ratings between the two topics; 4)

whether posttest topic order influenced posttest ratings; and, 5) whether vignette order/level of performance seen at pretest affected posttest ratings. These analyses revealed that these elements of the design, apart from training, demonstrated different patterns of outcomes.

Topic and Order Effects

The results indicated that there was a systematic effect for differences between topics. While the two topics were intended to represent similar levels of performance the CS lecture was consistently rated higher than the SFP lecture. Additionally, the topic seen first at pretest, which as designed exhibited a higher level of performance, influenced ratings by dramatically elevating the posttest ratings of the same topic. Therefore, a context effect was demonstrated which appears to have developed through a halo impression regarding the topic. Murphy, Balzer, Lockhart, and Eisenman (1985) and Smithers, Reilly, and Buda (1988) suggested that a contrast effect would result such that when trainees viewed the greater performance in either topic the subsequent posttest ratings of the identical topic would appear much lower given the contrast between the superior and average performance on the same topic. Yet, the opposite occured here. It may have been that the topic was more salient a feature of the presentation than the differences between performance levels. However, in their research the differences in ratings were across

one ratee on the same topic, whereas here the similarities in ratings were across one topic with different ratees.

The results also clearly demonstrated that the vignette order/level of performance presented at pretest affected posttest ratings. Trainees who saw the CS lecture as the superior performance example (e.g., pretest order A1) rated both posttest topics higher than did those who viewed the SFP lecture as the superior performance example (e.g., pretest order A2). This difference can be explained as evidence of both contrast and assimilation effects. While pretest topics were intended to demonstrate similar levels of performance within high and low categories, the experts had rated the superior SFP lecture used during training higher than the superior CS lecture ($\bar{x}$ = 43 vs. $\bar{x}$ = 52, respectively). These differences in superior performances seem to have affected posttest ratings by showing greater contrast between pretest order A2 and created an assimilation effect for pretest order A1.

In the case of the superior SFP training vignette (pretest order A2) there was a larger difference between the superior performance at pretest and the average performance at posttest, thus both posttest ratings were more moderate. For the superior CS training vignette (pretest order A1) the difference between pretest and posttest was not as great which seems to have caused the average performances of both topics to be elevated at posttest which may suggest an assimilation effect. The posttest performances

may have been viewed as more similar in pretest performance and thus the level of performance was assimilated at posttest.

There is literature which supports the conclusion of contrast effects and suggests that this may occur due to shifts in judgemental standards (Murphy et al., 1985). Additionally, as suggested by Feldman (1981), when incoming information is highly discrepant from previous information (in this case the level of performance observed) then the new information appears more salient and is encoded in more detail than consistent information. Subsequent evaluations may reflect this difference. The assimilation effect can occur when incoming information does not appear as extreme (Feldman, 1981; Smithers et al., 1988). However, these interpretations can only be offered as tenative due to the inherent confounding of topic order and level of performance at pretest. An alternative but less plausible explanation would be that there was some unknown influence in the CS high performance lecture (i.e., liking for the topic) which, when seen first, elevated posttest topic ratings by placing trainees in a better frame of mind or mood during the study.

These effects were unpredicted. Relative to one another, the videos were trichotomized by the experts during true score generation. Additionally, within and across performance levels the pattern of ratings were consistent. Superior performance in one topic was considered such, in comparison to average and low

performance in the same topic. Topic levels within conditions were also relative to one another.

There were no practice effects between the third and fourth rating occassions which occured after training. This was indicated by the lack of a significant posttest main effect. This is considered a positive finding. The two pretest practice sessions may have served to bring trainees up to a certain level of proficiency in rating abilities that did not then lead to carryover between the two final posttest ratings. Additionally, the insignificant results of posttest order ensured the efficacy of collapsing over posttest order in subsequent analyses.

The demonstration of rating differences between dimensions, and across topics within dimensions was also reassuring. This indicates that the trainees were rating each of the behavioral dimensions independently within vignettes and across topics. This suggests a degree of motivation or attention to each individual dimension as opposed to applying a "general impression" across or within topics. This also suggests that the trainees were, at some level, aware of the distinctions between behavioral dimensions which was an intent of the training. These results also partially support the original intent of the vignette's design. They were developed to demonstrate equal levels of overall behavior between vignettes, but varying levels of behavior within vignettes (Murphy et al., 1982). In this

study there were varying levels of behavior within vignettes and only somewhat similar levels across vignettes.

A limitation of the analysis of the measurement situation was the confounding of topic and level of performance presented at pretest in investigating order effects. In order to separate these two variables into distinct components it would have been necessary to counterbalance the order of presentation of high and low performance, separate from topics, such that trainees saw a low CS performance and a low SFP performance first at pretest, in addition to viewing these topics first in the high performance condition. While this confounding was intentional to reduce sample size requirements, the results here preclude the knowledge of whether pretest order, apart from level of lecture presentation, had an independent influence on the differences demonstrated.

The posttest ratings indicated differences between training conditions. This result was expected given that each training condition employed a different manipulation of the training content. However, differences between topics were also exhibited across the six training conditions. This suggests that the effects of the training conditions also varied with topic. As reported in Chapter IV, the effects of the training conditions were not consistent across topics as indicated by a different rank order of topic ratings between conditions. This then served as a warning to further investigate topic differences when

analyzing the results. Additionally, this interaction was explored further through the investigation of individual difference variables that might have contributed to rating differences and condition differences.

The benefits of this aspect of the research project are very direct. In all rigorous research, what cannot be controlled should be examined. This was the case herein. The different lecture topics were necessary to the research as were the presentation of differing levels of behavior. However, these design elements were counterbalanced as a method of control, and explored in an attempt at identification. Both proved fruitful. It is clear that the different topics were rated differently at posttest even though they were employed here as similar levels of behavior, and are utilized in other research without acknowledgement or investigation of how the different topics might influence the outcomes of the investigation. Also context and contrast effects were clearly present. The order of vignettes/levels of performance presented during training did influence subsequent ratings.

These elements of the research design need to be considered carefully in future research, given the potential for differential patterns of results based on simple design feature decisions. For example, those studies that employ these exact stimuli (i.e., Athey & McIntyre, 1987; McIntyre et al., 1984; Murphy et al.,1982a, 1982b) do not clearly indicate which tapes

were utilized in their research. Or, they do not always indicate the presentation order of levels of performance, which topics were used in which order, or they state that the vignettes were randomly selected. It may be the case that the occurence of a contrast or context effect systematically altered subsequent posttest ratings or that a random order of presentation biased differences between conditions apart from the intended manipulations. If the vignettes used during training differed significantly across conditions, for example, one vignette had more clearly delineated behaviors than another, then the effects of the independent variables may have been due to greater clarity of the behaviors observed in the videos rather than due to the superiority of one type of training.

Additionally, in the work of McIntyre and his associates, trainees viewed four vignettes, the first as practice and the remaining three as posttest measures. In all cases, the trainees observed the vignettes in the same order. In an effort to explain a lack of effect for a correlational accuracy index they discovered that the accuracy measure for one of the three vignettes was lower than for the other two. The results found here would suggest that this may have been due to a topic, contrast, or even a context effect. Unfortunately, they failed to identify the topics or levels of performance presented during or after training. As such, the ability to suggest or identify explanations for the differences are limited. Thus, this also

addresses the necessity for clearly describing these design features in published research and investigating the manifestation of these elements elsewhere in training results. This exploration can either support the adequacy and equality of stimulus materials and condition differences or can alert the researcher to search for its manifestation in subsequent results.

Individual Differences: Condition and Dimension Effects

Individual difference measures were of two types: 1) demographic variables; and, 2) attitudinal variables. The effects of both groups of variables were investigated for differences across conditions and within performance dimension ratings. The pattern of results were interesting in relation to the topic effects discussed above. Nine of the 16 demographic variables demonstrated significant differences between conditions. Yet, the only demographic variable to indicate a significant difference between performance dimension ratings was sex. Furthermore, there were significant topic main effects within 11 of the demographic variables, but there were no topic by demographic variable interactions. The relationship between the demographic variables and the condition effects is most certainly an artifact of the assignment of classes to conditions. However, taken together these results suggest that the demographic differences may have influenced the topic by training condition interaction discussed above. For example, in Condition

1 67% of the trainees had prior exposure to the SFP topic, while in Condition 3 40% of the trainees had prior exposure to this topic. Thus, this demographic difference between the two conditions may have influenced how trainees experienced the topics in each condition. However, the relationship between these three variables (demographic differences, training conditions, and topic ratings) is not simple or easily explained in all cases.

The second group of individual difference variables employed were rater attitudes about the lecture and lecturer. In this case, there were no condition differences in rater attitudes. There were significant correlations between attitudes and dimension ratings. In general, the results demonstrated a positive relationship between attitudes and dimension ratings such that higher posttest dimension ratings were related to more favorable rater attitudes. However, the magnitude of the relationship was different between the two topics. The relationship between attitudes and behavioral dimension ratings was stronger for the CS lecture than the SFP lecture. This suggests that it may have been the degree of positive attitudes toward the CS lecture that was contributing to the topic effects evidenced throughout some of the analyses. There is support in the literature that the more positive the rater attitudes are toward the ratee, the more positive the ratings become (Landy, 1985). Therefore, the positive attitudes toward the CS lecture may have elevated the CS ratings or the higher ratings may have

contributed to more positive attitudes. More importantly, however, the pattern of correlations between the attitudes and dimension ratings were similar across both topics. Thus, both topics were similarily influenced by rater attitudes.

Intercorrelations Among Accuracy Indices

Another aspect of this project was to investigate the interrelationship among accuracy indices, since the literature has recently questioned the relation between these measures and whether each offers unique information, or if results of one measure can generalize to another (Becker & Cardy, 1986; Roach & Gupta, 1990; Sulsky & Balzer, 1988). The relationship among these seven indices was tested here through calculating a correlation matrix of the indices assessed across all conditions.

The results demonstrated that the majority of the accuracy indices were correlated with one another; 76% of the 21 correlations were statistically significant. SA seems to have provided the most distinct information, demonstrating the fewest significant correlations with other measures, whereas the other indices evidenced between four to six significant relationships with one another. While there was a large number of significant relationships, the magnitude of the correlations were low to moderate in the majority of cases. There was also the identification of a large number of negative relationships. This discussion will first explore the results among the Cronbach indices and

then among DISTA, LENA, and ABVH, finally they will be taken together. Conclusions and recommendations will be offered last.

Elevation (EL), Differential Elevation (DE), Differential Accuracy (DA), and Stereotype Accuracy (SA)

Among the four Cronbach measures three of a possible six correlations (50%) were statistically significant, indicating that among the indices there is some uniqueness between measures, yet there is also shared information. SA was not correlated with any of the other measures, and only DE and DA were positively. SA's lack of any association with the three other measures, as suggested by Cronbach (1955), demonstrates that this index offers distinct information about rating accuracy; that the ability to identify ratings across all ratees at the dimension level is unrelated to other rating abilities. The relationship between DE and DA suggests that the ability required to make distinctions between ratees goes beyond a general rank order ability, or general impression of each ratee and is related to the ability required to decompartmentalize judgements for each dimension between ratees.

What is surprising among these indices is the occurence of negative correlations. There was evidence of a negative relationship between EL and DA and DE. In this case the results support the conclusion that accuracy in estimating overall group ratings is contrary to the ability required to accurately distinguish between ratees either in overall performance or at

the dimension level. It is possible that the cognitive processes involved in these judgements are different for the different measures. Accuracy in determining overall ratings may involve a simpler decision heuristic (an automatic judgement) whereas decisions between ratees may involve more complex integration of behaviors and weighting strategies to form evaluations. Using one type of decision strategy may serve to impede the other.

Distance Accuracy (DISTA), Leniency Accuracy (LENA), and Absolute Value Halo (ABVH)

The three additional indices demonstrated a great deal of overlap with one another. DISTA, ABVH, and LENA were all positively related to each other, and DISTA and LENA demonstrated near complete agreement (r= .92). The strong relationship between DISTA and LENA accuracy is not surprising since both are based on a difference score between experts and raters over ratees. DISTA requires the calculation of an absolute average deviation and can be thought of as a "city block metric" (Davison, 1985) of how far raters are from experts over all dimensions and ratees, whereas LENA is also a distance measure but is based on the differences between overall average true scores and overall average rater scores. LENA looks at leniency in ratings using the experts' rating as a profile of true leniency or severity. Sulsky and Balzer (1988) found this same relationship in two other data sets. These measures appear to share much common accuracy information. Conceptualizing ABVH and LENA as measures of error,

the positive but moderate relationship between the two ($\underline{r}$= .33) is supported in the literature. Generally, there are similar but inconsistent effects of both due to training (see Smith, 1986 for review). Thus, while related there appear to be different abilities necessary for the accurate determination of these judgements.

All Indices

An interpretation of the relationship between the four Cronbach measures and the three other indices together is more complex. In this case, of 12 correlations 10 were significant (83.3%). This indicates a lack of completely independent information regarding different aspects of rating accuracy. However, the magnitude of the absolute median correlation was moderate ($\underline{r}$= .22). The strongest relationship occured between EL and DISTA and LENA ($\underline{\tilde{r}}$= .96). This suggests that the comparison of these three with one another is redundant. In this case, as conceptualized the rating abilities required for each index are consistent with one another, and in fact their operationalizations are very similar. Sulsky and Balzer (1988) report that LENA squared provides results identical to EL, as was the case here, but suggest that LENA is a more interpretive measure since it goes beyond revealing a systematic difference between rater judgements and expert scores by including knowledge of systematic elevation or depression in scores.

SA, which was unrelated to the other three Cronbach measures, did positively correlate with DISTA and ABVH. While the

correlations were low there does appear to be some similarity in these indices. Regarding SA and ABVH this is not surprising given the conceptual and operational definitions of these measures. The relationship between SA and DISTA (r= .22) is more confusing. This suggests that the accuracy of dimension ratings (SA) is related to an overall difference measure between experts and raters of ratings across ratees and dimensions (DISTA). DISTA is conceptualized as similar to EL, yet there was no relationship between EL and SA. Thus, in one case there is evidence of related abilities between these two accuracy indices (SA and DISTA) and none between the other (SA and EL). This anomaly supports conclusions in the literature that different operationalizations of the same or conceptually similar indices can result in divergent results (Becker & Cardy, 1986; Sulsky & Balzer, 1988).

Five of the correlations between the Cronbach indices and the additional three accuracy measures were negative. DA and DE demonstrated negative relationships with DISTA and LENA, and the relationship between DA and ABVH was also negative. As measures of accuracy this suggests that the raters who are more accurate at more general judgements over all ratees and dimensions (DISTA and LENA) can less accurately distinguish differences between or within ratees. This interpretation is supported by similar results between EL and DA and DE discussed above. Additionally, those raters who can accurately determine halo and leniency vis a vis experts' judgements of these rating properties are less

accurate in identifying differences between or within ratees. Additionally, if ABVH and LENA are reconceptualized as error indices within the knowledge of expert true scores, this represents further support for the belief in the literature that decreasing error does not lead to an increase in accuracy.

Overall

The results of the relationship between these indices coincide in some respects and differ in others from that reported by Roach and Gupta (1990) and Sulsky and Balzer (1988) in similar investigations. With a BARS rating format Roach and Gupta (1990) reported a negative relationship between EL and DA, which was also found here. Conversely, Sulsky and Balzer (1988) found a positive correlation between these indices. Roach and Gupta (1990) also demonstrated positive correlations between measures of SA and DA, and Sulsky and Balzer (1988) found a negative correlation between these indices. Here, SA presented the most unique information and fewest signficant correlations. Sulsky and Balzer (1988) also demonstrated three more of the same relationships as found herein. These occured between LENA and EL, DISTA, and DA. However, for one of the three correlations (LENA and DA) they found a positive relationship whereas here the relationship was negative.

Taken together these three studies show convergence and divergence between results. This suggests that the pattern of

results are not necessarily stable across studies, and that there may be some third, unknown, influence affecting these relationships. Suggestions as to variables that may interact with the interrelationships would include differences across rating scale formats (e.g., Roach and Gupta demonstrated differences in patterns of results between a BARS format and a Graphic Rating Scale), differences in the saliency of the behaviors that are observed, or simply differences in types of performance vignettes (i.e., lecturer performance, assembly line work, and managerial behavior).

Roach and Gupta (1990) and Sulsky and Balzer (1988) conclude that each of the accuracy indices present distinct information pertaining to raters' abilities. The results in this research suggests likewise. However, in this data set there exists a large number of significant correlations between the indices, albeit of moderate magnitude. These interrelationships should not be ignored in the search for developing greater understanding of shared cognitive abilities and decision strategies across accuracy indices. Even more critical between these data sets is the identification of different patterns of relationships among those indices that are operationally similar. This suggests that these indices do not generalize and while similar are still capturing different aspects of rater judgement.

DISTA, LENA, and ABVH appear to contribute some new information to the Cronbach indices with the exception of the

high correlations with EL. However, Sulsky and Balzer (1988) question whether at the conceptual level ABVH should even be considered an accuracy measure given that the comparison between ratings and true scores occur between variances rather than between actual ratings as do the other measures. Additionally, the preponderance of negative correlations between these measures and the Cronbach indices argue for their inclusion as additional information leading to finer interpretations of the facets of rating ability.

The results found here are important for several reasons. First, and foremost, these results support previous research (e.g., Becker & Cardy, 1986; Murphy & Balzer, 1986; Roach & Gupta, 1990; and Sulsky & Balzer, 1988) which points to the lack of generalizability among the various indices. In previous research this conclusion is drawn based on few and low intercorrelations between the indices. Here this same conclusion arises from a different accumulation of evidence. In this case there were many more, and often stronger, relationships. Yet, the relationships in this data occured between different indices than in other data sets, suggesting that there may be other mediating variables (i.e., the manipulations, scale format) that interact and contribute to the evidence of rating accuracies or abilities and the relationship between indices. Furthermore, the large amount of negative relationships should be a warning that enhancing one type of accuracy may actually decrease that of

another. However, without some base rate of comparison it would be difficult to determine if this was a pre-existing difference or occurred due to aspects of the manipulations. For those indices that are unrelated to other indices, this supports the notion that these measures are clearly tapping different abilities. This also demonstrates the lack of generalizability across studies that utilize varying measures, or even studies that utilize the same conceptual accuracies but employ different operationalizations.

This research in combination with others serves as both a warning and a direction for future research. The warning is to consider carefully the choice and operationalization of accuracy scores. Whether training or other types of manipulations appear successful may depend upon the accuracy measures assessed. For example, if the purpose of training is to encourage raters to utilize a task focused observation schema to determine success of individual behaviors then the accuracy indices used for evaluation should reflect these distinctions and included SA and perhaps DA and ABVH. However, if the goal of training is to provide a person focused observation schema that encourages overall distinctions between individuals then DE might be used along with DISTA or LENA. The purpose of the training should guide the choice of indices.

The suggestion offered here is to include as many dependent measures as possible and to explore fully their relationships. As a direction for future research, the pattern of accuracy results should be explored as it relates to training programs, both in method and intent of training, in an attempt to discern explanations for training influences on different rating abilities, both those that enhance and reduce training effectiveness. For example, investigating the underlying cognitive processes promoted through different training methods can be identified as they increase or decrease individual rating decision strategies that the indices encompass. Knowing these properties should help in better understanding training effects. Finally, it is hoped that with the accumulation of future evidence we will eventually be in a position to match effectiveness indices with training outcomes and be able to choose rationally the most parsimoneous indices to employ in our research.

Hypotheses Testing

The first three hypotheses investigated the superiority of different combinations of the components of FOR training. Superiority was operationalized as training that increased indices of accuracy and decreased indices of error. Overall there was little support for any of the three hypotheses. However, the pattern of results are suggestive of differential effects for several of the training conditions. Additionally,

the method of investigation employed in this research made it possible to separate the effects of different methods and suggest the underlying processes involved in the components of FOR training. These will be discussed as they shed light on the differential influence of methods and intents of performance appraisal training that exist in this area of research, and as they relate to our current understanding of the cognitive processes involved in developing performance appraisal ratings.

Hypothesis 1

The initial hypothesis was that the complete FOR training program, utilizing all components originally proposed by Bernardin and Buckley (1981), would be superior to those conditions where one or more training components were omitted. With regard to the accuracy indices this hypothesis was only partially supported. With regard to the error indices the converse relationship was found to be true.

Accuracy was investigated through both aggregate scale measures and trichotomized factor scale measures. It will be recalled that the three factors were broken into scale items that addressed: 1) personal aspects of the lecturer (Factor 1); 2) cognitive aspects of the lecture (Factor 2); and, 3) both personal aspects of the lecturer and cognitive aspects of the lecture (Factor 3). In terms of the aggregate accuracy indices, only DE demonstrated a difference between conditions. This difference occured between Conditions 1 and 3. In this case

trainees who received the entire FOR training program were more accurate than trainees who did not receive feedback of true scores and behavioral rationales, and participate in discussion. These condition differences were also found for DE1, DE2, DA1, and ABVH1. In addition, Condition 1 was also more accurate than the remaining four conditions on DE1. Conversely, Condition 1 was less accurate than all other conditions on DE3; was less accurate on SA3 as compared to Conditions 2, 4, and 5; was less accurate on DISTA3 as compared to Conditions 3 and 5; and, was also less accurate than Condition 5 on LENA2. Overall, of the 21 trichotomized accuracy indices Condition 1 was most accurate on four measures and least accurate on four. There were no condition differences on the remaining 13 accuracy indices.

There were no significant differences between the complete FOR training program and the other training component combinations on any of the error indices except for Variance Halo. However, the pattern of results for Variance Halo are completely opposite of what was hypothesized. In this case there were significant differences between Conditions 1 and 3 on the CS lecture for the aggregate scale measure and for Factor 1. There was also a significant difference between Conditions 1 and 3 on Factor 1 when the topics were combined. However, what is interesting here is that in all cases Condition 1 evidenced the most Variance Halo and Condition 3 evidenced the least. No differences between conditions were found for Rater x Ratee

interactions, the second index of halo. However, reconceptualized as a measure of discriminant validity, Condition 5 had the greatest discrimination and Condition 1 had the least. Condition 4 resulted in the second highest discriminant validity index. In general, no significant differences were found for Interrater Reliability, and no differences were found for Midpoint Leniency.

A Cognitive Perspective of FOR Training

Through reviewing the cognitive literature specifically pertaining to the performance appraisal process (i.e., Borman, 1978; Cooper, 1981; DeNisi et. al., 1984; DeNisi & Williams, 1988; Feldman, 1981; Ilgen & Feldman, 1983) several interpretations for the results here appear plausable and informative. By addressing the components of FOR training with a cognitive perspective, the underlying processes of each component can be discussed as they may have contributed or detracted from the rating abilities considered inherent in the different accuracy and error indices. This approach to interpretation is offered as tentative since FOR training was not originally proposed as a cognitive training strategy nor was the investigation here intended to be primarily cognitive in nature. Additionally, this discussion will focus mainly on Hypothesis 1 since the two other hypotheses demonstrated fewer results and interpretation is facilitated through a more thorough discussion of the first hypothesis.

It is generally accepted that individuals use different organizing schemata or categorization strategies to observe,

encode, retrieve, and evaluate information which facilitates greater efficiency in understanding (Taylor & Crockett, 1981). There are many different schemata that can be used by an individual for the purpose of remembering information. Examples of types of schemata include person, event or role organization (Cardy et al., 1987) and worker or task blocked strategies (Williams, DeNisi, Meglino, & Cafferty, 1986). It is clear that the underlying organizing schema promoted in FOR training is task blocked. That is, the goal in this training is to develop in raters the understanding that performance is comprised of distinct dimensions of behavior. Yet, this training is also criterion-referenced since each performance dimension can have varying levels of poor, average, and superior performance. The training involves developing in raters a greater understanding of the distinctions or "frames-of-reference" of what constitutes these performance standards. The thread that runs throughout the training is the focus on dimensions of behaviors (tasks) and the development of consistent behavioral rationales for distinguishing varying levels of behavior (task focus). The steps of FOR training used here promoted this framework to varying degrees and in different manners.

The complete version of FOR training was more accurate than training that did not include feedback of true scores, behavioral rationales and participant discussion on accuracy indices that reflect raters' abilities to distinguish between ratees across dimensions

on the aggregate measure (DE) and on the personal (DE1) and cognitive (DE2) factors. This was also true for the personal factor accuracy indices that reflect the ability to differentiate between ratees within dimensions (DA1) and the intercorrelation within dimensions (ABVH1). It is likely that the personal aspects of performance were easier to identify or more clearly delineated in the performance vignettes than cognitive aspects of the lecture or cognitive and personal aspects combined. Even as naive raters, individuals are more likely adept at identifying personal characteristics of performance (i.e., comfort level, voice intonation) than more cognitive characteristics of performance (i.e., logical transitions) or dimensions that combine both.

Conversely, when the complete version of FOR training was least accurate this occured for factor accuracy indices that combined personal aspects of the lecturers' performances and cognitive aspects of the lectures. It is likely that the observation, encoding, retrieval, and evaluation of these aspects of performance are more difficult even for the most experienced raters. However, when the complete version of training evidenced the least accuracy there was no one best combination of training components across all indices. It appears that by removing various components that serve to facilitate the organization of the task related schema, or the understanding of the behavioral

rationales for evaluating performance levels, certian types of accuracies were increased.

Feedback of true scores and behavorial rationales, and discussion. The purpose of true score feedback and behavioral rationales was to provide trainees with greater understanding of the evaluative schemata used for encoding and evaluating performance judgements. The true score information should have provided knowledge of actual performance levels and the behavioral rationales should have helped in identifying aspects of behavior that were relevant for evaluating a performance dimension. Research demonstrates that once an item is labelled as relative to a schema it tends to be used in subsequent decisions (Higgins, Rholes, & Jones, in DeNisi et al., 1984). Additionally, it is suggested that reinforcement of the schemata should increase its saliency. Both these aspects were among the goals of this component, along with increasing the evaluative strategy.

This training component led to greater accuracy on the aggregate measure and the personal factor. DeNisi et al. (1984) suggest that if the saliency of a schema is high and as a rater's experience increases there will be a tendency to recall general impressions and overall evaluations. Additionally, if the feedback increased the saliency of the schema which aided in categorization, then it is more likely that the subsequent recall would involve more general impressions than specific behaviors (Nathan & Lord, 1983). This seems to have been what occured

through the incorporation of these components when the accuracy indices reflect distinctions between ratees in overall performance (DE and DE1) and distinctions within ratees in patterns of performance (DA1) that are highly salient, as well as the intercorrelation of highly salient performance dimensions (ABVH1). The feedback most likely led to a schema oriented and evaluative general impression that contributed to greater accuracy on indices that may actually be comprised of behaviors that profit from this effect.

Omitting feedback of true scores and behavioral rationales, along with participant discussion decreased halo. Contrasting this with the accuracy indices that demonstrated the inclusion of these components contributing to accuracy, it is possible that judgements of halo, in part, help achieve greater accuracy in some rating indices. This speaks to the issue of "true halo" as representative of true similarities in behavior levels between interrelated task dimensions. This also lends support to those who believe that decreasing halo should not be a goal of performance appraisal training (Borman), as well as those who demonstrate the inverse relationship between accuracy and error (Becker & Cardy, 1986; Fisicaro, 1988; Murphy et al., 1982).

Scale Orientation. The most consistent finding was that omitting scale orientation led to higher accuracy on the combined factor indices for overall distinctions among ratees (DE3), accuracy of dimension ratings across ratees (SA3), and on the

absolute value of the difference between trainees and experts across all dimensions and ratees (DISTA3). The literature suggests that scale orientation serves as a framework for raters to use in observing and encoding information (Ilgen & Feldman, 1983) and should serve a priming function. In this case, the scale was organized by behaviors and thus should have provided a task blocked schema for encoding behavior. That is, the introduction of the scale would prime trainees to be alert to specific behaviors as opposed to traits or general impressions, and that subsequent evaluations would involve independent evaluation of behaviors. A task blocked schema has been shown to have greater accuracy in performance ratings (Cafferty et al., 1986) and by increasing controlled processes of observation, leads to greater discrimination between ratees (Williams, Wickert, & Peters, 1985).

In the case of the combined personal and cognitive factor, to employ the task focused observational strategy encouraged through scale orientation would involve a complicated process of encoding together behaviors relevant to two different behavioral features, and then retrieving and evaluating it through a complicated weighting schema. This process would also have to be superordinate to the schema which calls for independence of behaviors. For DE3, SA3, and DISTA3 the task oriented priming and the notion of independent dimensions may be too difficult to follow through observation, encoding, and evaluation. The

accuracy indices all tap the ability to determine co-relationships either within ratees, dimensions or both. Removing this component may discourage the independence of dimensions and may encourage the detection of relationships. Furthermore, in another study of the priming effect of scale orientation, Cardy et al. (1987) failed to find an increase in rater accuracy by including this component. Thus, while the theoretical literature suggests its usefulness, research does not appear to support this notion, especially when the judgements require integration of less salient or more complicated information.

Removing scale orientation also demonstrated the greatest discriminant validity. Discriminant validity addresses distinctions between ratees (person blocked). Therefore, the effectiveness index was inconsistent with the organizing schema and scale format (task blocked). By removing this component the task blocked sequence of processing may have been less distinct as a strategy to follow and thus increased the raters abilities to discriminate between ratees. Additionally, by attending closely to the rating schema raters may not have been able to attend as closely to the behaviors demonstrated on the performance vignettes.

Behavioral Rationales. Accuracy ratings of specific dimensions of the combined factor were increased in one case by removing feedback of behavioral rationales, and by not requiring trainees to write out behavioral rationales in another. The

purpose of the behavioral rationales was to facilitate greater understanding of the distinctions in performance levels and how to combine different aspects of behavior that were relevant to a task. Here again, the underlying schema was task blocked. However, the presentation during training was within ratees since each of the two practice occassions were complete and distinct units.

The behavioral rationales provided by the experts and expected of the trainees in the combined factor required the integration of both cognitive and personal aspects of performance. The behavioral rationales may have generated information that was too complicated to understand and integrate during the short time span of training. This is particularly true when the aspects of performance that contributed to the task ratings were difficult to detect and categorize. By removing these components, accuracy may have increased by allowing a more automatic and streamlined response to ratings of dimensions across ratees (SA3). Not providing or requiring a detailed and multifaceted rationale may have allowed for a more implicit evaluation. Removing these components also increased the accuracy of distinctions between ratees in overall performance on the combined factor (DE3). It is suggested that this automatic response also led to greater accuracy in this case by allowing holisitic distinctions between ratees. Additionally, research has demonstrated that raters naturally organize information around persons (Srull & Brand, 1983) and removing this component

may have facilitated encoding, retrieval, and evaluation of less salient information in a less controlled, and in this case, more accurate manner.

Omitting the written behavioral rationales for each performance dimension resulted in the second highest discriminant validity index. Again, there is inconsistency between the organizing schema and the effectiveness index. Additionally, when the trainees knew they had to provide an explanation for behavior there may have been preoccupation with searching for behavioral rationales and this may have detracted from attending to how the behavior functioned in determining a ratee's overall performance.

Taken together these results suggest that the most effective combination of components is dependent upon the type of accuracy or error judgements solicited, and the type of behavioral dimensions assessed. For example, if it was important to assess overall differences between individuals for the purpose of assigning bonuses, and the rating judgements involved behavioral dimensions which referred to personal attributes, then the complete FOR training program would be most effective. However, if the same overall differences between individuals involved judgements that were more complex and that integrate different types of information about how well the individuals knew their job and how comfortable they were performing the job, then the complete FOR training program would lead to the least accuracy.

In this example, an effective and also more efficient method would be one that contained true score feedback but omitted explanation of the behavioral rationales leading to these scores and participant discussion. Additionally, since no differences were found for the accuracy of the average rating (EL), in those situations when accuracy of the rater is important or accuracy of group ratings, apart from detecting differences in dimensions or ratees, then training that involved the least effort in time and money would be as accurate as training that was more involved. Or, it may even be that to achieve accurate overall average ratings, no training is required. For overall accuracy, a rater's implicit schemata may be better than one provided through training. Thus, the search for one best training method or strategy may be an unrealistic and unnecessary goal. Rather the results here support the pursuit of identifying the best training paradigm that matches the intended purpose of the ratings or the decisions that need to result and the type of judgements that are instrumental in these decisions.

Hypothesis 2

Hypothesis 2 proposed that the more active training conditions would be superior to the more passive training conditions. Activity and passivity were used loosely here to refer to the amount of participation and activity trainees were allowed during training. There were no differences between the more and less active conditions on any of the criterion measures,

with the exception of Interrater Reliability. However, in this case the results appear to be data dependent. The differences in Interrater Reliability found between conditions may have been due to the large differences in sample size across combined conditions and may not generalize to other comparisons with more similar sample sizes. Therefore, activity as defined here, was not a critical distinction between training components. It would be interesting in future research to assess whether trainees perceived that they were more or less active in these conditions. A manipulation check on their perceived activity level would provide evidence as to their own determination of the degree of participation in the training, and how that alone may have influenced the training outcomes.

Hypothesis 3

Hypothesis 3 proposed that within the less active training conditions there would be a hierarchical ordering of effectiveness, such that the least active, that which did not include feedback of true scores and behavioral rationales, and participant discussion, would be less effective than the condition that included these components but excluded writing out the behavioral justifications for the ratings. There was no support for this hypothesis within the accuracy indices. In regard to the error indices there were some significant differences which support the converse of this hypothesis. For Variance Halo the results again demonstrated more halo for the condition that

included feedback and discussion. This occured for the personal factor on the CS lecture, and the two lectures combined. Thus the more active training condition resulted in more halo error. Yet, it should be noted that there is a confound here. The condition which included the more active training, also excluded writing out the behavioral rationales. While it can be suspected that this exclusion may have contributed to greater Variance Halo, the pattern of results reported thus far support the inclusion of the other components as contributing to the halo results. Additionally, the more active condition demonstrated more discriminant validity than the more passive condition.

Based on the pattern of results identified thus far, it is suggested that the important distinction among these two components is the true score and behavioral rationale feedback. Omitting this component led to a reduction in Variance Halo, however it also led to less discriminant validity. This component may have demonstrated to trainees, through the explanation of how different pieces of information are weighted together and through a low range of variability across dimension scores, the degree of true interrelationship between dimensions. Demonstrating the true interrelationship of performance dimensions should also increase the saliency of the rating schema as suggested by DeNisi et al. (1984) by increasing the retrieval of general impressions. While these general impressions serve to increase Variance Halo, this does not necessarily mean it is

inaccurate. This interpretation is supported by the finding that while Variance Halo was decreased through removing feedback, discriminant validity was also decreased, and accuracy in overall distinctions between ratees (DE) was decreased as well. By comparison, the more active component which included feedback but excluded the writing of behavioral rationales had more Variance Halo but also had greater discriminant validity. Thus, the saliency of the dimension interrelationships may have been increased but this may have also led to greater discrimination. By not requiring written justification for the ratings the halo engendered by the feedback was allowed to remain strong.

Overall

Overall there was not strong support for the three hypotheses investigating the superiority of the different components of FOR training. Therefore, those conclusions that are drawn can only be presented as tentative. However, within the significant results they present a pattern that does allow for some general and suggestive inferences. First, as stated above, there appear to be differential effects for the various methods and underlying processes of the six components included in FOR training. This suggests that there is no one best combination of training components. Rather, the most effective combination of training methods and strategies involves a somewhat complex interrelationship between the types of behaviors to be rated and the goals of the training in regard to the

criterion measures. Also, it appears that when the judgements involve integrating more than one type of information, the complete version of FOR training may actually decrease accuracy by presenting a task focused organizing schema for observing, encoding, retrieving, and evaluating information that becomes too cumbersome a strategy for identifying less salient information. This appears especially true when the accuracy ratings are addressing a general impression effect or more holistic ratings of complex behavior. These components may contribute to decomposed ratings which are not always the goal of the accuracy indices. In these cases the trainees may become confused which reduces accuracy.

However, this research does not address whether the information offered in the training leads to lower accuracy when trainees utilize the information, or whether the confusion leads trainees to reject the training and rely on some other judgement strategy. The more information trainees receive regarding schemata constructions the less certain they are in determining what information to utilize in forming their decisions (Feldman, 1981). Different methods of investigating this issue would be to provide stop points during the training program to ascertain whether trainees understood the elements of each training component (i.e., providing tests of learning); simply asking them to rate the importance of the information they learned in each component in arriving at their final ratings (i.e., "was this component

helpful?"); and asking them about their confidence in their ratings. Additionally, it is suspected that if they did reject the information, they would then utilize a more natural person focused strategy which may lead to increased accuracy on some indices.

Also, it is unclear whether trainees accepted the true score feedback and behavioral rationales as valid information. Trainees were told that the experts were industrial/ organizational psychologists with experience in developing performance appraisal systems and, as academics, with experience in evaluating lecturer performance. The social influence literature suggests that information from credible sources is more easily accepted as true (Aronson, Carlsmith, & Turner, 1963), and should influence trainees' opinions on the validity of this information (Moscovici, 1985). It seems reasonable that this was the case here. If they did not accept the information as accurate research suggests that they would reject the schemata and rely on prior strategies (Lord, 1985); most likely person focused.

Conversely, when the judgements involve more direct or salient information FOR appears to enhance ratings. Additionally, when the amount of information that needs to be integrated is limited the important elements in FOR training for increasing accuracy appear to be feedback of true scores, explanation of the behavioral rationales, and participant discussion. Yet, in regard to halo, these components increased halo and decreased

discriminant validity. The components of FOR training did not appear to differentially influence other traditional indices of error.

Possible explanations for the lack of support for these hypotheses arise from a few avenues. First is the use of students as trainees. The student sample may have lacked motivation or perspective in regard to performance appraisal training. This was addressed by reinforcing the dual purposes, to design a training program for the College to use to increase student accuracy in teacher evaluations, and to provide feedback to the Chairman regarding the job candidates for the next term as well as feedback to the candidates themselves. It is possible that this motivation was not strong enough, even though anecdotal evidence appeared otherwise (See Chapter III). Purpose of performance appraisal has been found to influence ratings. Ratings for adminstrative or counseling purposes tend to be less severe than ratings for experimental purposes (Zedeck & Cascio, 1982). In this study it is possible that while the purpose stated was motivational, it may also have served to temper ratings which could have detracted from the detection of significant differences between conditions; given relatively low accuracy indices.

It is also possible, and more likely, that the manipulations were not strong enough. The differences between some conditions were slight and may not have been different enough to result in

significant contrasts. When the differences were very obvious (i.e., between Conditions 1 and 3) distinctions between indices were detected. Additionally, the majority of results occured for the factor indices suggesting, as was suspected, that collapsing over dimensions may obscure detection of differences. This also indicates that a more behaviorally focused investigation of interrelated dimensions can lead to greater demonstration of results and potential interpretations. Perhaps a different combination of factors, based on different rationales, or even empirically derived factors would demonstrate different or more interpretable results. Finally, FOR training as originally proposed was intended for those raters who exhibited idiosyncratic frames-of-reference regarding performance levels. FOR training was designed to help these raters recalibrate their judgements to adopt the appropriate evaluative standards, so as to not have individual differences among raters influence ratings. Thus, if in this population many of the trainees already shared a similar frame-of-reference then this would explain the insignificant results failing to demonstrate the superiority of FOR training.

Furthermore, another constraint exists in the fact that this study was conducted in a laboratory setting with a very limited time frame. In an applied setting observation and encoding would occur over a longer period of time and would involve many more samples of behavior. Thus the information gathered would be

richer and also include more contextual variables. It is likely that the task focused organizing and evaluation strategy would be more firmly established through the passage of time and by viewing behaviors across more ratees on numerous occassions. In this case the vignettes were rather sterile, did not include contextual variables, and the training may not have been as successful at breaking down pre-existing schemata in a short period of time and with limited practice.

Finally, in this study the comparison condition was the complete FOR training program. It would have been interesting to include a condition that did not receive any training. This may have provided evidence as to whether trainees were utilizing the information received in training if no differences occured when compared to this condition. Or, it may have offered a clearer indication of the benefits of certian components if they were superior to no training.

As the first investigation to systematically vary the elements differentiating methods and processes of performance appraisal training and to explore the components with respect to many criterion measures, the results are promising. Future research should continue this avenue of investigation to provide more explanation and understanding of the components of performance appraisal training. Two goals of this type of research should be: 1) to determine the optimum combination of training components necessary to achieve certain types of

accuracy; and, 2) to explore training at the component level to determine more specifically what cognitive processes the components are facilitating as they contribute to different abilities inherent in indices of accuracy and error. The former goal could be addressed through employing the same methodology used here, but as opposed to using a subtractive model, different combinations of components should be explored. The latter goal could employ training on each component individually and determining the effect of each on the evaluation indices. Finally, it would be interesting to explore a method of documenting the decision strategies utilized in determining rating judgements to see if trainees actually employ the training techniques or return to a different strategy when the judgements are difficult. This may be a simple matter of incorporating manipulation checks at each stage, and/or increasing the difficulty level of the rating judgements.

Individual Differences As Related to Accuracy

Both sets of individual difference variables, demographic and attitudinal, were investigated for effects on accuracy. There was only one demographic variable (e.g., language) that demonstrated a relationship with two of the accuracy indices (e.g., DA and ABVH). Included in the remaining demographic variables were questions pertaining to previous experience with performance appraisals, both as the appraiser and appraisee, and

questions pertaining to life situations that may have increased sensitivity to the CS lecture. While finding no relationship with accuracy for the latter variables is considered a positive result, no relationship for the former is surprising. Previous research has demonstrated that increased experience with performance appraisals is related to greater accuracy (Cardy et al., 1987) however, this was not supported here. A possible explanation for the lack of support is that while both the samples were comprised of students, even for those who did have more experience in this sample, the experience may have been more limited than that in the Cardy et al. (1987) sample. The predominate lack of effects of the demographic variables on the accuracy results is reassuring given the effect demographic variables had on topic differences. Thus it is reasonable to conclude that the demographic differences detected in the topic effects did not influence the accuracy indices.

Rater attitudes were investigated separately for each lecture topic. Overall, the pattern of results were similar across four of the seven accuracy indices. Additionally, the variable measuring whether raters would like to learn more about the topic was not related to any of the accuracy indices. Once again there were no significant relationships for SA, which supports the conclusion that SA stands apart from the other indices and may offer the most unique accuracy information.

Accuracy indices calculated over all ratees (e.g., EL, DISTA, LENA) were positively related to attitudes towards the lecture and lecturer, whereas these same indices were negatively related to positive attitudes towards the lecture topic. It appears that as attitudes regarding the lecturer and his performance became more positive, the accuracy of the rater in making overall ratings increased. Conversely, as the topic became more interesting, the accuracy of the rater decreased. It is possible that as positive attitudes toward the lecturer increase rater accuracy increases, whereas, as the rater's attention is drawn into the topic itself, overall accuracy decreases. This pattern was evident for both topics. This suggests that the observation, encoding, retrieval, and evaluation process differs in these two cases. A tentative explanation may be that as attitudes towards the lecturer became more positive the rater was inclined to pay careful attention to specific behaviors in order to provide more helpful feedback to the lecturer. They were told that this was one of the reasons the lecturer agreed to be videotaped. This increased "helpfulness" may have led to greater accuracy. Conversely, as interest in the topic increased the rating task and careful observation may have been set aside leading to decreased accuracy. Future research may profit from investigating differences in rater attitudes toward the person, or the activity

itself, through manipulating both likeability of the performer and an interesting versus a more mundane task.

The relationship between attitudes and accuracy of distinctions between and within ratees (DA and DE) were not consistent across the two topics. Distinctions between ratees (DE) on the SFP lecture demonstrated a similar pattern of results as described above, yet the correlations were very low. However, on the CS lecture the relationships were reversed. In this case as attitudes became more positive regarding the lecturer, accuracy decreased, but as interest in the topic increased, accuracy increased. This relationship appears consistent with the underlying ability DE is thought to measure. Since DE refers to overall distinctions between individuals it suggests that as attitudes toward the CS lecturer increased the distinctions between lecturers decreased. Most likely the trainees were inflating the CS lecturer's ratings. Conversely, as interest in the CS topic increased the trainees were still able to make accurate distinctions between ratees apart from topic effects. DA also demonstrated a similar pattern of results to DE on the CS lecture. Again, this is understandable given that DA measures the pattern of performance dimension distinctions within ratees. This suggests that positive attitudes may be leading to a general impression effect. However, DA was not related to any attitudes on the SFP lecture.

Previous research has demonstrated that attitudes influence ratings (Landy & Farr, 1980). This research goes further in providing evidence that suggests attitudes also influence accuracy. This also demonstrates that attitudes have a differential effect on accuracies which assess different rater abilities. Thus increasing raters' favorable attitudes toward the ratee can enhance or impede accuracy. Positive attitudes may enhance the overall ratings of an entire workgroup, but impede accuracy when distinctions between employees within a group are necessary.

The somewhat different pattern of results for DE and DA between the two topics suggest that the attitudinal differences may be accounting, in part, for the topic differences in posttest ratings. However, since no differences in attitudes were found across conditions, and the relationship between attitudes and dimensions ratings were similar across both topics, it is safe to tentatively conclude that these differences did not significantly affect training results. Future research should further investigate the effect of attitudes on accuracy and training manipulations perhaps by including descriptive information on each ratee varying dimensions of likeability. Future research should also explore this area in an attempt to understand the underlying cognitive processes and search strategies linked to attitudes and accuracy relationships.

The Influence of Memory on Accuracy

The fourth and final hypothesis of this investigation was that heightened observational memory, or knowledge for what was observed, would be related to increased accuracy on the seven indices of rating effectiveness. Partial support was found for this hypothesis. Heightened observational memory, operationalized as the accuracy of performance frequency ratings, was related to heightened accuracy on behavioral dimension ratings or performance evaluation ratings. However, this relationship only occured across ratings that assessed the same type of abilities. That is, when the memory accuracy pertained to overall ratings or ratings across ratees, heightened knowledge of this type influenced behavioral accuracy ratings that also pertained to overall ratings or ratings across ratees. Conversely, when the behavioral accuracy ratings addressed distinctions between dimensions or ratees, heightened knowledge of the opposite type decreased behavioral accuracy.

For example, heightened memory on Stereotype Accuracy (SA) was related to more accurate ratings on Elevation (EL), Distance Accuracy (DISTA), Leniency Accuracy (LENA), and Absolute Value Halo (ABVH), all of which occur across ratees. Yet, heightened memory on SA was also related to the least accurate ratings on Differential Accuracy (DA) and Differential Elevation (DE) where raters need to make overall distinctions between ratees and between dimensions within ratees. This pattern of results is

consistent across all accuracy indices, with the exception of the relationship between Memory Elevation Accuracy and LENA. In this case high memory accuracy was related to low behavioral accuracy. However, this is not an artifact of the data, but rather this relationship is operationally dependent since the square root of LENA is equal to EL. Thus, even this relationship is consistent.

These results partially support those found by Murphy et al. (1982b). Using the same measure of observational knowledge, based on observational frequency, they found correlations between observational accuracy and performance evaluation accuracy for EL, DE, and DA. They propose that knowledge accuracy is related to evaluation accuracy, which is the same conclusion drawn here. However, they found a relationship between DE and EL which was not found here. Additionally, the research here employed more operationalizations of the different components of accuracy, all of which demonstrated consistent relationships or lack of relationships. Furthermore, the analyses here were different. Murphy et al. (1982b) present correlational evidence, whereas here MANOVA's and post hoc analyses were conducted and observational knowledge was trichotomized into high, medium, and low memory. It is possible that these analyses were more sensitive to differences across indices, as well as not being dependent upon the assumption of linearity. While this research supports their conclusions, it adds a conditional clause. That is, memory or observational accuracy is related to accuracy of

performance evaluations but only when the same cognitive abilities underlie the rating judgements.

It appears that positive effects of heightened memory or knowledge does not necessarily increase all types of accuracy. Thus greater memory, in itself, does not necessarily lead to greater accuracy. Greater memory that derives from the same type of rating abilities as the accuracy index assessed, leads to more accurate ratings. However, certain types of memory or knowledge can also decrease accuracy indices that require opposing abilities. Therefore, increasing observational memory, in the abstract, will not always increase accuracy. The results herein demonstrate that observation, as a measure of knowledge, itself does not generalize to all accuracy indices. Rather, training that attempts to increase observational memory must be linked to the criteria of effectiveness. The observational strategies employed must be linked to the goals of the training and the type of abilities training proposes to increase. For example, training in the observation of performance distinctions between ratees should employ components that facilitate this process and should be evaluated through indices of DE and DA.

Summary of Conclusions

Overall, there are several conclusions that can be drawn from this research. First, within the typical performance appraisal training research paradigm, several design features can influence results. The conclusions here suggest that: 1) the

choice of stimulus materials is very important; and 2) order (context and contrast) and topic effects were demonstrable and should be allowed for in the design of future research. While these results did not appear to significantly affect the training effectiveness indices they do suggest that design elements need to be considered carefully in subsequent research. As stated above, they also should serve as a warning to future researchers that elements of the design must be controlled and/or investigated in conducting training research. The employment of manipulation checks on design elements would serve as an important control and interpretive mechanism. Additionally, these design elements need to be more adequately described in the published research so that the consumer can evaluate the research and have a more critical understanding of potential influences and idiosyncracies inherent in any one or combination of research designs and training outcomes.

Second, in this research it was possible to explore the potential influence of individual differences on the results. This was provided through analysis of both demographic and attitudinal variables. The pattern of results demonstrated that the demographic variables failed to systematically influence dimension ratings or accuracy indices. However, rater attitudes did demonstrate significant correlations with both dimension ratings and accuracy indices. Their relationship with dimension ratings appear to suggest that either positive attitudes toward a

lecture topic elicited higher ratings, or positive judgements of performance contributed to positive attitudes. Rater attitudes also demonstrated a differential relationship with the various accuracy indices. The results suggest that positive attitudes enhanced accuracy indices that tapped a rater's ability to assess overall ratings, whereas positive attitudes detracted from a rater's ability to make distinctions between ratees or between dimensions within ratees. Interest in the topic itself detracted from a rater's overall accuracy. A possible area of research might be to go further in investigating the relationship between attitudes and accuracy in an attempt to understand the cognitive processes that promote this link and investigate methods of removing this influence from ratings.

Third, this research presents additional information regarding the relationship among accuracy indices. It is suggested here that the pattern of interrelationships may be more significant than commonly believed. While many of the relationships were of low magnitudes there was a consistent pattern herein of results suggestive of some shared information among the indices. There were positive relationships among conceptually similar measures and negative relationships among dissimilar measures. It was proposed that different cognitive abilities may serve to enhance or impede accuracy. Accuracy on one measure does not necessarily indicate that there will be a parallel accuracy on another measure, and parallel accuracies

will most likely not occur if the indices are dissimilar. It was also suggested that comparisons of some measures may be redundant and not offer much distinct or new information to the investigation.

These results also serve as a warning regarding the lack of generalizability across studies using different measures given that the interrelationships found were of moderate magnitude. Training results may be very dependent upon the criterion measures chosen and this may, in part, account for the lack of consistent results across investigations. Several suggestions for future research were offered including the continued investigation of these interrelationships given the difference in results found aross this study and previous research (Roach & Gupta, 1990; Sulsky & Balzer, 1988), and specifically linking the investigation of training programs to particular indices in an effort to determine what rating abilities are facilitated within and across training programs. Finally, it was proposed that it is very important to link accuracy indices with the intent of the training program since without this consideration outcomes may go undetected.

Fourth, support for the hypotheses investigating the differential effectiveness of the components of FOR training was very limited. It appears that the relationship between the effectiveness of the various components involves an interplay of several elements including: 1) the types of rating judgements required; 2) if they involve integration of similar

types of information or integration of different types of information; 3) the underlying cognitive processes and strategies promoted by each component; and, 4) the abilities assessed within the accuracy indices.

The complete FOR training program appears to have the most positive outcomes when the judgements require integration of similar information. When the judgements require the integration of several categories of information FOR training appears to provide too much information or too complicated an organization schema that may overwhelm raters. When FOR training fared best it appears due to including feedback of true scores and behavioral rationales, as well as discussion in the training. When FOR training led to inaccuracy there was no consistency in results demonstrating the superiority of a different set of training components. In these situations removing scale orientation increased accuracy, as did omitting the writing of behavioral justifications for ratings, as well as not providing true score and behavioral rationale feedback along with discussion.

In regard to error indices, these either failed to demonstrate significant differences between training conditions or when differences were detected they indicated the inferiority of the complete version of FOR training in decreasing halo or increasing discriminant validity. Omitting feedback of true scores and behavioral rationales along with discussion led to the

least halo and the most discriminant validity in ratings. While these components increase error they seem to have a differential effect on accuracy. No significant differences were found for leniency.

Finally, the relationship between observational memory or knowledge was explored as it was related to accuracy in performance evaluations. It appears that observational knowledge does interact with evaluation accuracy but it is ability dependent. That is, memory that increased overall observation ratings also increased overall evaluation ratings. Conversely, increased memory in overall observations demonstrated decreased performance evaluation abilities that required assessments between individuals. This suggests that not all types of memory enhancement are desirable since they can also serve to impede certain rating abilities. Thus, training that teaches observational strategies or attempts to enhance recall may evidence differential effects across accuracy indices. It is important that the strategies be linked to the types of accuracies important in a given rating situation. It would be fruitful to explore this relationship within those training programs that are observationally or strategy based.

One of the primary goals of this research was to support and expand the results demonstrated in two other studies on FOR training in the performance appraisal literature (e.g., Athey & McIntyre, 1987; McIntyre et al., 1984). As reported in Chapter II, using a modified version of FOR training researchers found

that this training led to greater accuracy on measures of DISTA and ABVH and found no effect for LENA. The condition here most similar to their version of the training was Condition 4. While this condition was not specifically tested against the other conditions it was compared to the complete version of FOR training and compared to the conditions that did not include feedback of true scores and the corresponding behavioral rationales. The latter comparison was similar to a contrast they performed with what they termed an "Information Only" condition.

In the present study, Condition 4 was superior to the complete version of FOR training on the factor accuracy index of SA3, but was no better on ABVH and DISTA. When compared to the conditions omitting feedback no differences in accuracies were detected, but this condition was superior at decreasing Variance Halo and increasing discriminant validity. Therefore, the results here lend only limited support to those of McIntyre and his colleagues. However, there are several limitations that must be noted.

First, McIntyre and his associates' work compared a modified version of the training against other types of training paradigms (e.g., RET and no training). Here the comparisons were between the modified version and a complete or more severely modified version of FOR training. Thus, they found FOR training superior to that of other paradigms. Here, the modified version was superior (on three dependent measures) to other versions of the

same basic training paradigm. Additionally, if this modified version were compared to all other versions of the training its superiority might have been demonstrated. However, this is unlikely given the pattern of results discussed above. Secondly, the goal here was to support and expand this previous work by demonstrating the superiority of FOR training as originally designed by Bernardin and Buckley (1981). This support was not found. It appears that, as discussed, the original design may be providing too much information for evaluating less salient behaviors or tasks that encompass different types of abilities. This appears to be leading to decreased accuracy in subsequent ratings.

There were other differences between the research here and that previously conducted. Trainees here viewed two vignettes depicting high and low performance and then rated average performance vignettes. In McIntyre's research trainees viewed one practice vignette of an undisclosed performance level and rated three vignettes also of unknown performance levels. Additionally, trainees here participated in discussion whereas there was none in the other research. Overall, the conclusions that can be drawn from the research on FOR training is that the modified version may actually be better at increasing accuracy and decreasing error in many instances (i.e., when the ratings involve integrating complex information) than FOR training as originally conceptualized.

Two goals of this project were to bring greater clarity to the interpretation of effects of performance appraisal training, and to present a framework with which to consolidate research in the field. To varying degrees these were achieved. The question that now needs to be addressed is how this research contributes to the applied use of performance appraisal training. Table 26 presents a summary of the effects of the training components as they related to the accuracy indices. The Table also suggests underlying processes for the components and how these may relate the assessment of accuracy judgements and to the goals of training and performance appraisal use.

A question can also be asked concerning whether this investigation helps to narrow the gap between research and practice. Several specific limitation of this study are presented below. The question now is more general and concerns whether this training program and the variables investigated serve a practical applied purpose. It is the belief here that they do.

The trainees in this study will never be expected to return to a work environment and use their newly acquired knowledge to conduct more valid and accurate performance appraisals, nor was that ever the purpose of this study. However, the utility of this research is that as a prototype for performance appraisal training the information gained here can be utilized for training programs in the field. It is commonly agreed that there are

Table 26

Summary of Component Effects on Accuracy Indices and Relationship to Training Goals

Index	Ability Tapped	Index Relationship to Training Goals and PA Use	Component Increasing/ Decreasing Index Effectiveness	Suggestions
EL	Determination of overall ratings across ratees.	To increase accuracy of group ratings when ratings are to be compared across work groups, or performance level of an entire group is important.	No effect on this index.	
DE	Overall distinctions between members of a group.	To increase overall distinctions between group members when ratings are to be used for promotions, salary increases, bonuses, etc. When employee comparison techniques are important.	FB of true scores and behavioral rationales, and discussion increased accuracy on DE, DE1, and DE2. Same component decreased accuracy on DE3, as did the individual contributions of each of the components.	May have contributed to a general impression effect across salient behaviors, but provided a complicated organization schema for behaviors not easily observed and evaluated together. Best to use for highly salient behaviors. FOR training may not serve any of the goals for DE when behavior integration and evaluation are difficult.

Table 26 continued

Index	Ability Tapped	Index Relationship to Training Goals and PA Use	Component Increasing/ Decreasing Index Effectiveness	Suggestions
SA	Accuracy within dimension ratings across ratees.	To increase accuracy of individual dimension ratings when the detection is important to identify average ratings of specific task behaviors. When job analysis information is important or distinguishes between dimensions in a job.	Scale orientation, writing behavioral rationales, receiving FB of behavioral rationales and discussion each decreased accuracy of SA3.	May have provided a complicated organization and evaluation schema for behaviors not easily observed and evaluated together. Identifying how to integrate performance distinctions (i.e., behavioral rationales) may not be as important as identifying task levels.
DA	Accuracy of the rank-ordering of dimensions performance levels between ratees.	To increase accuracy of distinctions between ratees on different performance dimensions, to determine differences between ratees in patterns of performance. When training and development information on employee comparison information is important or for appraisal feedback purposes.	FB of true scores and behavioral rationales and discussion increased accuracy on DA1.	May have contributed to a general impression effect within related and highly salient behaviors that helped in evaluating dimension rank ordering between ratees. FB of rationales and true scores may have demonstrated how behaviors were related and evaluated.

Table 26 continued

Index	Ability Tapped	Index Relationship to Training Goals and PA Use	Component Increasing/ Decreasing Index Effectiveness	Suggestions
DISTA	Overall absolute value difference between trainees and experts.	To increase similarity between true overall ratings and trainees' overall ratings when a common perspective is necessary to compare across work group performance in the assignment of group bonuses or projects.	Scale orientation, FB of true scores and behavioral rationales, and discussion each decreased accuracy on DISTA3.	May have provided an organization and evaluation schema that detracted from accuracy estimates for behavior that is difficult to observe, encode together and evaluate; provided inconsistency between task organization format and person/task index. May have detracted from a general impression or simpler strategy for overall ratings. While DISTA meets the goals of FOR training, this training may not quickly increase evaluations of complex behaviors.
LENA	Accuracy of overall rating indicating the amount of under/ over estimation in scores as compared to experts. Thought of as a profile analysis.	To increase similarity between true overall ratings and trainees' overall ratings by considering a consistent evaluative standard across raters. To decrease leniency or severity in overall ratings across raters In order to compare across work groups using the same standards. Uses coincide with DISTA and EL.	Scale orientation decreased accuracy on LENA2.	May have provided an organization schema that was inconsistent with a person/ task index. Under/over-estimations may have been increased for behaviors not easily observed or evaluated Scale orientation presents task focused format without indication of evaluative integration of behaviors.

Table 26 continued

Index	Ability Tapped	Index Relationship to Training Goals and PA Use	Component Increasing/ Decreasing Index Effectiveness	Suggestions
ABVH	Accuracy of the true level of inter-correlation between performance dimensions.	To increase accuracy of true inter-relationships between performance dimensions, when job analysis information is important or when task interrelationships for job training is necessary.	FB of true scores and behavioral rationales and discussion increased accuracy on ABVH1.	May have provided a general impression effect regarding related and highly salient behaviors then led to perception of interrelatedness. Rationales may have led to clearer perception of dimensions similarities.

several principles of learning important for effective training. These include: 1) motivation; 2) similarity between the training stimuli and the actual task; 3) practice; 4) feedback; and, 5) identification of the important elements in training (Goldstein, 1974). FOR training, if designed properly encourages four of these five principles. The fifth, motivation, is a critical issue in this training paradigm but is best left as a consideration between the individual, the trainer, and the organization's support for the transfer of training.

It is believed here that this research investigates in the laboratory many of the critical elements necessary to develop a practical and efficient method of training ratees that can be employed in the field. As stated above, the organization must choose which elements to include based on the purpose of training and the types or complexities of behavior that need to be observed and evaluated. It is the task of the organizational practitioner to provide appropriate behavioral vignettes and prototypes for communicating behavioral standards (Banks & Murphy, 1985) to be used during training.

This research and the training program itself identifies issues critical to the utilization of this type of training. As was identified by Banks and Murphy (1985) as a necessary ingredient for application, the results here suggest which components contribute to the decision of what to observe and how this information is integrated to form evaluations. Where this

research falls short in terms of its generalizability is through the exclusion of consideration of boundary variables, or existing organizational constraints that impede accurate observation and evaluation. This includes consideration of "noisy" observational environments, organizational policies and politics, competing demands, and rater/ratee interactions and preexisting relationships. Unfortunately, investigation of these elements are best addressed, if they can be addressed at all, in field research. Additionally, their influence is probably the among most critical elements contributing to the effectiveness of performance appraisal training programs.

Limitations

A potential constraint in this study was the reliance on expert derived true scores and behavioral rationales gathered through consensus groups. As discussed in Chapter II there is controversy in the literature as to whether whom the field considers "experts" truly have greater abilities in determining true value performance scores, or whether that subjective estimates can be regarded as "true values" (see Reilly et al., 1989 and Sulsky & Balzer, 1988). Unfortunately, in laboratory research and with behavior that is not objectively quantifiable, expert scores remain the best measure available. Additionally, recent research by Reilly et al. (1989) demonstrates support for the notion that individuals' familiar with the job and the

performance appraisal literature do generate true score estimates that closely approximate objective measures of the true performance level.

In the present study expert scores were developed through averaging consensus ratings gathered from two different groups of experts. Developing consensus ratings across a group of four to five individuals has potential confounds in the very process since social influence could engender group conformity that would detract from valid true scores (Yukl, 1981). Soliciting consensus ratings from two different groups raises the risk of confounds exponentially. As described within Chapter III numerous steps were taken in order to reduce these potential sources of bias. Several measures of both reliability and validity were assessed and compared favorably to estimates generated in other research with the same performance vignettes and rating scale (e.g., Athey & McIntyre, 1987; McIntyre et al., 1984; Murphy et al., 1982). However, the level of interrater agreement was rather low (r= .58) and indices of convergent and divergent validity were .58 and .40, respectively. As measures of true scores they are certainly imperfect.

Furthermore, the behavioral rationales that the experts provided in support of their ratings were relayed to trainees as explanation of what information was retrieved in forming their judgements, and how it was cognitively evaluated to determine performance levels for each scale dimension. The experts did not receive frame-of-reference training before viewing, rating, and

justifying their ratings. They were informed of the purpose their ratings and rationales would serve and the purpose of FOR training. They also received scale orientation which should have served as a task oriented organizing schema consistent with the training. Additionally, the frames-of-reference the trainees received were taken directly from the behavioral rationales supplied by the experts. Thus, there was consistency in feedback of true scores and the accompanying justifications.

It is possible, however, that the reliance on "experts" and consensus ratings for developing true scores and behavioral rationales may itself have contributed to confounds in this study. First, group generated true scores and rationales may be systematically different from those generated by individuals. Yukl (1981) suggests that groups may generate more extreme decisions and rely on different information supporting these decisions than individuals. Thus, there is the possibility that the information relayed to trainees may have appeared inconsistent with their own observations and evaluations. Second, there is evidence that individuals with more experience in performance appraisals (Cardy et al., 1987) or with the job (DeNisi & Williams, 1988) may use different organizing schemas than those with less experience. In this manner the expert information regarding evaluation and organization may have appeared very different from the implicit schemata of the

trainees, and it is unclear how trainees would resolve this inconsistency if it occured.

Attribution theory literature (see Feldman, 1981) and the cognitive literature (see DeNisi et al., 1984) suggest different conclusions. The attribution literature suggests that if the information is perceived as coming from a credible source and is not too far removed from the recipients region of acceptance it will be assimilated into the recipients belief system (Aronson et al., 1963). Thus, it is possible they believed and integrated the information and schemata of the experts. The cognitive literature suggests that if the schemata appear very different or present information that appears very different from their own schemata it can result in schemata being abandoned or altered. Thus, trainees may reject their own schema in favor of the experts. However, schema are also difficult to alter and the length of time it would take to alter or abandon schemata and then rebuild schemata is unknown. Additionally, it may be easier for those more familiar with the job and the rating schema to observe, encode, retrieve and evaluate information than those less familiar. Therefore, experts may be identifying aspects of behavior that are difficult for trainees to identify.

Thus, it remains unknown whether trainees accepted the expert true scores and evaluation or rejected this information. Or, if trainees did accept the information whether they could assimilate it during the time span of the training. In the

future it is clear that a manipulation check should be used to evaluate trainees' schemata and/or acceptance of the expert information. The implications of this possibility could be present in two ways. This could account for lower levels of the accuracy indices. It could also account for the number of insignificant results between condition differences.

The selection of vignettes based on expert ratings also deserves discussion given the evidence of the contrast effect reported above. It was necessary to select the best combination of eight possible vignettes to obtain a balance of topics and high, medium, and low performance levels across four different lecturers. In this manner, order of topic presentation could be counterbalanced, level of performance could be counterbalanced within topics, and trainees would not view any lecturer more than once across the four rating occassions. The six vignettes utilized were chosen based on these considerations and design necessities. While the difference between superior performance levels was not ideal, based on the experts' ratings the difference does appear to indicate true differences across performance levels, and similarities within levels.

Several general limitations of this research must also be noted. First is the use of a student sample. It is obvious that this type of sample has limited generalizability to that in a field setting. There are several reasons for this potential lack of generalizability. One can question the motivation level of

students in the training program even though efforts were taken to increase their motivation and highlight the personal meaning training could have for them. Many of the training conditions were long and much information was presented. Whether or not students attended to the entire training module is questionable. Therefore, it is uncertain whether the ratings actually reflect an influence of training. This seems particularly salient given the somewhat more error ridden and less accurate ratings rendered in the longest training condition. Additionally, students can not be considered as familiar with the performance domains of lecture behavior as managers may be of the positions they supervise. Thus, much of the information they are asked to rate may be new to them and not as easily coded, integrated, or recalled. Additionally, this type of appraisal involves upward evaluation whereas in an applied setting the evaluation is typically downward.

Conversely, students represent the most frequently used sample in this type of research. The task of rating professors is a fairly common procedure. For the students in this sample it occurs every semester. By conducting this research with students important issues can be detected in a controlled environment that can then be approached with other samples. Additionally, accurate teacher evaluations are an important concern for college administrators and thus in this context the results can be directly applied if student training becomes a reality.

There is also the concern of the limitations inherent in ratings of videotaped performance vignettes. As noted throughout this literature, performance appraisals occur in a "noisy" environment whereas ratings required of videotapes takes place in a very sterile environment. Thus, there are many more influences on ratings in applied settings than can possibly occur through videotapes. Therefore, in this situation the rating task is much simpler than it would be otherwise. Hence, it should be easier to demonstrate treatment effects, which was not the case here. Conversely, the vignettes offer limited information and the evaluations are required after viewing behavior demonstrated for a very short duration. Normally performance judgements are formed over time and through the accumulation of numerous performance examples. Therefore, raters were required to form judgements with very little information which may have impeded the detection of more or stronger training results.

However, many training programs do employ videotaped performance vignettes. These offer the utilization of accuracy indices given that the measures are difficult, if not impossible to assess in field settings. They also allow for control or manipulation of individual differences within ratees and environmental influences.

Third, the manipulation of the training components may not have been strong enough. While differences were detected across the more salient distinctions (i.e., including feedback of true

scores and behavioral rationales or omitting this feedback), those features that were not as obvious may have failed to have an effect for just this reason. Given the limited time available in a normal class the training may have been conducted too quickly and the various components may not have been reinforced sufficiently. This may account for some of the insignificant results.

Finally, trainees were required to perform posttest ratings immediately after training. In an applied setting employee ratings would generally be developed much later after training had occured. This research does not address the long-term effects of training. As discussed in Chapter II, previous research has not, in general, found long-term effects for training. Given the limited results found here this is likely to be the case as well. Alternatively, it is possible that once given a new understanding or frame-of-reference for evaluating performance, trainees might transfer this knowledge to the work setting when evaluating daily behavior. Given time to apply and integrate this knowledge, rating accuracy might be found to increase over time. That is, while the training may have been initially overwhelming, over time and once applied it may have become clearer and provide a framework leading to greater accuracy in ratings. Future research should investigate the long-term effects of FOR training in an applied setting.

Contributions

There are several contributions of this research which should be noted. This is the first research to systematically investigate the common methods and purposes employed in performance appraisal training and to identify the differential contributions of each. FOR training, as originally conceptualized, was not found to be significantly better than other abbreviated versions of this type of training. However, the pattern of results between the individual components and the effectiveness indices allowed for tentative explanations regarding the processes underlying the component effects. Additionally, by organizing the rating scale into behaviorally consistent factors greater interpretation of results were possible. Second, numerous dependent measures of both accuracy and error indices were employed. In this manner the differential effects of the methods and purposes of training could be explored as they influenced the most commonly used effectiveness indices. Therefore, the results here can generalize to other research that employs these same indices and training methods. Additionally, the use of these indices demonstrated that the effectiveness of training is dependent, in part, on the particular goals of training. Third, the relationship of individual differences and rater attitudes was investigated as a potential influence on training outcomes. Evidence was presented regarding the relationship between positive rater attitudes and accuracy and

dimension ratings. Fourth, many of the identical stimuli and rating materials were used here as are used in other performance appraisal research. Investigation of these materials revealed that they influenced order and topic effects. Fifth, relationships among the accuracy indices were investigated. While a conceptually consistent pattern of results existed, the indices did not generalize to one another. Additionally, some redundancy in indices was noted. Sixth, the influences of accuracy of observational memory on evaluation accuracy were explored. The results suggest that increased memory accuracy was related to greater evaluation accuracy on indices that tapped similar rating abilities, but it decreased evaluation accuracies that addressed different rating abilities. Numerous suggestions for future research were offered at each stage of discussion.

APPENDIX A

Lecture no._____Lecturer name_____Topic_____S.S. no._____

LECTURER BEHAVIOR EVALUATION SCALE

The presentation that you have just watched consisted of a lecture and a question and answer period. Please indicate how much you agree or disagree with the following statements about the lecturer.

A= Strongly Agree

B= Agree

C= Slightly Agree

D= Neither Agree nor Disagree

E= Slightly Disagree

F= Disagree

G= Strongly Disagree

The following statements refer to the lecture only.

1. He seemed interested in the topic.	A	B	C	D	E	F	G
2. He used clear examples to explain abstract ideas.	A	B	C	D	E	F	G
3. He presented the lecture smoothly.	A	B	C	D	E	F	G
4. He integrated the material effectively.	A	B	C	D	E	F	G
5. He followed an outline.	A	B	C	D	E	F	G
6. He followed a logical sequence of thought in his lecture.	A	B	C	D	E	F	G
7. He encouraged questions during the lecture.	A	B	C	D	E	F	G

The following statement refers to the question and answer period only.

8. He provided relevant answers to questions.	A	B	C	D	E	F	G

The following statements refer to both the lecture and the question and answer period.

9.	He was well prepared.	A	B	C	D	E	F	G
10.	He acted relaxed.	A	B	C	D	E	F	G
11.	He spoke clearly and distinctly.	A	B	C	D	E	F	G
12.	He spoke with vigor and enthusiasm.	A	B	C	D	E	F	G
13.	He emphasized important points by raising his voice.	A	B	C	D	E	F	G
14.	He made you interested in the material.	A	B	C	D	E	F	G

OVERALL, how would you rate the presentation that you've just watched?

1 2 3 4 5 6 7

Excellent Average Very Poor

APPENDIX B

PERFORMANCE DIMENSIONS

THOROUGHNESS OF PREPARATION: The extent to which the lecturer is prepared to deliver the lecture.

GRASP OF MATERIAL: The lecturer's mastery of, and knowledge about the subject matter.

ORGANIZATION AND CLARITY: The lecturer's design of the lecture and his/her arrangement of the lecture material such that the material is clear, smooth, and easy to understand.

POISE AND DEMEANOR: The lecturer's self-confidence and mood, reflected by his/her attitude and behaviors during the lecture.

RESPONSIVENESS TO QUESTIONS: The lecturer's ability and ease in answering questions from the audience.

EDUCATIONAL VALUE OF LECTURE: The lecture provides information that is both insightful and applicable to many situations.

RAPPORT WITH AUDIENCE: The lecturer's ability in interacting with the audience when lecturing and answering questions.

SPEAKING ABILITY: The lecturer's ability to speak clearly and smoothly at a moderate speed, and keeping the audience interested.

OVERALL RATING: Based upon all of the above performance areas, the rating that summarizes overall performance.

APPENDIX C

INTRODUCTION

Hello, my name is _____ and I'm here today to ask you to participate in a training experience in conducting performance appraisals.

A performance appraisal is a systematic review of an individual's performance which is used to evaluate the effectiveness of his or her work. Now, many of you may already have some experience in performance appraisals where you work. You may have been the one who's performance was reviewed, or you may have been the one completing a performance appraisal on someone else. If neither of these situations have happened yet, I can pretty well guarantee you that they will someday.

Now, if all of you have been at this school for at least one semester then I'm sure that you've had the opportunity to complete student evaluations on your professors. These student evaluations, or performance appraisals, are a very important tool for the school. They provide the student's point of view of a professor's teaching effectiveness, help identify weaknesses in performance and set goals for improvement, and, count toward whether a professor will acheive tenure. For adjunct faculty the performance appraisal takes on even more importance since it plays a large part in whether or not they will be rehired for the following semester. Additionally, at this early stage in an

adjunct's teaching career this important performance feedback can really have an effect on how they teach students in the future. So you can see just how important these appraisal are, and how crucial it is to make sure that they are as accurate as possible.

Well, we've been trying to determine ways of improving the accuracy of student evaluations and felt that one method may be to train students in actually rating performance. It just seems to make sense that rating someone's performance should be a trainable skill.

Fortunately, the timing of this training program worked out well since there are four candidates for adjunct faculty positions who where willing to be videotaped while conducting a lecture. These people are up for positions here and really want your feedback on their performance. So we're really able to accomplish several things at once here. We can train students to conduct more accurate reviews, give you first hand experience in a real live research study, get your feedback on four job candidates, and provide the candidates with this feedback.

So, if everybody is ready and willing to participate we can begin with the performance appraisal training.

What is going to be happening is that first we'll go over the duties and qualifications of the adjunct facult position. Next, we'll review the new student evaluation form and define the

dimensions of behavior covered in this form. Then, you'll get the opportunity to practice rating two different adjunct's performance by watching them deliver lectures. (We've given them a choice between two lectures so that we could better standardize the subsequent hiring process.) After this I'll tell you how the experts rated the adjuncts and discuss why they rated them this way. And then, since you will have been sufficiently trained you'll get to rate the performance of the last two candidates.

DEBRIEFING

The study that you just took part in, is in fact, a study on performance appraisal training. Everything that you learned is part of an actual training program that I'm testing out in order to determine how effective this type of training is, and to answer why it is effective. The training involves real skills that you can take with you and apply to conducting a performance appraisal on your subordinates at work, or in better understanding your own performance appraisal.

What I've been doing is systematically manipulating various components of the training program in order to determine which components are most effective. For example, in this group you received all components of the training except (insert component) so I'll be able to compare your results against a group, or condition, that did receive that part of the training to see which group generate appraisal scores closer to the scores that the experts generated. If you did as well as the other group then the results would suggest that this component is not really adding anything to the training. However, if the other group had scores closer to the expert's scores then the results would suggest that this component was an integral, and important, part of the training.

What I did, in fact, exaggerate was that we were contemplating developing training programs to increase the accuracy of the student evaluations. This is not to say that we're not

interested in achieving the highest levels of accuracy possible, but unfortunately we don't have the funds to train students in conducting performance appraisals. Therefore, we have to rely on your best efforts when you complete the student evaluations. Additionally, the lecturers you observed are not apply for positions here. They are actors who performed the lectures using a carefully designed script.

Please, in the interests of research (and my dissertation) do not talk about this study with any of your friends here at school since they may be participating the training at some point during the semester. If they participate in this study with prior knowledge about the conditions or the actual purpose of the study it would invalidate the results.

For those people interested in the results of this study and the effectiveness of the training you participated in I will be posting the results outside my office (the 18th Street building, room 1108) once all the data has been collected and analyzed. Additionally, if anybody is interested in talking about this further either now or after you've received the results please do not hesitate to contact me. My office number is 725-3200.

Thank you so much for your participation.

APPENDIX D

S.S. no._____

PERSONAL HISTORY

The following questions are for research purposes only and will be kept strictly confidential. Please answer each question by checking the appropriate category.

1. Sex:
   ____male ____female

2. Age: ____

3. Ethnic/Racial Heritage:
   ____White ____Native American
   ____Black ____Pacific Islander
   ____Hispanic ____Other
   ____Asian

4. Enrollment Status:
   ____Full-time Day Student
   ____Part-time Day Student
   ____Full-time Night Student
   ____Part-time Night Student

5. Class:
   ____Freshman ____Sophomore
   ____Junior ____Senior
   ____Masters

6. Area of Academic Major:
   ____Business Administration ____Education
   ____Social Science ____Mathematics
   ____Humanities ____Science
   ____Other

7. Language used most often:
   ____English ____Spanish
   ____Chinese ____Other

8. How long have you lived in the New York area?
   ____under one year ____5-10 years
   ____1-2 years ____10 or more years
   ____2 1/2-4 years ____entire life

9. Living Situation:

____with family	____alone
____with spouse and/or children	____with roommate
____with significant other	

10. Not including yourself, how many people do you live with?

____one	____6-10
____two	____10 or more
____3-5	

11. Do you live in a:

____one family home	____one bedroom apartment
____two family home	____two bedroom apartment
____studio apartment	____three or more bedroom apartment

12. To what extent do you feel your living situation is crowded?

1	2	3	4	5	6	7
Not at all						Extremely crowded

13. Employment Status:

____Full-time ____Part-time

14. At work:

have you ever had your performance formally reviewed ?

____Yes ____No

have you ever formally reviewed another person's performance?

____Yes ____No

15. Before today, had you had any formal prior knowledge of either of the topics presented?

A. Stress and Crowding: ____Yes ____No

B. Self-fulfilling Prophecy: ____Yes ____No

APPENDIX E

Please indicate where you would place yourself along each of the following statements.

1. In general, the lecture was:

1	2	3	4	5	6	7
Very Interesting						Very Boring

2. This lecture topic is:

1	2	3	4	5	6	7
Very Dull						Very Interesting

3. I thought the lecture was:

1	2	3	4	5	6	7
Very Enjoyable						Very Unpleasant

4. I would take a class taught by this lecturer.

1	2	3	4	5	6	7
Strongly Agree	Agree	Slightly Agree	Neutral	Slightly Disagree	Disagree	Strongly Disagree

5. I would like to learn more about this topic.

1	2	3	4	5	6	7
Strongly Disagree	Disagree	Slightly Disagree	Neutral	Slightly Agree	Agree	Strongly Agree

6. I would recommend him for the job of lecturer.

1	2	3	4	5	6	7
Strongly Agree	Agree	Slightly Agree	Neutral	Slightly Disagree	Disagree	Strongly Disagree

APPENDIX F

x_{ijk} = rating given to i^{th} ratee on j^{th} dimension by the k^{th} rater, i= 1... n, j=... d, k= 1... m

t_{ij} = "true score" for i^{th} ratee on j^{th} dimension

$x_{i.k}$ = mean rating for i^{th} ratee by k^{th} rater

$x_{.jk}$ = mean rating on j^{th} dimension by k^{th} rater

$x_{..k}$ = mean rating by k^{th} rater

$t_{.j}$ = mean true score on dimension j

$t_{i.}$ = mean true score for i^{th} ratee

$t_{..}$ = mean true score

s^2_{xij} = variance of $x_{i.k}$

s^2_{xjk} = variance of $x_{.jk}$

s^2_{xijk} = variance of $x_{ijk} - x_{i.k} - x_{.jk} + x_{..k}$

s^2_{ti} = variance of $t_{i.}$

s^2_{tj} = variance of $t_{.j}$

s^2_{tij} = variance of $t_{ij} - t_{i.} - t_{.j} + t_{..}$

z_{ktxj} = z - transform of r_{ktxj}

$$r_{ktxj} = \frac{\frac{1}{n}\sum_{i=1}^{n}\left[(x_{ijk} - x_{\cdot jk})\,(t_{ij} - t_{\cdot j})\right]}{\sqrt{\left[\frac{1}{n}\sum_{i=1}^{n}(x_{ijk} - x_{\cdot jk})^2\right]\left[\frac{1}{n}\sum_{i=1}^{n}(t_{ij} - t_{\cdot j})^2\right]}}$$

$$r_{kxjj'} = \frac{\frac{1}{n}\sum_{i=1}^{n}\left[(x_{ijk} - x_{\cdot jk})\,(x_{ij'k} - x_{\cdot j'k})\right]}{\sqrt{\left[\frac{1}{n}\sum_{i=1}^{n}(x_{ijk} - x_{\cdot jk})^2\right]\left[\frac{1}{n}\sum_{i=1}^{n}(x_{ij'k} - x_{\cdot j'k})^2\right]}}$$

$$r_{tjj'} = \frac{\frac{1}{n}\sum_{i=1}^{n}\left[(t_{ij} - t_{\cdot j})\,(t_{ij'} - t_{\cdot j'})\right]}{\sqrt{\left[\frac{1}{n}\sum_{i=1}^{n}(t_{ij} - t_{\cdot j})^2\right]\left[\frac{1}{n}\sum_{i=1}^{n}(t_{ij'} - t_{\cdot j'})^2\right]}}$$

$z_{xjj'k}$ = z - transform of $r_{kxjj'}$

$$EL = (x_{..k} - t_{..})^2 \qquad (1)$$

$$DE = \frac{1}{n}\sum_{i=1}^{n}[(x_{i.k} - x_{..k}) - (t_{i.} - t_{..})]^2 \quad (2)$$

$$= s^2_{xik} + s^2_{ti} - 2Cov\,(x_{i.k}\ t_{i.}) \qquad (3)$$

$$SA = \frac{1}{d}\sum_{j=1}^{d}[(x_{.jk} - x_{..k}) - (t_{.j} - t_{..})]^2 \quad (4)$$

$$= s^2_{xjk} + s^2_{tj} - 2Cov\,(x_{.jk}\ t_{.j}) \qquad (5)$$

$$DA = \frac{1}{nd}\sum_{i=1}^{n}\sum_{j=1}^{d}[(x_{ijk} - x_{i.k} - x_{.jk} + x_{..k}) - (t_{ij} - t_{i.} - t_{j.} + t_{..})]^2 \qquad (6)$$

$$= s^2_{xijk} + s^2_{tij} - 2Cov\,(x_{ijk}\ t_{ij}) \qquad (7)$$

$$DISTA = \frac{1}{nd}\sum_{i=1}^{n}\sum_{j=1}^{d}|x_{ijk} - t_{ij}| \qquad (8)$$

$$\text{DIFFERENTIAL CORRELATIONAL ACCURACY} = \frac{1}{d}\sum_{j=1}^{d} z_{ktxj} \qquad (9)$$

$$\text{CORRELATIONAL HALO ACCURACY} = \frac{2}{d(d-1)}\sum_{j=1}^{d}\sum_{\substack{j'=1 \\ j<j'}}^{d}(r_{kxjj'} - r_{tjj'}) \qquad (10)$$

$$\text{ABVH} = \frac{1}{nd} \sum_{i=1}^{n} \left| \sum_{j=1}^{d} [(x_{ijk} - x_{i\cdot k})^2 - (t_{ij} - t_{i\cdot})^2] \right| \quad (11)$$

$$\text{LENA} = \frac{1}{d} \sum_{j=1}^{d} (x_{\cdot jk} - t_{\cdot j}) \quad (12)$$

APPENDIX G

Lecturer no._____Lecturer name_____Topic_____

Behavior Frequency Scale

Rate the frequency with which each of the following behaviors occured on each tape. Use the following scale to rate frequency.

1	2	3	4	5	6	7
Never	Almost Never	A Few Times	About Half of the Time	Often	Most of the Time	All of the Time

1. Examples were presented which were clearly related to the central topic. _____
2. Lecturer used purposeful nonverbal behaviors (smiles, points) to emphasize points. _____
3. Lecturer stops in mid-sentence. _____
4. Lecturer is hesitant, says "eh" or "am." _____
5. Lecturer loses eye contact with audience. _____
6. Lecturer provides facts or evidence to support broad generalizations. _____
7. Lecturer speaks fast. _____
8. Lecturer acts nervous. _____
9. Lecturer speaks in a monotone for a sustained period. _____
10. Lecturer gives clear answers to questions. _____
11. Lecturer varies his facial expression. _____
12. Lecturer appears unsure of what he is saying. _____

REFERENCES

Aronson, E., Turner, J. A., & Carlsmith, J. M. (1963). Communicator credibility and communication discrepancy as determinants of opinion change. Journal of Abnormal and Social Psychology, 67, 31-36.

Athey, T. R., & McIntyre, R. M. (1987). Effect of rater training on rater accuracy: Levels-of-Processing theory and social facilitation theory perspectives. Journal of Applied Psychology, 72, 567-572.

Banks, C. G., & Murphy, K. R. (1985). Toward narrowing the research-practice gap in performance appraisal. Personnel Psychology, 38, 335-345.

Becker, B. E., & Cardy, R. L. (1986). Influence of halo error on appraisal effectiveness: A conceptual and empirical reconsideration. Journal of Applied Psychology, 71, 662-671.

Bernardin, H. J. (1978). Effects of rater training on leniency and halo errors in student ratings of instructors. Journal of Applied Psychology, 63, 301-308.

Bernardin, H. J., & Buckley, M. R., (1981). Strategies in rater training. Academy of Management Review, 6, 205-212.

Bernardin, H. J., & Pence, E. C. (1980). Effects of rater training: Creating new response sets and decreasing accuracy. Journal of Applied Psychology, 65, 60-66.

Bernardin, H. J., Villanova, P. (1986). Performance appraisal. In E. A. Locke (Ed.) Generalizing from laboratory to field settings. Lexington, MA.: Lexington Books (pp. 43-62).

Bernardin, H. J., & Walter, C. S. (1977). Effects of rater training and diary-keeping on psychometric error in ratings. Journal of Applied Psychology, 62, 64-69.

Bitner, R. H. (1948). Developing an industrial merit rating procedure. Personnel Psychology, 1, 403-432.

Borman, W. C. (1975). Effects of instructions to avoid halo error on reliability and validity of performance evaluation ratings. Journal of Applied Psychology, 60, 556-560.

Borman, W. C. (1977). Consistency of rating accuracy and rating errors in the judgment of human performance. Organizational Behavior and Human Performance, 20, 238-252.

Borman, W. C. (1978). Exploring upper limits of reliability and validity in job performance ratings. Journal of Applied Psychology, 63, 135-144.

Borman, W. C. (1979). Format and training effects on rating accuracy and rater errors. Journal of Applied Psychology, 64, 410-421.

Borman, W. C., & Dunnette, M. D. (1975). Behavior based versus trait-oriented performance ratings: An empirical study. Journal of Applied Psychology, 64, 412-421.

Brown, E. B. (1967). Influence of training, method, and relationship on the halo effect. Journal of Applied Psychology, 52, 195-199.

Cafferty, T. P., DeNisi, A. S., & Williams, K. J. (1986). Search and retrieval patterns of performance information: Effects on evaluations of multiple targets. Journal of Personality and Social Psychology, 50, 676-683.

Campbell, J., Dunnette, M., Lawler, E., & Weick, K. (1970). Managerial behavior, performance and effectiveness. New York: McGraw-Hill.

Cardy, R. L., Bernardin, H. J., Abbott, J. G., Senderak, M. P., & Taylor, K. (1987). The effects of individual performance schemata and dimension familiarization on rating accuracy. Journal of Occupational Psychology, 60, 197-205.

Cleveland, J. N., Murphy, K. R., and Williams, R. E. (1989). Multiple uses of performance appraisal: Prevalence and correlates. Journal of Applied Psychology, 74, 130-135.

Cooper, W. H. (1981). Ubiquitous halo. Psychological Bulletin, 90, 218-244.

Costin, F. (1974). Measuring lecturing behavior of college instructors. Professional Psychology, 1, 106-108.

Cronbach, C. J. (1955). Processes affecting scores on understanding of others and assuming "similarity". Psychological Bulletin, 52, 177-193.

Davis, B. L., & Mount, M. K. (1984). Effectiveness of performance appraisal training using computer assisted instruction and behavior modeling. Personnel Psychology, 37, 439-452.

DeNisi, A. S., Cafferty, T. P., & Meglino, B. M. (1984). A cognitive model of the performance appraisal process: A model and research propositions. Organizational Behavior and Human Performance, 33, 360-396.

DeNisi, A. S. & Williams, K. J. (1988). Cognitive approaches to performance appraisal. In Research in Personnel and Human Resources Management (Vol 6, pp. 109-155). Greenwich, CT: JAI Press.

Fay, C. H., & Latham, G. P. (1982). Effects of taining and rating scales on rating errors. Personnel Psychology, 35, 105-116.

Feldman, J. M. (1981). Beyond attribution theory: Cognitive processes in performance appraisal. Journal of Applied Psychology, 66, 127-148.

Fisicaro, S. A. (1988). A reexamination of the relation between halo error and accuracy. Journal of Applied Psychology, 73, 239-244.

Hedge, J. W., & Kavanagh, M. J. (1988). Improving the accuracy of performance evaluations: Comparison of three methods of performance appraisal training. Journal of Applied Psychology, 73, 68-73.

Ilgen, D. R. & Feldman, J. M. (1983). Performance appraisal: A process focus. In B. M. Staw and L. L. Cummings (Eds.), Research in Organizational Behavior (Vol. 5, pp. 141-197). Greenwich, CT: JAI Press.

Ivancevich, J. M. (1979). Longitudinal study of the effects of rater training on psychometric error in ratings. Journal of Applied Psychology, 64, 502-508.

Kavanagh, M., MacKinney, A., & Wolins, L. (1971). Issues in managerial performance. Psychological Bulletin, 73, 34-49.

Landy, F. S. & Farr, J. L. (1980). Performance rating. Psychological Bulletin, 87, 72-107.

Landy, F. J., & Trumbo, D. A. (1980). The Psychology of Work Behavior (rev. ed.), Homewood, IL.: Dorsey Press.

Latham, G. P., Wexley, K. N., & Pursell, E. D. (1975). Training managers to minimize rating errors in the observation of behavior. Journal of Applied Psychology, 36, 29-33.

Lee, C. (1985). Increasing performance appraisal effectiveness: Matching task types, appraisal process, and rater training. Academy of Management Review, 10, 322-331.

Levine, H. Z. (1986). Performance appraisal at work. Personnel, 63, 63-71.

Levine, J., & Butler, J. (1952). Lecture versus group discussion in changing behavior. Journal of Applied Psychology, 36, 29-33.

Lord, R. G. (1985). An information processing approach to social perceptions, leadership and behavioral measurement in organizations. In B.M. Staw and L.L. Cummings (Eds.), Research in Organizational Behavior (Vol 7, pp. 87-128). Greenwich, CT: JAI Press.

Lord, R. G. & Nathan, R. G. (1983). Cognitive categorization and dimensional schemata: A process approach to the study halo in performance ratings. Journal of Applied Psychology, 68, 102-114.

McIntyre, R. M., Smith, D. E., & Hassett, C. E. (1984). Accuracy of performance ratings as affected by rater training and perceived purpose of rating. Journal of Applied Psychology, 69, 147-156.

Murphy, K. R., & Balzer, W. K. (1981, August). Rater errors and rating accuracy. Paper presented at the Annual Convention of the American Psychological Association, Los Angeles.

Murphy, K. R., & Balzer, W. K. (1986). Systematic distortions in memory-based ratings: Consequences for rating accuracy. Journal of Applied Psychology, 71, 39-44.

Murphy, K. R., Balzer, W. K., Kellam, K. L., & Armstrong, J. G. (1984). Effects of the purpose of rating on accuracy in observing teacher behavior and evaluating teaching performance. Journal of Educational Psychology, 76, 45-54.

Murphy, K. R., Balzer, W. K., Lockhart, M. C., & Eisenman, E. J. (1985). Effects of previous performance on evaluationsof present performance. Journal of Applied Psychology, 70, 72-84.

Murphy, K. R., & Cleveland, J. N. (1990). Personal conversation.

Murphy, K. R., Garcia, M., Kerkar, S., Martin, C., & Balzer, W. K. (1982). Relationship between observational accuracy and accuracy in evaluating performance. Journal of Applied Psychology, 67, 320-325.

Murphy, K. R., Martin, C., Garcia, M. (1982). Do behavioral observation scales measure observation? Journal of Applied Psychology, 67, 512-519.

Pulakos, E. D. (1984). A comparison of rater training programs: Error training and accuracy training. Journal of Applied Psychology, 69, 581-588.

Pulakos, E. D. (1986). The development of training programs to increase accuracy with different rating tasks. Organizational Behavior and Human Decison Processes, 38, 76-91.

Pursell, E. D., Dossett, D. L., & Latham, G. P. (1980). Obtaining valid predictors by minimizing rating errors in the criterion. Personnel Psychology, 33, 91-96.

Roach, D. W. & Gupta, N. (1990, April). Relationship among components of rating accuracy in a realistic setting. Paper presented at the fifth annual meeting of The Society for Industrial and Organizational Psychology, Miami Beach.

Saal, F. E., Downey, R. G., & Lahey, M. A. (1980). Rating the ratings: Assessing the psychometric quality of rating data. Personnel Bulletin, 88, 413-428.

Scrull, T. K. & Brand, J. F. (1983). Memory of information about persons: The effect of encoding operations on subsequent retrieval. Journal of Verbal Learning and Verbal Behavior, 22, 219-230.

Smith, D. E. (1986). Training programs for performancce appraisal: A review. Academy of Management Review, 11, 22-40.

Smith, P. C., & Kendall, L. M. (1963). Retranslation of expectation: An approach to the construction of unambiguous anchors for rating scales. Journal of Applied Psychology, 47, 149-155.

Smither, J. W., Barry, S. R., & Reilly, R. R. (1989). An investigation of the validity of expert true score estimates in appraisal research. Journal of Applied Psychology, 74, 143-151.

Smither, J. W., Reilly, R. R., & Buda, R. (1988). Effect of prior performance information on ratings of present performance: Contrast versus assimilation revisited. Journal of Applied Psychology, 73, 487-496.

Sulsky, L. M., & Balzer, W. K. (1988). Meaning and measurement of performance rating accuracy: Some methodological and theoretical concerns. Journal of Applied Psychology, 73, 497-506.

Taylor, S. E., & Crocker, J. (1981). Schematic bases of social information processing. In E. T. Higgins, C. P. Herman, & M. P. Zanna (Eds.), Social Cognition: The Ontario Symposium on Personality and Social Psychology. Hillsdale, NJ: Erlbaum.

Thorton, G. C. & Zorich, S. (1980). Training to improve observer accuracy. Journal of Applied Psychology, 65, 351-354.

Vance, R. J., Kuhnert, K. W., & Farr, J. L. (1978). Interview judgments: Using external criteria to compare behavioral and graphic scale ratings. Organizational Behavior and Human Performance, 22, 279-294.

Warmke, D. L., & Billings, R. S. (1979). Comparison of training methods for improving the psychometric quality of experimental and administrative performance ratings. Journal of Applied Psychology, 64, 124-131.

Williams, K. J., DeNisi, A. S., Meglino, B. M., & Cafferty, T. P. (1986). Initial decisions and subsequent performance ratings. Journal of Applied Psychology, 71, 189-195.

Williams, K. J., Wickert, P., & Peters, R. D. (1985). Appraisal salience: Effects of instructions to subjectively organize information. Proceedings of the Southern Management Association Meetings, 124-126.

Zedeck, S., & Cascio, W. (1982). Performance decision as a function of purpose of rating and training. Journal of Applied Psychology, 67, 752-758.